The Scientific Nature of the Mind-Body Problem

– Reflections on Supervenience –

BY: PEDRO FONSECA DIRECTOR: DANIEL ANDLER

TABLE OF CONTENTS

— September 2000 —

"*The second constraint I have followed is to* take science seriously. *I have not tried to dispute current scientific theories in domains where they have authority. […] For example, I have not disputed that the physical world is causally closed or that behaviour can be explained in physical terms". (Chalmers, 1996: xiii)*

"*On the other hand, it is clear by now that* all *interpretations of quantum mechanics are to some extent crazy." (Chalmers, 1996: 356)*

SUPERVENIENCE[1], in its most general formulation, is a logical relation between two sets of properties[2] defined as: a set of properties M supervenes on a set of properties P *if and only if* there can be no differences in M without there being differences in P. In a more intuitive fashion, it just means that two indistinguishable states or events at the subvenient level cannot be distinguished at the supervenient level. One standard example is shape: shape supervenes on the microphysical structure of a physical object, since two objects with exactly the same microstructure must have exactly the same shape. It might seem that this is a necessary relation, that is, that it would be a logical truism that two physical objects exactly alike in all its parts would have an equal shape. However, regarding other properties, like shadows or weight, we cannot say the same, since they are not the same for every possible . Consequently, to assert the necessity of a property like weight we would have to consider a system of objects exactly alike. In this way, if we would individuate the objects in two identical systems in the same manner, we would obtain exactly the same weight relations for each object in both systems. The kind of necessity involved (strong or weak, logical or natural), as well as the scope (local or global) of the supervenient relation are strongly dependent on the kind of properties involved. There is no consensus on both of these topics regarding mental and physical properties.[3]

---

[1] Usually it is said that the concept of supervenience appeared first with the emergentists of the twenties but especially in the moral discourse of Moore (in 1922) and Hare (in 1952) who, in particular, introduced the word in the moral philosophy discourse. But Kim argues that "Moore and other emergentists [in the 20's] were the first, as far as I know, to develop a generalized concept of supervenience as a relation, and their concept turn ou to be strikingly similar to that in current use, especially in philosophy of mind." (Kim 1990: 138). However, the concept was forgotten and it was only introduced in contemporary debate by Davidson in 1970.

[2] We will follow the literature and speak always of 'properties' to simplify the speech. But it should be noticed that these properties, when applied to physical objects, are usually understood in terms of states or events. Although, as we will see later, this is perhaps not a good description of what would happen if we really would try to provide a description of a supervenience base of a physical system.

[3] For instance, Chalmers argues that mental properties supervene only naturally on physical properties. While Kim argue that only a strong version of supervenience (which holds across

Our work, however, will not focus on these well known problems. We will try to put in evidence a characteristic of supervenience that has, to our knowledge, passed unnoticed to philosophers. According to our argument it is a fact that supervenience does not apply in any simple manner to future events, unless certain special constraints are assumed. We will also try to sketch some of the consequences of this feature to contemporary debate. We will start by a small introduction, and by elucidating some of the classical problems regarding supervenience and their most common answers, and we will then pass on to the critical part of our work.

1. Mereological supervenience[4] as the minimal commitment to physicalism.

What is an object? If it is merely a collection of parts it must be a logical truism that the same parts (in the same arrangement) must produce the same (an identical) object. If we accept this definition then mereological (or **m-supervenience**) must seem a logical truism, because, as the result of the definition, there would be no logical space for an object to have the same parts but not the same intrinsic (i.e. non-relational) properties.[5] But suppose that we were to accept that some kind of su-

---

worlds) of the mental over the physical could yield an attractive relation of dependence. There is also a further difficulty concerning semantic properties: due to their relational character, they would seem to demand global supervenience, while our assumption that events spatially unrelated (the locality assumption) cannot have causal intercourse would seem to demand local supervenience (that is, we would expect that events that can affect semantic properties of the propositional attitudes of an individual but not the individual physical state, would be unable to change its behaviour, therefore they would be causally inefficacious).

[4] Except when otherwise noted mereological supervenience applies only, in our text, to the relation between physical properties. Note that Kim assumes that mereological supervenience is at the base of 'standard type-physicalism' (see Kim 1994: 582). Our use of mereological supervenience will show to be much closer to Chalmers conception of logical supervenience, that we will discuss later on.

[5] This is what Chalmers (1996) argues to be the case regarding all physical properties. However, as we will see in the next chapter, there are many exceptions to this rule, and it is also not demonstrated that this is in fact the case. Chalmers would call this 'logical' supervenience. Besides considering a notion of logical necessity at odds with contemporary debate (see Chalmers 1996: 52) it as also led him to differ from the usual categories of supervenience drawn by Kim (1994, 1998), and also to redefine the notion of metaphysical (or a posteriori)

pernatural forces act on the world. Let's suppose that Witch Mary, for instance, has the power to produce in certain physical systems the mysterious property W, which can be measured (by some kind of physical apparatus). If we had lived in a supernatural world (with lots of witches and magic spells) we would have to choose a different definition of object. In this case, it would be a collection of parts and something else – the magical properties that witches and gnomes (and other strange looking creatures) would send over them.[6] In this world m-supervenience clearly does not apply, because, for instance, two indistinguishable brooms at the elementary level could exhibit quite different properties (for instance one could levitate). Seen through this perspective, m-supervenience looks like a very strong thesis, it is the thesis that not only witches, ghosts, gnomes, fairies, angels and interacting souls, do not exist (or at least that they cannot alter any intrinsic properties in physical objects), but that everything that exists are jumbles of elementary physical particles (in the measure that these are the only entities which can be included in the subvenient level). Since it is a thesis about the (in)existence of certain objects in the world it cannot, obviously, be considered a logical thesis, since its truth logically depends on what there is.

