VII*—THE SUPERVENIENCE OF MENTAL CONTENT

by Manuel García-Carpintero

It seems pretty obvious to many people that mental content cannot be a relational property; still more if the relational property is some variety of causation, however hedged. This is put sometimes by claiming that mental content is an 'intrinsic' property. But expressing it in this way does not give us a clue about the arguments that are supposed to back the contention, for relational properties could be 'intrinsic' in the sense of being *essential*—at least in some deflationary senses of 'essential': thus, it could follow from the concept of mental content, as manifested by some explicit definition, that mental content is relational. To express more properly the intuition, it was formerly customary to resort to epistemic considerations: we could not be as certain as we are regarding our own instantiating content properties if those were relational. Nowadays, the conceptual basis for claims like this one has be rendered so shaky that even the philosophers closest to the views that used to be sustained by those claims would rather rely on more direct metaphysical considerations. Like this: mental content cannot be relational, for two individuals who share content properties might still differ in relational properties.

Of course, unless supported by further considerations, an assertion like this would just be begging the question against, say, the defender of that variety of functionalism which holds that its being in certain causal relations is *analytic* for a state to instantiate content properties. Nowadays very popular considerations to back the claim that two individuals could share content properties while still differing in relational properties ('Twin Earth' considerations) are based on a famous thought-experiment. I will present them in a more detailed way later. The general idea goes like this. (i) Content properties are supposed to be *causally efficacious*; they play an

*Meeting of the Aristotelian Society, held in the Senior Common Room, Birkbeck College, London, on Monday, 31st January, 1994 at 8.15 p.m.

essential role in causal explanations. (ii) For a macro- property to be causally efficacious in a particular causal transaction, however, it must *strongly supervene* on microphysical properties accounting for that particular causal transaction.[1] (iii) But relational properties do not strongly supervene on microphysical properties; this is what the 'Twin-Earth' thought-experiment allegedly shows. Therefore, there must be intrinsic, non-relational, content properties, for there to be genuine causal explanations which appeal to content properties. Content properties which are causally explanatory cannot be relational.

People who accept this argument do not always reject a sense of 'mental content' such that mental content is relational. They even give it a name, 'broad content' or 'social content'. This is wise, for, curiously enough, the 'Twin Earth' considerations were brought into the philosophical scene in the first place precisely to establish that content, as ordinarily understood, *is* relational.[2] What they claim is that broad content cannot play a role in psychological causal-explanatory generalizations; and, therefore, that there must be a non-relational variety of content which does. (Unless, of course, there are not genuine psychological generalizations, a conclusion that other philosophers have drawn sometimes from similar arguments.) Hence, the philosopher better known for having defended these views, Jerry Fodor, introduces what, following Putnam, he calls 'narrow content': narrow content is just content which supervenes on intrinsic microphysical properties and therefore can play a role in psychological laws.[3] The concession is not to

1 For the concept of *strong supervenience*, see Kim, 1984b. The properties in family M are said to be *strongly supervenient* on the properties in family P iff any event of type M_i, for some *i*, is also an event of type P_j, for some *j*, and, necessarily, any event of type P_j is also an event of type M_i. There is an important and largely undiscussed problem related to how the 'necessarily' in the definition should be understood. It will loom large subsequently in my discussion.

2 See Putnam, 1975. Putnam relied on the principle that two content-vehicles (utterances of sentences, occurrences of mental states or parts thereof) which have a different extension (truth-value, reference) cannot have the same content, and invoked Twin-Earth considerations to contend that 'meanings ain't in the head.' For two intrinsically identical people living in different environments could use the same content-vehicle in such a way that (our intuitions say) the extension it has in each use is nonetheless different.

3 See, for instance, Fodor 1987, ch. 2. I must say that Fodor has changed his views; in a seminar he gave in Ciudad de Mexico in August 1992, he put forward a view similar to the one I myself will defend later. My own views having developed independently of his (I had the opportunity to defend them in that same seminar, finding unexpected

be gladly taken by believers in broad content, though, for it is not at all clear what role a content should play which has nothing to do with explaining behaviour. Several attempts to find such a role should not be considered successful,[4] I believe, although I will not argue for this here.

