# Mental Contents, Tracking Counterfactuals, and Implementing Mechanisms

Josep E. Corbí and
Josep L. Prades

In the ongoing debate, there are a set of mind-body theories sharing a certain physicalist assumption: whenever a genuine cause produces an effect, the causal efficacy of each of the nonphysical properties that participate in that process is determined by the instantiation of a well-defined set of physical properties. These theories would then insist that a nonphysical property could only be causally efficacious insofar as it is physically implemented. However, in what follows we will argue against the idea that fine-grained mental contents could be physically implemented in the way that functional properties are. Therefore, we will examine the metaphysical conditions under which the implementing mechanism of a particular instance of a functional property may be individuated, and see how genuine beliefs and desires—insofar as they track the world—cannot meet such conditions.

In the current controversy, the crucial question is this: how can mental properties be causally relevant and, nevertheless, respect the causal primacy of physical properties? We use the phrase '*Causal physicalism*' to refer to the kind of answer that the dominant perspective provides for this question. 'Causal physicalism' is meant to pick out a set of theories that, despite their divergences in other respects, share a certain assumption, namely: that every case of nonphysical (and, hence, mental too) causation has to be conceived as systematically dependent on

certain physical processes which underlie it. For whatever the causal efficacy of a nonphysical property may be *in a particular case*, it cannot go beyond the causal powers of the physical properties that are involved *on that particular occasion*. Physical properties appear to the causal physicalist as the basic properties of the world, while the rest of the properties come up as dependent upon those putative basic properties.[1]

Following up Jaegwon Kim,[2] we can see this view as being motivated by two rather plausible principles, namely: Physical Closure and Explanatory Exclusion. The first principle affirms the causal closure of the physical world: every physical effect has a complete physical cause. Explanatory Exclusion, on the other hand, stresses that ". . . [t]wo or more complete and independent explanations of the same event or phenomenon cannot coexist."[3] This principle just sounds like a reasonable way of expressing the conviction that overdetermination cannot abound and, consequently, that complete causes must not only be sufficient for their effects but also necessary for them. It trivially follows that there cannot be more than one complete causal line for each physical effect and, given Physical Closure, this causal line must be physical. Hence, insofar as a property can have a causal influence on the physical world (as mental properties do), its physical import, its causal relevance on the physical world, cannot be independent of the causal powers of the physical line that constitutes the complete cause of the physical effect at stake. In other words, the causal physicalist would insist that a nonphysical property F could only be causally efficacious if, for each tokening of F, there is a set of physical properties that constitutes its implementing mechanism, that is, a set of physical properties about which it is true that if this set had not taken place, the effects of the relevant tokening of F would not have occurred. Accordingly, Charles' present desire of reading *Ulysses* could only causally explain Charles' behavior of moving toward the bookshelf if there is a set of physical properties that constitutes the physical implementation of that desire, that is, if there is a set of physical properties that bears the relevant counterfactual link with that behavior.

This is, however, a picture that we would like to resist. In fact, we intend to demonstrate that causal physicalism is internally inconsistent because it relies on two incompatible assumptions, namely: (a) that each effect has a complete cause and (b) that effects are not generally overdetermined. We will thus argue that, insofar as the causal physicalist is bound to stick to the notion of complete cause, she lacks the conceptual resources to individuate causal processes in such a way that massive overdetermination is averted. This upshot will crucially affect the conditions under which nonphysical properties may be causally relevant.

For it will invite us to reconsider the metaphysical principles by which implementing mechanisms could at all be individuated. The crucial point is that, after a proper elucidation of such metaphysical principles, we will conclude that there is no set of physical properties that could bear the relevant counterfactual link with the fact that an individual has a certain fine-grained mental content, so that it could count as the implementing mechanism of that mental content. In short, we will claim that fine-grained mental contents are individuated in such way that they cannot be physically implemented.

## I. COMPLETE CAUSES AND OVERDETERMINATION

The striking of a match may, in the circumstances, be a necessary and sufficient condition for the explosion of the building. The match's striking will thus be, together with plenty of other nonsuperfluous causal factors, one of the ingredients of the complete cause of that explosion. At first sight, the presence of air may come up as one of those causal factors. But, indeed, air cannot be the nonsuperfluous causal factor. From the point of view of the causal efficacy of the match's striking, the particular combination of oxygen plus hydrogen existing in the air is not necessary. It is only oxygen which is necessary for the match's lighting. So, it is oxygen, and not air, that should form a part of the causal line. But, of course, not all the oxygen that is present in the combustion may be strictly necessary for the combustion to take place. So, part of the oxygen that is in fact in the room may be neglected. Yet this attempt to discount the superfluous elements seems to undermine the possibility of individuating a causal line that should be counterfactually necessary for a given effect. For what applies to oxygen, goes for other putative causal factors actually given in the circumstances, so that various combinations of different putatively nonsuperfluous causal factors will give rise to several causal lines composed of disparate ingredients.