Now, the straight reaction to this would be: OK, we can see that supervenience is false if dualism is true. But dualism cannot be considered an acceptable choice. Due to

---

necessity. In our text, we will argue that m-supervenience can be considered a logical truism only if we have appropriate constraints on the description of the supervenience base. However, has will become clear from our treatment of natural laws, we think, ultimately, that this description is unattainable.

[6] Notice that it is not clear that this imaginary situation would violate any causal relations between physical events. Although many philosophers have argued that any kind of dualism would necessarily violate physical laws like the principle of the conservation of energy (see for instance Kim (1998: 31) Dennett (1991: 35), Lewis (1988: 513), among many others) it seems clear that all depends on what past physical events entails for future physical events. For if the past entailed only sets of possibilities (imagine that (micro)physical state p would conduce either to flying or non-flying atoms in object B) then magic spells could operate on these sets of possibilities without any violation of the physical laws (conservation of energy, statistical determinism, etc). What it does violate is the causal *closure* of the physical.

massive empirical evidence (and explanatory problems[7]) we can see it's not even a debatable option. So, although we might say that there is a sense in which supervenience is an empirical theory this would be misleading if we would not also add that it is the only theory that fits the data and is reasonable enough to be true. In other words, supervenience has perhaps an empirical character but that doesn't change the fact that it is our only reasonable choice.[8]

1.1. Supervenience as covariance is not sufficient for physicalism.

Notice that this general position might seem to be near to what Kim has called the commitment to minimal physicalism: "the minimum commitment that anyone who calls herself a physicalist should be willing to accept." (Kim, 1998: p.38). According to this position supervenience is a sufficient and necessary condition of physicalism.

> "mind-body supervenience is inconsistent with more extreme forms of dualism, such as, Cartesian dualism, which allow the mental to float freely, unconstrained by the physical domain. Thus mind-body supervenience can serve as a useful dividing line: it can be viewed as defining

[7] There are two main difficulties that a dualistic theory has to face. The first, the most commonly held, is to provide a conceivable account of mind-body interaction. But we think the most important is perhaps its lack of explanatory power. If materialism is true than there is a hope of explaining consciousness scientifically, but if dualism is true, how could we account for consciousness, how could we explain its apparition in the world? If consciousness should prove not to be the result of some physical process its existence would be more enigmatic then the most enigmatic physicalist explanation could show it to be.

[8] This is very close to the position defended by Lewis (1988: 507) [What experience teaches]: "Minimal Materialism is a supervenience thesis: no difference without physical difference. … But we materialists usually think that Materialism is a contingent truth. We grant that there are spooky possible worlds where Materialism is false, but we insist that our actual world isn't one of them. If so, then there might after all be two possibilities that are alike physically but not alike *simpliciter* … . Spooky worlds could differ with respect to their spooks without differing physically." Notice that Lewis does not distinguish between covariance and dependence claims in the concept of supervenience. Notice also, that the acceptance of spooky stuff without causal powers does not obviously denies supervenience (see our discussion on these topics later on).

minimal physicalism."[9] Kim (1998: 15, *italics* in quotations are always from the original texts.)

The difference between our view and the one Kim is expressing, however, is that Kim tries to use a minimal version of supervenience according to which it just asserts a certain form of covariation between mental and physical events. As we will show, this won't do as a definition of physicalism.

> "The notion of supervenience we introduced[10] simply states a pattern of *covariance* between the two families of properties, and such covariances can occur in the absence of a metaphysical dependence or determination relation."[11]

We will argue that, by abandoning the reference to any kind of dependence relation, Kim's definition opens the way to a perspective in which we would end up by being obliged to include under the definition of minimal physicalism views that he

---

[9] Notice that this implies that supervenience is both necessary and *sufficient* for minimal physicalism (it implies minimal physicalism – previous quote– and it is incompatible with perspectives in which physicalism is denied, i.e., in which the mental in 'unconstrained' by the physical – this quote). We will try to show in the next few lines why, if physicalism is taken in its usual formulation, supervenience-as-covariance is not sufficient (and, therefore, why it can't serve as a *definition*) for physicalism.

[10] Kim gave the following definition in: "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." (1998: 9)

[11] Notice that Kim's views have changed a on this respect. In Kim (1984) supervenience is defined in terms of a dependence relation: "The core idea of supervenience as a relation between two families of properties is that the supervenient properties are in some sense *determined by*, or *dependent on*, the properties on which they supervene." (1984: 98) (Although we should notice that the definition that follows in that text is just the classical covariance claim.) It was only on Kim (1990: 140-148) and specially on Kim (1993: 167) that he writes that "it now seems to me a mistake, or at least misleading, to think of supervenience itself as a special and distinctive type of dependence relation". As we will argue later, Kim's change of view is based on the mistaken view that the dependence relation between the two properties is always a metaphysical (and not a logical) claim.

himself calls "substantival dualism" (1998: 29) is accepted. To show this, we may start by noticing that the parallelism espoused by Malebranche and Leibniz exhibit the same type of property covariation (between physical and mental entities) that, for instance, identity theory would exhibit, only the dependence relation would change. In fact, in a previous text, it is Kim himself who remarks that:

> "Even Leibniz and Malebranche had something to say about this [supervenience of the mental over the physical]: the observed property covariation is due not to a direct dependency relation between mind and body but rather to divine plans and interventions." (Kim, 1994: 582).[12]

In fact, the covariation is there, since, due to God's powers, two indiscernible states at the physical level would necessarily correspond to two indiscernible states at the mental level. And so, if supervenience is just supervenience-as-covariation (let's call it **c-supervenience**), there is no way we can deny that Malebranche and Leibniz views accept the c-supervenience of the mental over the physical and, therefore, their views should be counted under (minimal) physicalism! Now it's time to ask if we are still prepared to consider c-supervenience as the 'dividing line' between physicalism and everything else.