Narrow contents are thus those allegedly intentional but individualistically individuated aspects of mental states which are causally explanatory of the production of behaviour. More specifically they are what Ned Block calls 'short-armed inferential roles';[5] or even the phenomenal contents of the old empiricist tradition. 'Narrow content' thus is in fact 'non-broad' content, content individuated independently of objects, properties, natural kinds and causal relations going on in the 'external,' actual world. It is notorious that neither of those notions has been made perspicuous; as a matter of fact, I am convinced by Robert Stalnaker's arguments that Fodor's 'narrow contents' lack any acceptable intelligibility.[6] I think that the empiricist, today out-of-fashion, variety of narrow content at least outruns the more fashionable ones in this respect. This will not bother me, however, because I think that we have not been given any acceptable reason to have recourse to any such notion in our explanatory undertakings. Contrary to many externalists, who have tried to rebut the argument by rejecting one of the first two premises,[7] I will try to argue for this while accepting them. Particularly, I accept the supervenience constraint, in the strong form I have put it above. What I will try to show is that the Twin Earth considerations do not establish what they

agreement), there are, as far as I can tell, considerable differences of detail. Firstly, I rely heavily on a teleological account of content. Secondly, I totally disagree with Fodor's extreme atomism: content depends not only on causal connections with the external world but also on internal connections. However, those points affect only the plausibility of the view, not the substance of the rejection of the Twin-Earth argument.

4 See, for instance, McGinn 1982 for such an attempt.

5 Block 1986, 636.

6 See Stalnaker, 1989, 1990 and 1991.

7 Burge 1986 tries several lines of argument, but one of them is a distinction between *individuation* and *causation* which I do not think of any help in this case. Dretske 1988 tries to make do with a different kind of causation. Elsewhere I have tried to show that his answer is the same I will be offering, phrased in a different way. And, of course, many people have tried several weakenings of the supervenience requirement. I think that Kim 1989a rightly shows that global supervenience is not enough for claims of causal efficacy as to macro-properties, in general. Quite irrelevant (for causal-explanatory purposes) characteristics of an event could globally supervene on its physical properties.

purport to establish, namely, that content properties relationally individuated do not strongly supervene on micro-physical properties. For all that we know, it is reasonable to ascribe causal efficacy to broad contents, even when we accept the supervenience constraint in its strongest form.

Let me present first the argument for narrow content in a more detailed form. We assume, for *reduction* purposes, that (mental) content is externally individuated; for instance, that one cannot have a thought about water unless one is somehow related to water, the stuff present in more or less pure form in the lakes, rivers and oceans of our environment. That is to say, for me to have the belief that there is a glass of water in front of me and the desire of water, I must be, leaving aside some complications, causally connected with water, that real stuff. Now, the thought that that belief and desire are causally efficacious in the explanation of why my arm moves as it does now reaching for the glass is part of our every-day intuition that mental properties are causally productive. Of course, we know that there is also a neurophysiological story to tell explaining the same event, and also a microphysical one; and perhaps even a computational one, 'above' all these and 'beneath' the folk-psychological story. But we still think that the folk-psychological account is as valid as any of those (and more relevant).

It will be useful to take the opportunity to point out that the argument I will be discussing should not be confused with an argument to the effect that only the more basic, microphysical properties of events are 'truly' and 'seriously' causally efficacious. There are important considerations to this effect (mostly the 'explanatory exclusion' considerations forcefully defended among others by Jaegwon Kim),[8] and they deserve careful philosophical examination. I myself do not find them ultimately convincing; but, be it as it may, they do not constitute an argument *for narrow content*. Were these considerations accepted, they would ruin the causal efficacy of any non-basic non-reducible property, and so that of narrow content together with broad content. (Assuming, as the three types of defenders of narrow content I mentioned before would undoubtedly assume, that narrow content is not reducible to

8 See, for instance, Kim 1989a and 1989b.

microphysical properties.) It is important to keep this in mind, for sometimes the 'Twin Earth' argument is mixed up with the 'explanatory exclusion' considerations. An example is provided by Jerry Fodor's famous argument in 'Methodological Solipsism as a Research Strategy for Cognitive Science,' based on the combined facts that scientific psychology is computational and that computational processes 'do not care' about the broad meanings of the representations they operate upon. This line of argument leads to the 'explanatory exclusion' consideration; for computational processes are scientifically taken as realized by physical process which 'do not care' either about computational, largely functional properties.[9]

Let us then set aside this other argument, and consider only the effects of the Twin-Earth thought-experiment—which are worth discussing by themselves. The usual setting for the argument is as follows. It seems that there could be a planet very much like Earth, with rivers, lakes and oceans, and someone very much like me (in fact, my molecule-by-molecule twin) having in the equivalent of this very moment beliefs and desires very much like mine, who, being as he is in front of a glass of what seems to be water, behaves exactly as I do. The difference is, there is no water on this planet; the stuff filling the lakes and running through Twin Earth rivers is not H_2O, but something else, something that a knowledgeable chemist would be able to tell apart from water but cannot be told apart with the bare eye by the layman.