The different causal lines may not only differ in the combination of the critical values for the same parameters. The presence of a certain parameter can make the presence of some others redundant. The critical, minimal value of a certain parameter may not only depend on the values of the rest of parameters in its causal line, but also on whether certain other parameters are included or not in the causal line. This is, in fact, quite a common phenomenon. A lethal poison, for instance, produces many different chemical and physiological effects that, in combination, can kill a human being. There are many ways of grouping those chemical and physiological processes so that each of the resulting sets will still be lethal by itself, even if all of them are instantiated together.

Each of these lines will constitute a sufficient condition for the effect, but none of them will be necessary in the circumstances and, therefore, every effect will be massively overdetermined by its causal lines.

The causal physicalist concedes that the existence of several independent causal lines (each one of them enough to produce the effect) for the same effect must be very rare. Our worry is, however, that this rarity cannot be ensured if causal lines have to be individuated in terms of their physical determinations. Even when those causal lines differ only in having different combinations of values for the same parameters, the causal efficacy of each line does not require the instantiation of the other. This is more obvious when we accept, as in the explosion and the poisoning case, the existence of causal lines that involve some differential parameters. If the requirement of a complete cause has to be of any help to the causal physicalist, it cannot be reduced to the complete cause with all its physical determinations. For, otherwise, we would have more than one causal line in each case of causation, and this would involve massive overdetermination.

The causal physicalist might still insist that the different basic causal lines involved in the explosion do not constitute a genuine case of overdetermination. For they are not, properly speaking, like the case in which two darts simultaneously hit a balloon. The situation looks closer to the case in which the balloon is impacted by the distinct points of a single fork. So, they should be construed as different ingredients of the same causal process. This is, in our view, the right answer. The problem is that the different causal lines that we have picked out are complete and independent in the only sense to which the causal physicalist may accede, namely: each causal line is enough to produce the effect and, hence, none of them requires the instantiation of any other causal line to bring about the effect. In this sense (the only available one to the causal physicalist, we must insist), two causal lines may be independent even if not all of their parameters are. Yet, this notion of complete cause and its associated notion of independence entails, as we have seen, that effects are generally overdetermined.

A central implication of this upshot is that, if we wish to retain the idea that overdetermination does not abound, we must accept that certain higher-order properties may often be better candidates to the role of cause than any set of physical, basic causal lines. This claim will surely guide our analysis of the conditions under the implementing mechanism of an instance of higher-order property may be individuated and, consequently, our treatment of mental causation. We shall thus conclude that, in those cases where a fine-grained mental state causally explains the agent's behavior, there is no set of neurophysiological

properties that may bear with that mental state the relevant counterfactual link to count as its implementing mechanism.

## II. IMPLEMENTING MECHANISMS AND TRACKING COUNTERFACTUALS

To elucidate the notion of 'implementing mechanism', let us focus on an apparently unrelated issue, namely: the conditions under which the causal powers of objects are individuated. It is clear, to begin, that objects are extremely robust concerning their properties: they tend to preserve most of their functional properties through highly diverse spatio-temporal conditions. A car, a table, a carburetor, a key, a brake do instantiate those functional properties even if we would take them to very remote areas in space. Needless to say, the robust stability of a particular key regarding the functional property 'being a key' presupposes a more general kind of stability: a key cannot continuously change its shape or its mass. For, if it did so, it could not preserve the ability to open a certain kind of lock in a wide range of circumstances. The kind of stability that is proper to objects is connected, as we shall see, with the possibility of individuating implementing mechanisms for functional properties.

If, here and now, placing a certain object in the right slot is counterfactually necessary for the production of a functional effect, then there is, here and now, another counterfactual link: every property in the bunch that is formed by the stability of the object, is counterfactually necessary for the effect. If the causal efficacy of a particular key is judged to be, here and now, necessary for the production of an effect, a lot of properties are judged to be features of this particular causal process, just because they are bundled by the stability of the key. In other words, we have here a counterfactual link between different properties of an object and an effect generated by the mere stability of the object, even if some of those properties are not causally relevant features for the production of the effect at stake. It is in this precise sense that the redness of the key might be described as a necessary feature of the process by which this particular key, here and now, opens this lock: this necessity does not depend on the causal relevance of the color of our key. We can call this link between the different features of an object a 'stability link' (*S-link*) and, thus, we can say that many features of a causal process are counterfactually necessary for the effect, not because they are the causally relevant features, but because they are S-linked to certain causally relevant features.