In our presentation it seemed that we had good reasons to upheld a similar view, but notice that we've started with *m-supervenience*, which asserts much more than simple covariation. We will try to show that, in fact, what serves as a dividing line between physicalism and everything else is not the covariation thesis between the mental and the physical, but the dependence relation. To see this more clearly we propose to go a little beyond what Kim says and check if a covariation relation in which *the physical depends on the mental* can be conceived. So imagine we have a possible world where what we might call dictatorial idealism[13] applies. Now a world where dictatorial idealism applies is such that every mental state creates a physical state. Now, we can imagine the covariance relations between mental and physical

[12] See also Kim (1998: 123, n.22).

[13] In fact we should call it Kant's intellectual intuition. Since that, for Kant, if the spontaneous intellect would be capable of intuition it would be capable of creating its own objects.

states at will. Either random (the same mental state would generate unpredictable physical states), one way supervenience of the physical over the mental (indistinguishable mental states would generate indistinguishable physical states, although the same physical state could be created by different mental states), or supervenience of the mental over the physical (two indistinguishable mental states could generate different physical states, although indistinguishable physical states must be generated by indistinguishable mental states[14]), either supervenience relation in both ways (each mental state would generate one and only one physical state[15]).

The point is that, whatever may be the pattern of covariance between the mental and the physical, the dependence relation stays the same. Or, in more general terms, if supervenience encloses two claims: a) covariance and b) dependence, then covariance does not imply anything about dependence. Asserting a certain kind of covariance won't tell us nothing about the dependence relation there is between the supervenient and subvenient properties.[16]

---

[14] Notice that this is just the normal supervenience relation with an inverted dependence relation. Therefore, we suppose that each mental state creates a particular physical state from a constrained set of physical states available (to assure multiple realizability) such as each of those physical states can only be generated by a single mental state. We must also stipulate – since indistinguishable mental states would be able to generate different physical states – that it is in principle impossible to know in advance which physical state from the constrained set will be created by a given mental state. Nothing of this seems incoherent.

[15] Notice that supervenience of the mental over the physical is not incompatible with supervenience of the physical over the mental: "in general, the supervenience of A on B does not exclude the supervenience of B on A." (Kim, 1998: 11) It is not clear if this also applies to mereological (or other forms of) supervenience, this seems to depend on how we individuate high-level properties.

[16] In fact, Kim would agree with us here, in fact, to Kim, saying that "the mental supervenes on the physical … says nothing about just what kind of dependence is involved in mind-body supervenience." (1994, p.582). Our novelty is to point out that this is incompatible with using supervenience to circumscribe physicalism. Notice also that, although covariance says nothing about dependence, dependence (as it is understood) entails covariance. This is a point Kim also establishes: "for there to be property dependence there must be property covariation." (1990: 148)

It seems clear then that, either we conceive of physicalism as a covariance thesis, either we reject the idea that c-supervenience can serve as a useful guideline to circumscribe physicalism. But it seems clear that physicalism should not be able to apply to possible worlds functioning according to God's-enabled parallelism or what we've called dictatorial idealism. In general, we could say that physicalism demands more than a pattern of covariation, it demands an explanation according to which the pattern of covariation can be understood as deriving wholly from physical entities; i.e. we must be able to explain it in terms of physical entities and structures, and nothing else.[17] A covariance claim does not secure this, it allows all sorts of dependence relations, and therefore it is not sufficient to circumscribe physicalism.

1.2. What m-supervenience cannot contradict.

We have seen that trying to reduce the concept of supervenience to a covariance relation enlarges its scope of application to include a wide range of theories like some forms of idealism or parallelism.[18] But, according to Kim, supervenience should not be understood as being committed to any particular form of dependence like "causal dependence, reductive dependence, mereological dependence, dependence grounded in definability or entailment, and the like" (1998: 14). But why should we accept Kim's claim? In his 1998 text, Kim offers a reasoning that seems far from clear. He asserts that accepting one type of dependence would preclude us from accepting competing candidates for the dependence relation (Cf. with next quote). But Kim's objective is precisely to find a core definition of supervenience which would be compatible with any kind of dependence relation. What we will argue here is that Kim's perspective

[17] There is an 'under-the-table' way out for Kim. He might claim that c-supervenience is both necessary and sufficient for minimal physicalism and that minimal physicalism is only necessary (but not sufficient) for physicalism. The final result is the same: c-supervenience does not circumscribe effectively physicalism. We might argue that it is a necessary condition for physicalism, but it is certainly not a sufficient one, it doesn't work as a definition.

[18] In fact it seems possible to include versions with magical properties, such as all-powerful witches, gnomes, fairies and all kind of stuff, which would also be compatible with supervenience of magical properties over physical properties (specially when magical beings are all-powerful to change the world as they saw fit).

can be maintained, up to a point, but instead of using c-supervenience we should use m-supervenience as defining the core concept of supervenience. There are two reasons for this, first, m-supervenience has a completely different status from other forms of supervenience dependence, it seems to be *necessary* to accommodate the current view of scientific explanations (even in a non-reductionist perspective). Second, it would seem that m-supervenience, with a second constraint that we will only present at the end of this chapter – intelligibility –, is sufficient to assert physicalism

According to Kim, accepting a specific kind of dependence relation would be to deny other dependence relations:

> "mind-body supervenience [i.e. c-supervenience] is consistent with a host of classical positions on the mind-body problem; in fact it is a shared commitment of many mutually exclusionary mind-body theories. As we will see, both emergentism and the view that the mental must be physically realized [i.e. functionalism] … imply mind-body supervenience. … What is more obvious, type physicalism, which reductively identifies mental properties with physical properties, implies mind-body supervenience. Moreover epiphenomenalism … is apparently committed to mind-body supervenience. … If mind-body supervenience is a commitment of each of these diverse, and conflicting, approaches to the mind-body problem, it cannot itself be a position on this problem that can be set alongside these classical alternatives." (1998: 12-13)

But in what sense would the acceptance of a dependence relation from parts to wholes deny emergentism, functionalism, type-physicalism, epiphenomenalism, or any other physicalist view? If Kim wants to show that m-supervenience is too strict a view, he must do more than just declare it, he must provide some kind of argument. But that is precisely what we can't find in his texts. And, it should be clear, that we *cannot* find such an argument. Because, as Kim is obliged to admit, it is

> "mereological supervenience … that seems to underlie and support the enormous productive research strategy of micro-reduction in modern

theoretical science. And, conversely, the success of this research strategy reinforces our belief in mereological supervenience." (1984b: 77).

So, if functionalism, epiphenomenalism, or any other theory, were considered to be contradicting m-supervenience, they might as well be considered to be contradicting natural science.

"It seems to be a fundamental methodological precept of theoretical physical science that we ought to formulate *microstructural theories* of objects and their properties – that is to try to understand the behaviour and properties of objects and processes in terms of the properties and relationships characterizing their microconstituents." Kim, (1984b: 96)

So, either m-supervenience is not 'a fundamental precept of theoretical physical science', either it cannot refute any theory that would not be also refuted by the fundamental methodological commitments that underlie theoretical physical science.

But can we suppose that Kim has failed to see this rather glaring contradiction? We should start by noting that what first led Kim to the idea that covariance does not imply dependence was that

"property covariation by itself does not warrant the use of "because", "in virtue of", etc., in describing the relationship [of supervenience] any more than it warrants the attribution of dependence. Thus, if we want to promote the doctrine of psychophysical supervenience, intending it to include a claim of psychophysical dependence, we had better be prepared to produce an independent justification of the dependency claim" (Kim, 1990: 147).

Now, it is this explanation, or a rationale, of the dependence relation which, following Kim, seems that cannot be found regarding m-supervenience. We know (or at least don't doubt) that it works, we have considered it a principle of scientific explanation, but, why should it work? Now is a good time to note that Kim puts m-supervenience on a par with causation, as the 'metaphysical backbone' of scientific explanation. He does that in the following terms:

> "This Democritean doctrine of m-supervenience, or microdeterminism, forms the metaphysical backbone of the method of microreduction, somewhat in the same way that the principle of causal determinism constitutes the objective basis of the method of causal explanation." (Kim, 1984b: 96-7)

The comparison with causation gives us an insight on the particular way Kim sees m-supervenience. Just as causation is an assumption of science in the sense that it is required for scientific explanation but it is beyond scientific demonstration[19] so mereological dependence would be an indemonstrable principle at the basis of modern science:

> "The part-whole relation is also important; however, its importance seems to derive largely from the belief that many crucial aspects of a whole, including its existence and nature are dependent on those of its parts. That is, mereological relations are significant because mereological determination, or "mereological supervenience," is, or is thought to be, a pervasive fact." (Kim, 1984a, 54 - underline of our responsibility)

We can easily see how Kim seems to understand m-supervenience as an assumption beyond any kind of proof. It is in this perspective, and since dependence relations are also explanations to covariance patterns, that we can understand Kim's claim that m-supervenience does not express a proper dependence claim. Since, although it claims a certain pattern of covariance between parts and wholes, it cannot explain why it should be present. That's why, in the end, we think that Kim's argument can be summed in the following passage:

> "But supervenience or determination is one think, explanation quite another. … Mereological supervenience of the mental on the physical

[19] As a humean conception of causality can easily show, it is coherent to suppose that we can only show that regularities exist but that causality is something which we project (and not 'discover') on reality.

would not automatically promise us an intelligible account of why the particular mind-body supervenience relations hold." (1998: 18).

It would be hard to underrate the importance of this point. We think (and we'll try to show) that, by putting on a pair causality with mereological dependence, Kim is in fact blurring what could be (at least in this text) our most important definition of physicalism. In fact, we will claim, it is precisely in the fact that mereological dependence is not on a pair with causation, that reside our best reasons to uphold physicalism. And this will also give us the opportunity of giving a much more precise definition of physicalism than the one that is usually given. The asymmetry between causation and mereological dependence can in fact be seen as the crux of our whole paper.

1.3. Explaining mereological dependence.

Arguably, the most important dividing line that separates Chalmers and Kim's account of supervenience concerns logical supervenience.[20] According to Kim, as we have seen, supervenience relations involve always an *independent* claim of dependence. Since types of supervenience are not defined by their dependence relation, the difference between global, weak or strong supervenience must be completely dictated by a pattern of covariation. One of the undesired consequences of considering covariation as the primary distinctive trait of each type of supervenience is that global supervenience does not imply weak (or strong) supervenience. But this is counterintuitive, in fact, we might have expected (like Kim did) that, since global supervenience applies cross-worlds and weak supervenience applies only intra-worlds, global supervenience would entail weak supervenience. However, when considering a covariation pattern we must assume first that we can get the same pattern of subvenient properties in both worlds to consider if we have or not the same pattern of supervenient properties. The problem is when the subvenient properties are different between the two worlds in a way that is innocuous to the dependence relation. Suppose for instance a possible