The point of the thought-experiment is that, assuming the externalist view of content espoused before, my doppelgänger does not in fact have mental states with the same content as I do. He cannot therefore have beliefs or desires about water, for he is not adequately related to water, but to something else. Nonetheless, he behaves exactly as I do. Therefore, mental states externally individuated do not causally affect the movement of my arm; for *I would have done the same, even if I had not had the mental states I actually have*. The thought-experiment apparently establishes that mental content broadly individuated does not supervene on microphysical condition; for I share the microphysical state with my doppelgänger, while differing in relationally individuated mental content.

9 The same confusion of the two kinds of arguments occurs in McGinn 1982.

There is an immediate refuge that almost everybody seems inclined to take when thinking about this argument. It is philosophically developed and defended by Daniel Dennett and by Brian Loar.[10] I think it is important to stop for a moment to notice that, in a way, it offers no relief, and, in another, it already concedes the point of the argument. The idea (which comes even more naturally in thinking of Tyler Burge's 'social' Twin Earth cases) is to distinguish, in Loar's terms, 'social' from 'psychological' content. The 'social' content attributed by using 'water' could indeed depend on external and environmental matters of interest to the community using that term, while the 'psychological' content would depend more closely on how the particular individual 'sees' the world. The 'psychological' content, in the present case, would have to do with the characteristics that the layman uses to recognize water. Under the conditions of the thought-experiment, it seems plausible to say that the 'psychological' content is the same for both me and my doppelgänger (something to the effect that there is, in front of me, part of the colourless, odourless and tasteless stuff present in the lakes, rivers and oceans of the planet I inhabit which quenches one's thirst), therefore the conclusion does not follow; for all that the thought-experiment establishes, 'psychological' content seems to be in good shape to satisfy a claim of causal efficacy.

This, however, plausible as it may sound, is, as I said, either a straightforward acknowledgement of defeat or totally unsuccessful. What we do is simply transfer the problem to the terms we use in the specification of the 'psychological' content. These will contain for the most part 'observational' notions. Now, if we think of their individuation in phenomenalistic terms, we have simple reenacted once again the empiricist project, the analyzability of every content in phenomenal terms. There is no more reason today to think that we will succeed were they failed, and, anyway, this is to concede the point that content must be narrow to be causally efficacious, in the most unpleasant form. On the other hand, if we individuate the content of observational notions externally, we will be exposed to exactly the same Twin-Earth type of argument. It will be useful to verify this point now, since my own reply to the argument will be

10 In Dennett, 1982, and Loar, 1987.

easier to accept when examined first regarding externally individuated observational contents, and only then extended to non-observational concepts.

The rebuttal of the Twin Earth argument will depend crucially on taking seriously some externalist account of content, therefore I will quickly summarize the one I take to be closer to the truth. My own variety of externalist account of content is teleo-functional. To apply a teleo-functional account to an observational concept, for instance the concept (or 'percept') of a colour or sound, we need first a characterization of the concept's content, the colour or sound themselves. Take the colour *red*, and think of our concept of it. For familiar reasons, no simple, non-disjunctive physical characterization of the property would do as the content we are looking for, and neither will physical dispositions like *reflectance*, the disposition of surfaces to reflect given percentages of the light in each wavelength. No such property would do justice to the way we classify objects as having the same or different colours, and even less to the relations we establish among colours in terms of saturation, brightness and hue.[11] We must define observational properties, I believe, as dispositions to produce given *actual* outputs in certain mechanisms. Of course, the mechanisms in question will be in the last resort the ones implementing the concepts we are about to characterize. There is therefore interdefinition here, which means that in any given circumstance we will have to characterize in the same breath both the concept and its content, but not vicious circularity.