Of course, there is another sense in which we could say that the properties of an object are counterfactually linked to an effect. The particular shape of our key, for instance, could be regarded as crucially relevant to the effect, in the sense that a similar key with a different shape would be unable to open the lock. Without this shape, our key would have lost its lock-opening abilities. Call this link 'the causally relevant features link' (*CRF-link*). This is, indeed, the kind of counterfactual link by means of which we identify the causally relevant features.

The distinction between these two kinds of counterfactual links (i.e., CRF- and S-links) makes it possible to claim that there are certain properties that can be both regarded as superfluous and, nonetheless, count as actual features of a causal process. Some might use our intuitions about the causal irrelevance of certain features to try to fix the implementing mechanism of a certain functional process only in terms of the physical features that are causally relevant. Yet, our arguments bring to light that nothing important is gained by this move. It is true that there is a difference in causal relevance between the color and other properties, but it is also true that we cannot assume that it is possible to fix an implementing mechanism in physical terms that could be free from the kind of counterfactual link we have called 'S-link'. In this respect, we have shown that our key can be described as having many different physical bases that fix its causal ability to open certain locks, but we have no reason to prefer one of them over the rest. If we want to avoid the self-refuting conclusion that there are many different implementing mechanisms of a particular instance of this functional property, we must put together all those different physical bases as being determinations of a single physical implementation of this functional property. In fact, those different physical bases are mentioned in the first place, as opposed to other putative ones, because they are implicitly recognized as being linked together: they are linked together by the particular stability of our key. It seems then that one can only obtain a sensible notion of 'implementing mechanism' insofar as one accepts that S-links do fix the size and boundaries of the relevant implementing mechanism, and one recognizes that the sort of 'over-determination' generated by S-links is not the kind of overdetermination that our causal ontology tends to exclude.

Let us now turn to mental causation, and see what we can learn from the distinction between S-links and CRF-links as to the determination of the implementing mechanism of a fine-grained psychological process. Imagine that Charles, moved by his desire to read *Ulysses,* goes to the bookshelf and picks up his copy of Joyce's text. How could we fix the implementing mechanism that underlies this particular instance of

a causally relevant mental property? The causal physicalist assumes that, insofar as there is a physical/neurophysiological description that is enough to fix any physical outcome, the causal efficacy of the desire has to be accounted for in terms of the causal efficacy of a certain physical/neurophysiological causal chain.

Now, we should notice an important difference between mental contents and those cases in which a particular object can be selected as the object that is doing the causal job. The difference lies in the fact that, insofar as we only pick up the implementing neural state by defining it as whatever neural state that is causally responsible for the relevant movements of Charles' body, we have not yet drawn the required distinction between the relevant S- and CRF-links. We are just trying to find the lower-level properties that back certain CRF-links, without using our intuitions about S-links. Then, it is not yet clear what we mean by 'implementing mechanism' in this case. To put it another way, our opponent must, at least, grant that a complex neurological state would only count as an implementing mechanism of Charles' desire if the ensuing story were true of it:

> Charles' desire is implemented by those neurophysiological structures that, here and now, do produce the changes in the material world that are proper to the causal powers of that desire. But, just as it happens with every higher-order property that is implemented by lower-level properties, the causal powers of the desire are individuated because, here and now, Charles would not have moved toward the bookshelf if he had not moved in the particular way he did.

This story simply seeks to apply to Charles' desire the conditions that have previously set for standard functional properties. We are, though, convinced that the analogy fails at a crucial point. Consider the last sentence of the paragraph. One may reasonably suspect that a counterfactual like "here and now, Charles would have not moved toward the bookshelf if he had not moved in the particular way he did" can never be true. In fact, it is a very common assumption that the causal powers of a genuine desire are individuated in such a way that the detailed way in which our bodies actually move are not counterfactually linked to the causal powers of the desire. To see this, we just need to recall an intuition that most people judge fundamental to respect, and that no approach to mental causation could plausibly deny, namely, that

> . . . if Charles had been a few more meters away from the bookshelf, his desire to read *Ulysses,* along with some perceptual beliefs, would have led him to take a few more steps so that he could reach the bookshelf.