[20] According to Chalmers (1996: 367) there are a number of authors who also argue for logical supervenience. It is the case of Horgan (1984), Jackson (1993), Kirk (1974) and Lewis (1994). However, in the time available, we could not check these articles, therefore our discussion will be entirely based on what Chalmers has to say regarding logical supervenience.

world exactly like ours down to every physical particle, except for the fact that some of the brooms, also exactly alike in every physical respect, could fly (due to magical powers). Now we would want to say that global supervenience would not apply between these two worlds. But, according to the covariance view, it would be enough to change just any little bit of matter in any corner of the universe (in fact we would just have to put one of the brooms flying around) to disturb the subvenient properties in such a way as to make the comparison between the two worlds impossible. Therefore we could not reject global supervenience. And this turns global supervenience into a somewhat useless concept.[21]

Chalmers view on supervenience is undoubtedly focused on the dependence relation. Although he starts with the usual formulation of supervenience:

> "B-properties[22] *supervene* on A-properties if no two possible situations are identical with respect to their A-properties while differing in their B-properties." (33)[23]

[21] This is a rather informal presentation of the problem. A formal presentation was made first by Kim in his (1987). See especially pp. 82-4. A deeper discussion can be found in (1993: 169-171). The problem also applies, of course, to the relation between global and strong supervenience (the problem is always the same, we don't need to have perfect covariance to see that the dependence relation does not hold). Here is Kim's way of putting the subject: (1987: 82-3). "Consider two worlds, w1 and w2, each with two individuals a and b. In w1, a has G and F, and b has G. I w2 a has G but not F, and b lacks G. It is clear that this pair of worlds is a counterexample to the strong supervenience of F on G (or, more verbosely, the unit set of F on that of G) . However, the example leaves open the possibility that F globally supervenes on G: since the two worlds are not G-indiscernible they cannot be a counterexample to the global supervenience of F on G. Therefore, global supervenience cannot entail strong supervenience." (Notice that F is usually considered to be the supervenient and G the subvenient property in current literature.)

[22] Unless otherwise noticed A-properties are the subvenient properties (or supervenient base), while the B-properties are the supervenient ones.

[23] When quoting Chalmers the numbers inside parentheses always refer to Chalmers (1996) except if otherwise mentioned.

He rapidly introduces the different kinds of necessity on which the argument of the book rests. So, *natural (or nomological) supervenience* holds

> "when the A-facts about a situation *naturally necessitate* the B-facts. This happens when the same clusters of A-properties in our world are always accompanied by the same B-properties, wherever and whenever this happens." (37)

Whereas *logical supervenience* holds

> "when the A-facts *entail* the B-facts, where one fact entails another if it is logically impossible for the first to hold without the second." (36)

In Chalmers case the dependence relation comes first and establishes the kind of covariance pattern to emerge. So, obviously, natural supervenience only applies contingently, based on *a posteriori* relations.[24] It involves laws that are not logically necessary, i.e. which contradiction is logically possible. So, natural supervenience cannot be taken to apply cross-worlds, since, for every case of natural supervenience, we can always imagine a possible world where the supervenient relation does not apply. On the other hand, since in the case of logical supervenience we cannot conceive that the subvenient properties are present without the supervenient properties being also present, it follows that we cannot conceive of a possible world in which the supervenience relation would not hold. And this entails cross-world covariance. We should also note that, in Chalmers text, logical supervenience should be regarded as global logical supervenience, while natural supervenience should be regarded as local natural super-

[24] Obviously, if *a posteriori* necessity had a similar character to logical necessity (more specifically, if, as Kripke argues, it applied to all possible universes), the distinction between the two kinds of supervenience would loose its appeal. It's to separate the two that Chalmers dedicates so much importance to re-explaining a posteriori necessity in a way that it becomes harmless for his theory.

venience.[25] Notice also that global supervenience normally includes not only space but also time (a 'spatiotemporal hunk' (73)).

Chalmers presentation of supervenience as a relation of dependence seems much more clear and less awkward than Kim's version, for instance, while if we start by accepting dependence, covariance automatically follows, in Kim's version we need two separate principles. However, we should ask if there is such a thing as logical supervenience. We must remark that Kim has understood m-supervenience (which is perhaps where we can find the most undeniable dependence relation) as a methodological assumption, used, not because it is obvious, but due to the efficacy of its results. Chalmers also relies on m-supervenience[26] but he has the opposite view regarding the necessity of the relation. Like he says:

> "'If the low-level facts turn out like this, then the high-level facts will be like that.' The facts specified in the antecedents of this conditional effectively include all relevant empirical factors. Empirical evidence could show us that the antecedent of the conditional is false, but not that the conditional is false. In the extreme case, we can ensure that the antecedent gives a full specification of the low-level facts about the world. The very comprehensiveness of the antecedent ensures that empirical evidence is irrelevant to the conditional's truth value." (55)[27]

[25] There is a difficulty here. It seems clear that logical supervenience should be associated with global supervenience, and that natural supervenience should be associated with local supervenience. However, Chalmers remarks that there is a problem regarding global natural supervenience, but we could not understand the passage. (See Chalmers, 1996: 363, n.11: "Global natural supervenience without localized regularity is a coherent notion on a non-Humean account of laws, although perhaps not on a Humean (regularity-based) account. Even on a non-Humean account, though, it is hard to see what the evidence for such a relation could consist in."

[26] Notice that m-supervenience is a special case of logical supervenience, however it is also the strongest. It is not casual that all the arguments Chalmers presents in favour of the logical supervenience of A-facts to B-fact are based on m-supervenience.