Now, the disposition corresponding to *red* is an objective property, a property that surfaces would have had even if there had not been any organism implementing the mechanism by means of whose outputs we have characterized the property. The same property could in principle be described in purely physical terms, as a matter of fact as a disjunction of several physical properties, or as a disjunction of more abstract physical properties like reflectance. Relative to this objective albeit dispositional property, we specify the concept whose content is constituted by that property as a structure which has the function of detecting that property. We

11 See Hardin 1988 for a clear summary of these facts.

understand 'function' in a way akin to the biological understanding of the term; the familiar idea here is that it is the performance of the function by the structure in the past that helps to explain the proliferation or simply the preservation of structures like that. The fact that the performance of the function is 'good' for the organism in which it is placed does not explain why it appeared in the first place (that could be a matter of chance), but it explains why it was preserved or/and reproduced. This explanation could take the form of a typical natural selection explanation, or, more to the point in the present case, that of a not so well understood acquisition by a learning process.[12] Finally, in speaking of a function of 'detecting' the property, what we mean in this very simple case is that the structure, when everything goes well, is a causal factor in the performance of behaviour successful for the organism precisely because the property is actually instantiated in the environment.

To put it simply, for the sake of avoiding all unnecessary complexity (complexity which, on the other hand, should not be forgotten, if the account is not to be unbelievably far from the facts): the colour *red* is one among a set of physical properties, all of which have in common the disposition to produce in normal conditions the same result in a certain mechanism; and the *concept red* is a structure which is kept because it causes, in normal conditions, behaviour of the organism in which it is placed 'sensitive' to those properties, behaviour that satisfies the needs of the organism in part because those properties are instantiated in its environment.

There is a case in which we can almost fill in all the details of this account, provided by the amazing research of Konishi and his colleagues on the auditory system of the barn owl.[13] We have here, first, behavioural evidence that the owl is able to identify sonorous objects and track them in three-dimensional space. This justifies the conjectural attribution of content, in the former sense: the presence of a structure which has the function of producing behaviour sensitive to certain properties instantiated in the environment. At this stage we can only guess what the external properties are. Then we have a functional (in the non-teleological sense of the word) account

12 See Wright 1973.

13 Conveniently summarized in Konishi et al. 1988.

of how that feat is accomplished, which involves a more precise elaboration of both the content and the concept; and, finally, a carefully worked out neurophysiological filling out of the functional model. This is what is required to attribute to the owl states with content, content answerable to the spatial location of sounds of given categories (to be specified relative to the outcomes of the mechanism).

This account of the content of observational concepts is externalist. Precisely because it is externalist, it does not have the problems besetting traditional phenomenalist accounts. For instance, we can render intelligible the normative dimension of content attributions; and we do not have any 'other minds' problem. I will not dwell on these aspects here. My strategy in this paper is not to defend the view that our intentional concepts should be understood according to the teleological explanation, but only the weaker claim that intentional concepts individuated according to a teleological explanation are not liable to the Twin Earth arguments—even if at first sight they seem to be so liable.

A typical way to show that observational concepts, teleologically understood, are liable to the Twin Earth considerations against the causal efficacy of content, relationally individuated, is to invoke 'inverted spectrum' cases in Twin Earth settings. As I stressed before, the content of a teleologically individuated observational concept is constituted by whatever physical properties *actually* cause the specified output in the indicated mechanism. Now, suppose we claim that my perceptual belief that there is something red before me (the radiant upper light in a traffic light) is, in virtue of having that intentional property, causally efficacious in the production of my body movement (my foot pressing the brake); and here we think of the intentional property as individuated according to the teleological account. It seems then easy to think of someone like me in all relevant respects (my physical or phenomenal twin), being in the same internal state I am, and therefore producing the same behaviour, who nonetheless is in a state whose function it is to produce behaviour appropriate to *green* objects (because, say, during the learning period in which the function was acquired he had 'colour inverting lenses' inserted in his eyes).

Actually, the detailed working out of the idea happens to be tricky. Inverted spectrum cases were designed to illustrate the

'converse' cases from the ones we need here, cases of people being in the same broadly individuated state but different physical or phenomenal states. When the two subjects in those examples, the 'normal' and the one with 'inverting lenses' are in the same phenomenal/physical state (when they both have a red *quale*) they are in states with different broad contents, but, typically, they also behave in different ways (one utters 'red', the other 'green'; one brakes, the other accelerates). To run a Twin Earth case, we need also similar behaviours. Ned Block has proposed a way to conceive of a case like that, having resort to an 'Inverted Earth.'[14]

> Inverted Earth differs from Earth in two respects. Firstly, everything has the complementary colour of the colour on Earth. The sky is yellow, grass is red, fire hydrants are green, etc. I mean everything **really** has these oddball colours. If you visited Inverted Earth along with a team of scientists from your university, you would all agree that on this planet the sky is yellow, grass is red, etc. Secondly, the vocabulary of the residents of Inverted Earth is also inverted: If you ask what colour the(yellow) sky is, they (truthfully) say 'Blue!'. If you ask what colour the (red) grass is, they say 'Green.' [...] Further, the intentional contents of attitudes and experiences of Inverted Earth are also inverted.[15]