In fact, this is the kind of counterfactual that one takes for granted as one may claim that Charles' desire to read *Ulysses,* together with some perceptual contents, causally explains his movement towards the bookshelf, and not just his movement towards a certain location.[4] More generally, we could say that whenever we explain someone's behavior in terms of her mental contents, we are assuming that a counterfactual of that kind must hold, that is, we presuppose that it is proper to a genuine desire (or a genuine belief) that it would have produced different physical movements if the circumstances would have partly changed. We may refer to this counterfactual as '*the tracking counterfactual*', since it comes to express the fact that mental states are individuated in such a way that they track the world. Thus, Charles' perceptual contents are assumed to track, say, certain variations in his position with regard to the bookshelf, while the intentional object of Charles' desire is individuated allowing for variations in the particular way it can been fulfilled and presuming that, in combination with the relevant perceptual contents, Charles will track some modifications in the way in which it can be accomplished. It trivially follows that the particular way in which Charles did moved towards the bookshelf was not counterfactually linked to his wish to read *Ulysses,* and that this is a constitutive feature of his behavior being causally explained by a certain desire of his.

The problem for the causal physicalist is that it is hard to understand how the existence of an implementing mechanism for fine-grained mental contents could be consistent with the fact that those contents are individuated on the assumption that they track the world. For such putative neurological state ought to satisfy what looks like two inconsistent demands. For, on the one hand, that state could not be identified as the detailed neurological state that could causally explain Charles' behavior in the fully determinate way in which it occurred, since, in this case, that neurological state would not sufficiently disregard the details, so that it could account for the fact that Charles' desire would still have moved his body in a very different way if the circumstances had partly varied. But, on the other hand, the neurological state in question should be detailed enough to fix the relevant outcome (say, that Charles picks up his copy of *Ulysses*), in contrast with some other effects (like going towards his desk, or lying on the sofa). It is hard to see how a neurophysiological state could fulfill these two requisites, how it could both (i) have the degree of dissociation with the details that the tracking counterfactual requires and (ii) be specific enough to causally account for the fact that Charles has picked up his copy of *Ulysses,* instead of doing something else. To see this, recall that, so far, the only available story about how (ii) could be satisfied is by providing a very detailed

explanation about how the exact position of Charles' limbs is fixed by a certain antecedent neural condition. The difficulty is, however, that this way of satisfying (ii) conflicts with the need to disregard the details that point (i) imposes.

There is, however, a more general reason why (i) and (ii) cannot be simultaneously satisfied. Suppose, by *reductio ad absurdum*, that there is an implementing mechanism of Charles' present belief B. Hence, that implementing mechanism must fix the huge list of disparate physical dispositions that, according to the tracking counterfactual, compose B's causal powers. Dispositions that concern the combination of this belief with an indefinite range of desires and other beliefs, as well as their corresponding implementing mechanism. It is now easy to show, though, that this demand is inconsistent with a distinction that sounds constitutive of the very notion of 'implementing mechanism'. Trivially, the idea of 'implementing mechanism' involves the possibility of multiple realization, that is, the possibility that a higher-level property were implemented by different mechanisms on distinct occasions. As a result, we may claim that, if B* is to count as an implementing mechanism of Charles' present belief, then it must be possible to delimit some circumstances where B could be implemented by B*, and some other situations were it should be implemented differently. The problem is that, in the case of fine-grained mental contents, this distinction cannot be drawn. For the list of dispositions that B* is supposed to account for, may be enlarged to include any type of counterfactual situation where Charles' behavior would be explained by his belief that his copy of *Ulysses* is on the bookshelf. So, if the causal physicalist should insist that all the dispositions on the list are fixed by the actual implementing mechanism B* of Charles' actual belief that his copy of *Ulysses* is on the bookshelf, she would have to accept that the current implementing mechanism should account for Charles' behavior under any kind of circumstance where Charles' behavior could be explained by that belief. It follows, then, that there is no way to individuate some types of circumstances where Charles' actual belief could not be implemented by the current mechanism. Consequently, the causal physicalist would have to accept that the same implementing mechanism would have to be involved in any counterfactual situation in which Charles might have had the belief that his copy of *Ulysses* is on the bookshelf.

To avert this distressing outcome, the causal physicalist may be now inclined to admit that the mechanism B* that implements Charles' present belief B is not supposed to fix the totality of the list of dispositions that compose B's causal powers, but only a proper sub-set of them. Implementing mechanism B* would then be individuated by its ability to

account for a certain portion of those physical dispositions. The problem is that the causal physicalist may have been deprived of any principled reason to privilege some of these dispositions over the rest. This is surely not a worry that may affect standard functional properties. For, in that case, implementing mechanisms have criteria of identity that are independent of the fact that they implement some instances of a functional property. The problem is that such independent criteria are unavailable for B*, since, as we have seen, the only metaphysical principle to which the causal physicalist may appeal to pick up B* is the fact that it accounts for some of the dispositions that Charles has, here and now, in virtue his belief that his copy of *Ulysses* is on the bookshelf. It is clear, though, that Charles presently instantiates all the listed dispositions in virtue of having belief B and, hence, it seems that any demarcation criterion that the causal physicalist might propose would be fully arbitrary from a metaphysical viewpoint.