[27] Underline of our responsibility. Notice that the low level facts (connected with A-properties) are the elementary entities described by contemporary physics, while the

Chalmers argues for the logical character of m-supervenience in three different ways: "using arguments that appeal to conceivability, to epistemological considerations, and to analysis of the concepts involved." (73)

The **conceivability** argument seems to rest on the lack of a counterexample. Just like in Kim, there is no rationale for the inexistence of counterexamples. Chalmers simply *asserts* (instead of demonstrating) that "this inconceivability does not seem to be due to any contingent limits in our cognitive capacity. Such a world is inconceivable *in principle*. Even a superbeing, or God, could not imagine such a world. There is simply no thing for them to imagine." (73)[28]

---

high-level facts (connected with B-properties) includes everything else made of these particles: "For our purposes, the relevant A-properties are usually the physical properties: more precisely, the fundamental properties that are invoked by a completed theory of physics. Perhaps these include mass, charge, spatio-temporal position; properties characterizing the distribution of various spatio-temporal fields, the exertion of various forces, and the form of various waves; and so on. The precise nature of this properties is not important. If physics changes radically, the relevant class of properties may be quite different from those I mention, but the arguments will go through all the same." (33)

[28] Since the argument seems to be nothing more than the inexistence of counter-examples, perhaps we should present Chalmers example: "What world could be identical to ours in every last microphysical fact but be biologically distinct? Say a wombat has two children in our world. The physical facts about our world will include facts about the distribution of every particle in the spatiotemporal hunk corresponding to the wombat, and its children, and their environments, and their evolutionary histories. If a world shared those physical facts with ours, but was not a world in which the wombat had two children, what would the difference consist in? Such a world seems quite inconceivable. Once a possible world is fixed to have all those physical facts the same, then the facts about wombathood and parenthood are automatically fixed." (73) The absence of any explanation to this state of affairs, however, is completely overlooked. We should also note that there is a simple reason why the wombat example works so well. The example demands us to think of the distribution of every elementary particle regarding the whole universe (and which includes the evolutionary stories of all the relevant events), but this is impossible even regarding small molecules. So this cannot be considered a good argument, because the example at hand is simply too difficult to analyse. In practice, Chalmers is just appealing to our intuitions. And, as we will shortly notice, Chalmers argument is in fact based, not on logical considerations, but on logical considerations con-

The **epistemological** argument is a little better because Chalmers notes the consequences resulting if 'there *were* a possible world' where m-supervenience would not apply. Notice, however, how the fact that we can know the consequences, imply that we can conceive of the situation. The epistemological argument, refutes the inconceivability argument. Not only is the falsity of the conditional conceivable, it has precise consequences to it! Note that this is a stronger argument then it seems. We can see this better by considering what the epistemological consequences of the failure of m-supervenience would be:

> "if I were in an alternative world [physically identical to ours], it would certainly *look* the same as this one. It instantiates the same distribution of particles found in the plants and animals of this world; indistinguishable patterns of photons are reflected from those entities; no difference would be revealed under even the closest examination. It follows that all the external evidence we possess fails to distinguish the possibilities. Insofar as the biological facts about our world are not logically supervenient, there is no way we can know those facts on the basis of external evidence." (74)

N.B., the failure of logical supervenience is conceivable, Chalmers conceives of it, and draws the appropriate consequences. However, just like in the previous case, we are still looking for a rationale: why should the lack of m-supervenience entail consequences in relation to third person observations? What is the connection, was it to be expected? In Chalmers arguments, these two strands (conceivability and epistemological considerations) appear disconnected, almost contradicting each other. The contradiction appears most clearly in the passage just following the one we quoted:

> "In actuality, however, there is no deep epistemological problem about biology. We come to know biological facts about our world on the basis of external evidence all the time, and there is no sceptical problem that arises." (74)

---

cerning the way the world is like! Logical supervenience might not be applicable to our world.

So, after all, the conditional (from low to high level properties) seems to be a matter of empirical results.[29] Where do we stand? The answer would plausibly be that the conditional does not depend on empirical evidence because the antecedent of the conditional 'effectively include all relevant empirical factors'. But the fact that the antecedent of the conditional has, or has not, all the relevant information, seems to be a question that can only be sorted out by empirical investigation. Chalmers, I think, would tend to agree with this. He would probably say that we could devise other possible worlds where, unlike in our world, the low-level facts are sufficiently imprecise *not* to yield (even with all the relevant information) all the high-level facts. But the point, Chalmers presumably stresses, is that this world is *unlike* our world, because, in our world, it is not "disputed that the physical world is causally closed or that behaviour can be explained in physical terms" (xiii). I think this is what Chalmers would say, but of course this would imply that logical supervenience is in fact based on assumptions that involve empirical claims. That is to say: logical supervenience would indeed express a logical relation, impossible not to exist, between low and high-level facts, but, this logical relation would itself be dependent on the way the world (contingently) is. So it would be a logical relation that would be based on empirical evidence. We think that it is possible to give m-supervenience a further support.[30] In fact it does seem that considerations on m-supervenience have a character of necessity that does not depend on empirical considerations. It is this character of necessity for which we will try to find a rationale in the next section. But lets first get back to Chalmers.

In the third and last strand, Chalmers considers analysability arguments.[31] He asserts that

---

[29] If there was this deep epistemological problem about biology (or any of the 'special sciences') then logical supervenience would not hold. And we can certainly imagine a world like this. For instance, the world as conceived by vitalists was surely like this!

[30] Besides, it is not at all clear that the world is causally closed at the physical level, but we wont focus that here.