This is indeed so, according to a teleological account of intentional content. Before proceeding, though, I would like to point out that the case is, in all probability, wrongly described. For the stuff filling the lakes, rivers and oceans in our planet to have a different colour (without changing the laws of nature) it should have a different chemical composition. That means, according to widely held Kripkean intuitions, that the stuff in question would not be water, but something else. The same with grass, and probably with 'the sky.' Therefore, what Block is trying to describe is something similar to Earth in some of the observational properties, but very different in everything else; which means that 'Inverted Earth' does not seem to be a nomically possible world. Block could also claim directly that there is water, H_20, in Inverted Earth, but that the laws are there very different, which explains that the colour it produces is different. This sounds intelligible, but I do not think it is, for I do

14 See Block 1990a.

15 Block 1990a, p. 62.

not understand very well what it means to say that the natural kinds could be the same although the laws they obey are different.

None of this matters very much to Block's goal in the paper I am discussing (which is merely to prove that inverted spectrum cases are conceptually possible, or conceivable, which I am ready to concede right away and will have nothing to do with my main point), but it will be seen to be very relevant to mine.

After describing Inverted Earth, Block tells an inverted spectrum story relative to it: just imagine my twin, raised in Inverted Earth with colour inverting lenses inserted in his eyes. It seems as if, when he is looking at the Inverted Earth sky, he is in a state with a broadly individuated content different from mine, when I am looking at the Earth sky; physically and phenomenally, though, he is in the same state I am, and our behaviours are likewise the same. Now, take again the claim that my perceptual belief that there is something red before me (the radiant upper light in a traffic light) is, in virtue of having that intentional property, causally efficacious in the production of my body movement (my foot is pressing the brake). By considering Inverted Earth, however, it seems plausible to contend that *I could have been in a state with a very different broadly individuated intentional content* (i.e., with the content that there is something *green* before me) *while carrying out the same braking behaviour.* The causal efficacy of the intentional content of my state, thus, seems to fade away, 'screened off' by its phenomenal, narrow or even physical properties. Once again, mental content, broadly individuated, does not seem to supervene on internal state.

This is why, as I said before, Loar's promising solution is no solution: to the extent that we have an externalist account of the observational concepts, the Twin Earth considerations apply to them too. And I should add the following: Once you see why the Twin Earth argument is unsound with respect to the observational concepts, you will see that it is also unsound for any other. The grain of truth in Loar's idea is this: the higher-order the concept, the easier it is to be misled by the Twin Earth considerations; the easier to conceive of the mental property as 'free-floating' over its physical basis. It is good strategy then to rebut the argument with respect to observational concepts, and only then to extend the rebuttal to the more theoretical and social ones.

There is another faulty attempt to parry the Twin Earth considerations that I have heard sometimes, whose examination will take us to the heart of the matter. In a nutshell, this reply has it that the Twin-Earth considerations do not prove that, say, water thoughts are causally inefficacious, but only that the behaviour caused by them could have had different causes (i.e., twater thoughts). The argument would then be ineffectual, for in almost any case of full-fledged causation the effect could have been brought about by different causes: it would not establish what it purports, namely, the causal inefficacy of content.

But this is to miss entirely the point of the argument. Consider this example. The patient claims that his depression has been caused by his feeling miserable, his girlfriend having left him; the doctor objects that that is not 'the' cause, that the real cause was his hyperthyroidism, that *even if his girlfriend had not forsaken him, he would still have had the depression.* Obviously, it would be a mistake to point out here that the doctor has not proven the patient wrong, that he has only shown that depressions (*types*) can be caused by hyperthyroidisms (*types*), leaving intact the fact that depressions are also caused by states of feeling forlorn. The doctor's point concerns *this* particular process; he is claiming in effect that *this same hyperthyroidism process* could have happened without a *state of feeling forlorn* also happening, and it would still have produced the same depression. The doctor's point is that the alleged causal character of the mental property is in this case successfully 'screened off' by physical properties.

This is then the heart of the matter: the Twin Earth arguer contends that his thought-experiments prove that the externally individuated intentional properties of the individual states are rendered causally inefficacious by being screened off by their narrow content properties, whatever they are. And let me remind you that this point is not to be confused with the 'explanatory exclusion' consideration, to the effect that the alleged causal efficacy of *every* non-basic property is successfully screened off by microphysical properties.