We can then conclude that the conditions under which the implementing mechanism of a higher-level property may be individuated, are undercut in the case of fine-grained mental contents. This is so because in the latter case there is no means to distinguish between the causal powers of a certain fine-grained content, and the causal powers of a particular instance of it. And the possibility of drawing that distinction is constitutive of the notion of 'multiple realization' and, consequently, of that of 'implementing mechanism'.

We should say, to close, that this line of reasoning does not preclude that, on some occasions, a brain state may appear as the cause or as the implementing mechanism of a given mental state. In fact, there are overall mental states for which a neurophysiological explanation should be expected. Think of the ingestion of lithium and its ability to improve depressive pathologies. Yet, even if we are glad to acknowledge that the lack of a suitable level of lithium may causally explain the proliferation of depressive thoughts, we are not thereby assuming a more refined discrimination in the values of lithium parameter could be counterfactually connected to the exact content of those depressive thoughts. This, indeed, the kind of transition that we have been trying to block in the present paper: even if there may be implementing mechanisms that account for the tokening of certain quite general mental states, this cannot be so for fine-grained mental contents. For the latter, and not the former, are individuated in such a way that they track the world.

*Josep E. Corbí, Departament de Metafísica i Teoria del Coneixement, Universitat de València, València, Spain 46080; josep.corbi@uv.es*

*Josep L. Prades, Department de Filologia i Filosofia, Universitat de Girona, Girona, Spain 17071; prades@skywalker.udg.es*

## NOTES

1. For a defense of causal physicalism, see Ned Block, "Advertisement for a Semantics of Psychology," *Midwest Studies in Philosophy*, vol. 10, ed. P. French, T. Uehling, Jr., and H. Wettstein (Minneapolis: University of Minnesota Press, 1986); Patricia M. Churchland, *A Neuro-computational Perspective: The Nature of Mind and the Structure of Science* (Cambridge, Mass.: MIT Press, 1989); Paul S. Churchland, *Neurophilosophy: Towards a Unified Theory of Mind/Brain* (Cambridge, Mass.: MIT Press, 1986); Fred Dretske, *Knowledge and the Flow of Information* (Oxford: Basil Blackwell, 1981); idem, *Explaining Behavior: Reasons in a World of Causes* (Cambridge, Mass.: MIT Press, 1989); Jerry Fodor, *The Language of Thought* (Hassocks: The Harvester Press, 1975); idem, *Psychosemantics* (Cambridge, Mass.: MIT Press, 1987); Jaegwon Kim, "Concepts of Supervenience," reprinted in *Supervenience and Mind: Selected Philosophical Essays* (Cambridge: Cambridge University Press, 1993); idem, "Mechanism, Purpose, and Explanatory Exclusion," reprinted in *Supervenience and Mind: Selected Philosophical Essays* (Cambridge: Cambridge University Press, 1993); idem, "Explanatory Exclusion and the Problem of Mental Causation," reprinted in *Information, Semantics, and Epistemology*, ed. E. Villanueva (Oxford: Basil Blackwell, 1990); Brian Loar, *Mind and Meaning* (Cambridge: Cambridge University Press, 1981); and Ruth G. Millikan, *Language, Thought and Other Biological Categories* (Cambridge, Mass.: MIT Press, 1984); idem, *White Queen Psychology and Other Essays for Alice* (Cambridge, Mass.: MIT Press, 1993). For a map of the most relevant stances concerning this issue, see Fodor, "Fodor's Guide to Mental Representation," *Mind* 94 (1985): 79–100; and William Lyons, "Intentionality and Modern Philosophical Psychology, I: The Modern Reduction of Intentionality," *Philosophical Psychology* 3 (1990): 247–69; idem, "Intentionality and Modern Philosophical Psychology, II: The Return to Representation," *Philosophical Psychology* 4 (1990): 83–102.

2. Cf., for instance, the various works by Kim, op. cit., above.

3. Kim, "Mechanism, Purpose, and Explanatory Exclusion," op. cit., 92.

4. Peacocke, "Externalist Explanation," *Proceedings of the Aristotelian Society* 92 (1993): 203–30, has forcefully argued that this phenomenon is the crucial phenomenon on which the externalist character of psychological explanation depends.