[31] On the epistemological argument, Chalmers also focus the possibility of coherently denying the existence of causation (in a Humean perspective) and also of solipsism. The first discussion will be considered at length later on and the second does not seem relevant to our discus-

> "Meanings are fundamentally represented by intensions, not definitions. The role of analysis here is simply to characterize the intensions in sufficient detail that the existence of an entailment [from low to high-level properties] becomes clear." (78)

Chalmers than provides a 'rough-and-ready' analysis according to which what determines the intension of a concept are structural and functional conditions. He then assumes that structural conditions can be wholly determined by microphysical facts, and argues that the same happens with functional conditions. So, regarding structural properties, Chalmers says it is just obvious that they "are clearly entailed be microphysical facts." (79) Regarding functional conditions Chalmers argues a bit to the conclusion that "functional properties are all derivable, in principle, from the physical facts." (79) To espouse and criticise this argument here, however, seems out of the question, since it would lead us to a very long discussion that is tangential to the main argument of this paper. It is sufficient to note, however, that the analysability argument seems to presuppose much more than the other two. Not only on semantics, but specially on the causal efficacy of high-level properties (which seems to lead to consider that all the objects of the special sciences are epiphenomenal – since their function could be, in principle explained wholly on the terms of elementary physics). In any case, this argument seems the weakest of all the three and we will not consider it further.

1.4. A rationale for m-supervenience.

Until now we have analysed classical conceptions of supervenience examining the rather opposite positions of Chalmers and Kim on them. These two last sections present a quite different material. We have seen, from Chalmers' and Kim's arguments that mereological dependence seems to be an undeniable and pervasive fact. However, neither Kim nor Chalmers were able to provide us with a rationale that explained the necessity of that relation. What we will try to provide here is one such rationale. As

---

sion. Basically, what Chalmers argues regarding causality and solipsism is that their existence does not logically supervene on our experience of the world (or on facts). Which seems rather correct although not remarkably important.

can easily be imagined (specially taking into account the nature of the problem), this is not a subject where a swift demonstration can be given, and these two sections will remain highly tentative and speculative. Whatever they value might be, the best we can hope is that they can give rise to some useful discussion.

The rest of the section will be devoted to arguing in favour of its main assumptions but the general form of the argument can be presented now: if all the parts of the world have causal efficacy, and all that has causal efficacy is a part of the world, then e-supervenience amounts to say that the same causes generate the same effects (which would be equivalent to 'the same parts generate the same composites'). Which is a logical principle. Although at first this might seem a little trivial or even naïve we think it will soon be clear that there are a host of problems involved in this.

Perhaps the first thing to notice is that we didn't spoke about the microphysical or macrophysical level. That is why we have not spoke about m-supervenience, but e-supervenience, an expression we've coined for reasons that will soon become apparent.

Now, there are two ways in which m-supervenience can be thought to apply: time and space. The first is the traditional way, concerning determinism. It was in this sense that Leibniz principle (the same causes generate the same effects) was generally used, to remark that, (using a modern situation) given two indiscernible physical worlds at a certain moment t_1, than all the future moments of each world would also be indiscernible. This is synonymous with determinism. An equivalent expression is found regarding static situations, i.e. we think that the same microphysical facts generate the same macrophysical facts (which is the principle of m-supervenience). But can we prove any of these thesis?

Our answer is that, we can't! There is no proof (although there is reasonable evidence), nor logically, nor empirically, for any of those thesis, at least for now. This fact can be easily seen if we consider that there is no logical contradiction in supposing downwards causation. In fact, it is not absolutely clear that there is no empirical evidence for it. Now, if downwards causation existed in some possible world, it is obvious that mycrophysical description would not be enough to establish the macrophysi-

cal description. We will still be back on this point on our next chapter, but for now, the only important think to notice is that it is not a logical principle. The empirical evidence in its favour will be pondered in the next chapter (albeit in a simplified fashion). But if its not a logical principle, how are we supposed to argue for m-supervenience?

Our trick here is just to replace 'elementary part' by 'part'. If there is a kind of supervenience that does not require any assumptions regarding its supervenient base, except that it should consist in all the causes of a the supervenient properties, then, it will be just a tautology to say that the same parts generate the same composites. We achieve therefore a rationale, not for every form of supervenience, but for what we will call e-supervenience (**elementary supervenience**).

It is simpler to get the idea of e-supervenience if we draw a parallelism regarding determinism. In the same way as physicalism, as we will see, can be expressed as a supervenience thesis (that if two world are alike physically then they are alike *simpliciter*), determinism can also be expressed as a supervenience thesis (all worlds alike regarding the past will be alike regarding the future; or: the future entails the past). Now, it is not a logical thesis that the past entails the future or that there must be in any way any specific covariation pattern. We might suppose that in some worlds there would be even a random covariation of past and future events.[32] The only thing we need to suppose is that it is not necessary to exist a dependence relation between the successive states, but it is difficult to see how that would be *logically* incoherent. In empirical research it is also difficult to prove that in practice, although we might suppose that it is the only reasonable hypothesis, if we think the world should be intelligible at all.

Now, the connection with e-supervenience is simple, instead of presuming that past facts entail present (or future) facts, we simply say that the all the causes of the present facts are sufficient to yield a complete description of the present facts. Now, this leads to a somewhat innocuous perspective of supervenience, because, we might

[32] So, that, from all possible states of the possible worlds, the probability of encountering a particular state at a certain particular moment would be independent of the particular states at other moments, that is, they would always be random.

think that, if gods, angels and whatever existed and created lighting and all that, this would not be considered a violation to e-supervenience, since they would automatically be included as the causes of present (or future) events. In fact, whatever existed in the world that caused future events would be automatically included in the supervenient base description. So, we might think, this also is not useful to adequately circumscribe physicalism (the first question with which we were concerned).