Here is a taxonomy of common examples in which an allegedly causally efficacious property is screened off by another:

(i) **No property**: (The inefficacy of astrological and other superstitious 'properties.') Although every predicate constructed according to the rules of the language has a sense, or expresses a

concept, not every predicate express a property. Goodman's 'grue' is a case in point. A good first stab at an analysis of the difference is the idea that to express a property, a predicate should appear in the formulation of true laws of nature. A reason for denying causal efficacy to an alleged property expressed by a predicate applying to the cause-event is that the predicate in question does not even express a property. My eating of the muffin, this particular event, satisfies the predicate 'being the eating of a muffin bought in *Safeway*', but it cannot be in virtue of this aspect that this event causes my stomach-ache. The causal efficacy of alleged properties expressed by predicates like 'flying on a Friday the 13th', 'being born under the influence of Mars' is denied on similar grounds.

(ii) **Nomically related effects of a common cause**, or **epiphenomena, old variety**: 'Drinking coffee in such-and-such an amount' probably expresses a causally efficacious property; but it is not in virtue of instantiating this property that a protracted 'event' causes a lung-cancer. The 'real' cause is another property of 'the same' event, namely, its being a case of heavy smoking; both properties happen to be nomically correlated, by being effects of a common cause. As a matter of fact, this knowledge would suffice to consider the description before as relying on a much too coarse characterization of the facts, and to separate 'the' event in, at least, two different ones, an event or events of drinking coffee and an event or events of smoking. In the last century, mental events were said to be epiphenomenal in this sense: they were thought of as separate effects of the real causes of behaviour. The mental properties were still causally efficacious; at least, they were effect-properties in laws of nature, even if they were never cause-properties.

(iii) **Constitutively related nonbasic properties**, or **epiphenomena, new variety**: In Dretske's inspired example, the property the soprano's singing has of being the singing of a word meaning *Help!* is perhaps causally efficacious for some effects, but it is not causally efficacious in the shattering of the glass; only the intensity and the pitch of the sound are relevant to that causal explanation. Block mentions an intriguing example.[16] The Wiedemann–Franz Law links thermal and electrical conductivity;

16 Block, 1990b, p. 147.

however, it is the thermal conductivity of a rod connecting a fire to a bomb which causally explains the explosion, and not the electrical conductivity of the rod. In those cases, we cannot 'split' the cause -event, for its having the efficacious property is constituted by the same facts as its having the inefficacious one. We cannot drive a spatial, or at least temporal, wedge between two events, perhaps two effects of a common cause. Mental properties are said nowadays to be epiphenomenal in this sense, for we do not want to leave any room for a mental 'substance.'

In all these cases, we express our disbelief in the efficacy of the allegedly causally efficacious property by resorting to counterfactuals similar to the ones that the Twin Earth thought experiment seems to force us to assert, for example that even if I had not had a belief that there is something red before me, I would have put the brakes on anyway. Similarly: even if you had not been born under the influence of Mars, you would still have been as aggressive as you are; even if you had not drunk any coffee at all, you would still have had the lung cancer; even if the song had not been a 'help!' cry, it would still have shattered the glass. In these latter cases, the truth of the counterfactual expresses the failure of strong supervenience of the macro-property on intrinsic microphysical aspects.

But there is a fundamental difference between the basis for the last three counterfactuals and the basis provided by the Twin Earth thought experiment for the first; and here we reach the crux of the issue. The difference is, to put it bluntly, that *in none of these cases is only a thought-experiment what lies behind the counterfactuals.* In all those cases there are *a posteriori*, scientific reasons to believe that *there are not explanatorily relevant links between the more basic causally efficacious property of the event that could be mentioned, and the alleged causally efficacious properties.*

This is what is missing in the Twin Earth cases, and, without it, they cannot achieve their goals. Assuming that some non-basic properties are causally efficacious, a sheer thought experiment should not be considered sufficient to establish that a non-basic property is inefficacious. If it were, we could cheaply prove the causal inefficacy of any non-basic property; the only requirement would be some imagination, together with the slack between the concept of the macro-property and those of the relevant basic

properties. Any causal claim about the efficacy of the temperature of a body, say, could be rendered false, just by thinking of a possible world in which temperature is not constituted by the mean kinetic energy of the particles, and so the body is microphysically as it is now without having any temperature; in Kripkean terms, the thought of an epistemic possibility that as a matter of fact is not a metaphysical possibility would be enough.

Consider the owl example above, and take a claim that the present movement of the owl is causally explained by its having a (teleologically individuated) belief about the location of a given type of sound. Of course, we could think, with some imagination, of 'Inverted Earth' cases such that the owl is physically, or neurologically, in the same state as it is now, while states of that kind lack the function of detecting the given type of sound. (They could have the function of detecting another sound, or no function at all.) But do these cases show that the teleological property is causally inefficacious? Of course not. The 'possible worlds' in question, we can claim with certain knowledge in this particular case, are just epistemic possibilities, allowed by the slack between the teleological and the neurological concepts. But they are not nomically possible: the neurological state could not have produced its 'inverted' contents, given its physical nature. Those worlds are not even metaphysically possible, on familiar Kripkean grounds which I cannot dwell on here: Once we know that *this* teleological state is *this* neurological state, we can claim that they are the same across possible worlds; for, after all, the causal powers of the teleological state are accounted for by its being a given neurological state, and an event is essentially something which has certain causal powers.

The point, simply put, is this: assume that observational concepts are teleologically individuated, as previously explained. Assume that there is an explanatory link between a brain structure having a particular physical, or neurological constitution, and its having one of those functions; namely, assume that we can *explain* the preservation and proliferation of the structure, by learning or natural selection, on the basis of its performing the function of contributing to the production of behaviour sensitive to the presence of the observational property. In those circumstances, the fact that we could think of fancy cases in which the physically individuated

structure lacks the function is entirely irrelevant to any claim about the causal efficacy of its having that function; in particular, it does not establish that its alleged causal efficacy is 'screened off' by its physical properties.

But perhaps, it could be argued, it is even physically possible that—by being struck by lightning, say—this tree should become a physical doppelgänger of the owl, as it is when we explain its behaviour by attributing to it the perception of given sounds in three-dimensional space. Fortunately enough, we do not need to get embroiled in an argument about what is physically possible to dispose of this reply. This is my answer. Macro-properties, like functions, typically take part in *ceteris paribus* laws. For the right sort of (nomic) supervenience to hold, then, we do not need a (counterfactually supporting) exceptionless claim: whenever something is in the same physical state as the owl is now, it has the functional property we attribute to it with causal-explanatory purposes. Any such claim is unrealistic, and is going to run into imaginative counterexamples of the kind envisaged. What we need is just an explanatory relation from the physical state of the owl to its possessing the functional property (the perceiving of the spatially located sounds). This will be a general supervenience law, but it could be as hedged by *ceteris paribus* restrictions as the macro-laws we want to account for in more basic terms already are.

Thus, the only (counterfactually supporting) general truth we need is something like this: whenever something is in such-and-such a microphysical state (the one instantiated by the owl), *and everything else is equal*, it is in a functional state with such-and-such characteristics (namely, a perception of a certain spatially located sound). Now, in what circumstances *ceteris paribus* clauses are justified, and when they are instead used as an *ad hoc* device to uncritically salvage a claim, is a thorny issue which need not concern us. For it seems safe to invoke the point to dispose of the lightning example—assuming, as I have done, that we do have a good explanation in at least neurological terms of the owl's instantiating the functional property.

I will finish by discussing briefly the issue of more complex contents, contents regarding, say, water, John Huston or arthritis. The grain of truth in the Loar-like reply I rejected above is that to have these concepts necessarily requires conceptual links with

other concepts (in the last resort, with observational concepts) of the sort exhibited when we express the Loar-like 'psychological content.' To have water-thoughts requires, roughly, to know that water is the substance responsible for the odourless, tasteless and colourless aspects of *this* stuff—pointing at concrete samples; to have thoughts of John Huston requires, roughly, to know that John Huston is the person referred to by 'John Huston' in *these* utterances—pointing at concrete cases; to have arthritis-thoughts requires, roughly, to know that arthritis is the illness referred to by 'arthritis' in *these* utterances—pointing at concrete cases. I said *roughly*: the exact account of the analysis is a very difficult subject about which I have not entirely made up my mind. And the knowledge in question must be roughly thought of under the 'tacit knowledge' model of, say, Evans 1985; the 'roughly' here qualified as the previous ones. I am just trying to give the essential aspects of the idea.

Now, suppose Manuel qualifies as having the thought that John Huston has arthritis, and we claim that this thought (together with some other mental states, etc.) causally explains some behaviour, say, some linguistic behaviour of his. Assume that his instantiating the conceptual abilities constitutive of the requirement is explained by his instantiating neurological state **N**, which is also a partial physical cause of the sounds he utters, etc. It is obvious that here we are much more in the dark as to how this 'explaining' goes than in the owl's case (or even in that of our own perceptual states). This is why it is even easier here to think of someone instantiating **N** without instantiating the mental property. And it is clear that this is enough for the existence of the *epistemic* possibility of Manuel instantiating **N** without instantiating the thought that John Huston has arthritis. (Just think of Burge's possibility, the linguistic community giving a different sense to 'arthritis'—or to 'John Huston'.) But the question is, is this also a nomic or metaphysical possibility? To answer in the negative, as the Twin-Earth contender does on the basis of a thought-experiment, is simply to beg the question. For the real question regarding the externalist's claim as to the causal efficacy of the thought that John Huston (*that man*) has arthritis (*that illness*) is whether having this thought has a physical explanation. And this is not to be decided by thought-experiments

or other appeals, however masked, to what is plausible. We already knew that externalism sounds very implausible.[17]

Departamento de Lógica, Historia y Filosofía de la Ciencia
Universidad de Barcelona
08028 Barcelona

REFERENCES

Block, Ned, 1986: 'Advertisement for a Semantics for Psychology,' in French, P., Uehling, T., and Wettstein, H., (eds.), *Midwest Studies in Philosophy*, Minneapolis: University of Minnesota Press.

Block, Ned, 1990a: 'Inverted Earth,' in J. Tomberlin (ed.), *Philosophical Perspectives, 4: Action Theory and Philosophy of Mind*, Atascadero, California: Ridgeview Pub. Co.

Block, Ned, 1990b: 'Can the Mind Change the World?', in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*, Cambridge: Cambridge University Press, 1990, pp. 137–170.

Burge, Tyler 196: 'Individualism and Psychology', *Philosophical Review* xcv, 1986, 3–46.

Dennett, Daniel, 1982, 'Beyond Belief,' in Woodfield, Andrew (ed.), *Thought and Object*, Oxford: Oxford University Press.

Dretske, Fred 1988: *Explaining Behaviour*, Cambridge, Mass.: MIT Press.

Fodor, Jerry, 1987: *Psychosemantics*, Cambridge, Mass.: MIT Press, 1987.

Hardin, C. L., 1988: *Colour for Philosophers*, Indianapolis: Hackett Publishing Company.

Kim, Jaegwon, 1989a: 'The Myth of Nonreductive Materialism,' *Proceedings and Addresses of the American Philosophical Association*, 63.3, pp. 31–37, 1989.

Kim, Jaegwon, 1989b: 'Mechanism, Purpose, and Explanatory Exclusion,' in J. Tomberlin (ed.), *Philosophical Perspectives, 3: Philosophy of Mind and Action Theory*, Atascadero, California: Ridgeview Pub. Co., pp. 77–108.

Konishi, Masakazu, et al., 1988: 'Neurophysiological and Anatomical Substrates of Sound Localization in the Owl,' in Edelman, Gall and Cowan (eds.), *Auditory Function, Neurological Bases of Hearing*, New York: J. Wiley & Sons.

Loar, Brian, 1987, 'Social Content and Psychological Content,' in *Contents of Thought: Proceedings of the 1985 Oberlin Colloquium in Philosophy*, R. Grimm and D. Merrill (eds.), Tucson: University of Arizona Press.

McGinn, Colin, 1982: 'The Structure of Content,' in Woodfield, Andrew (ed.), *Thought and Object*, Oxford: Oxford University Press, pp. 207–258.

Stalnaker, Robert, 1989: 'On What is in the Head,' in J. Tomberlin (ed.), *Philosophical Perspectives, 3: Philosophy of Mind and Action Theory*, Atascadero, California: Ridgeview Pub. Co., pp. 287–316.

Stalnaker, Robert, 1990: 'Narrow Content,' in C. A. Anderson and J. Owens (eds.), *Propositional Attitudes. The Role of Content in Logic, Language and Mind*, Stanford: CSLI, 1990, pp. 131–146.

17 The research for this paper was funded by the DGICYT, Spanish Ministry of Education, as part of the research project PB-0701-C03-03. I thank the participants in the seminar on the Philosophy of Mind held at the University of Barcelona in the year 1992–93, especially David Pineda, for their critical remarks and support.

Stalnaker, Robert, 1991: 'How to Do Semantics for the Language of Thought,' in B. Loewer and G. Rey (eds.), *Meaning and Mind. Fodor and his Critics*, Oxford: Basil Blackwell, pp. 229–238.
Wright, Larry (1973): 'Functions,' *Philosophical Review*, LXXXII, 139–168.