But there are two points to notice here: the first is that we would agree that physicalism needs a different formulation, for reasons that we will espouse in the next section. The second point is that e-supervenience does not allow for everything. If the causes must end in some effect, we must suppose that they will be in some sense compatible with one another, and we must also explain how they can interact. On this basis, dualism could be excluded, for instance (for being incapable of explaining how an entity with no physical properties (mass, magnetic charge, nuclear force, etc), could act on entities which seem to only be affected by physical properties).

Until now we have spoken about the limits of the rationale we have to present. It does not apply straightforwardly to micro-macro dependence relations. However, this shouldn't surprise us. Either supervenience is conceived as a logical relation (and that is the way Chalmers wants to use it), either it can be proved or disproved by empirical results.[33] But the relation of *absolute* dependency of the macro-level over the micro-level is in fact an empirical question still being debated, which proves, *a fortiori*, that m-supervenience, conceived as a logical dependence relation from the micro to the macro-level, is simply unsustainable. We won't get a rationale for it, but we argue also that there is no rationale for it, unless it is based on empirical evidence, in which case it is simply not true that we can speak of 'logical supervenience' (in Chalmers sense). So, what our rationale provides is a way to speak of logical supervenience, to show the conditions under which A-facts *logically entail* B-facts. And those conditions, we say, arise when the A-facts can be shown to be the necessary and sufficient conditions for B-facts.

[33] It's true: we simply do not take seriously enough the Kripkean a posteriori necessity stuff.

Now this would be just a linguistic twist if we wouldn't present a further argument: the argument is that, the concept of 'being part of an object' can be reduced to the concept of 'having causal effects on the object'. And, if this is true, then, when we are saying that in the supervenient base is the description of all the elementary parts of an object, then we are also saying (and we can see this by conceptual analysis) that we have a complete description of the causes of the object. And this leads us to the conclusion we desire: mainly to understand why it is a logical principle that the same parts always generate the same composite. Namely, because our notion of 'part' is interchangeable with the notion of cause. So, it is to this second part of the argument that we will now turn.

This second part of the argument is divided in three parts: **first** we will remark the kind of problems to which (conceivable) entities with no causal efficacy are associated. We will argue then that such entities, whatever their other properties might be, cannot be considered as parts of observable objects. This concludes the first part of the argument which shows that causal efficacy is necessary for an entity to be considered as part of an object[34]. The **second** part of the argument considers a view of supervenience in which causal powers are *sufficient* to consider an entity as a part of the object (i.e. if x has causal powers over object P, then x is a part of P). This is the most controversial part of the argument, which will only be fully clarified in the following section. In the **third** part of the argument, we will recapitulate the rationale and also extract some consequences for the current debate. In the following section we conclude this chapter by exploring some consequences of adopting this rationale to gather a revised definition of physicalism.

[34] To be more precise we will always use the word 'universe' or 'world' instead of 'object', this is to avoid problems regarding extrinsic properties. Therefore our argument will, unfortunately, only apply to global forms of e-supervenience. However, this will be enough for most purposes, recall that Chalmers, for instance, uses logical supervenience almost always connected with global supervenience. Besides, we think that a more complex formulation of the argument can be found, that also applies to local e-supervenience, although we will not attempt it here.

1. Our initial question will be: what are the necessary conditions for something to be considered as a part of the universe?[35] We will start by considering several possible candidates:

1. To be a state or event.
2. To have a spatio-temporal location.
3. To have a causal role.
4. To have a physical structure.
5. Other properties.

From a simple inspection of these candidates we should easily spot that we have only one plausible candidate. Let's start by considering the first hypothesis. Imagine that there is some state or event (it does not matter if it's physical or mental) that is outside of our universe. Perhaps in a possible world or, alternatively, in a real universe whose dimensions are orthogonal to ours. In any case, if the states or events are such that they cannot have causal effects in our universe (including in physical or in mental states), then, it would be difficult to consider them as parts of our universe, if it wasn't for anything else because we could not know of their existence.[36] Since it does not

[35] By universe I mean a closed system (with no relational properties). In some contexts 'universe' means 'everything there is', this is not implied by our definition. Notice that, in fact, not only it is plausible to think that there might be more than one universe but this is implied by several scientific theories. For instance the existence of 'universes' with orthogonal dimensions to ours, and, therefore, with no no possibility of intercourse with ours, was argued for by Stephen Hawking. In fact it seems that they can not only exist but be engineered! They would (expectedly) be created and maintained if we could create certain sizes of black holes according to certain conditions (although its creators would never, by principle, be able to see what's going inside). Some versions of the inflationary theory of the universe also suggest that our universe might be one among many other causally separated universes.

[36] To find out of their existence by observation would imply that they had some causal consequences. If not we would have no reason to assert their existence since everything would be the same, and could be explained without resorting to them. If we knew them directly (like in first person observation) they would have an effect in our mental properties and, plausibly, also on our behaviour (for instance we could remark (loudly perhaps) to ourselves: 'I just sensed something').

seem inconceivable that states or events can occur outside of the universe, and having no intercourse with it, we must consider that having the property of being a state or event is not *sufficient* to be part of the universe. The same reasoning can easily be applied to physical entities (if they exist outside the universe, i.e. if they have no causal effects, then they cannot be considered as parts of it), and also to entities that somehow can be referenced to a space-time location of the universe. Because, whatever the other properties of the entity in question, it seems undeniable that, if an entity does not have *any* causal powers, then it cannot be known.[37] And if an entity cannot be known, we have no reason to consider it as part of our universe.[38]

2. It seems clear, therefore, that causal efficacy is a *necessary* condition for an entity to be considered as a part of the universe, since, without it, no imaginable entity, whatever its attributes, could ever even be known to exist. We now have to ponder if having causal powers is *sufficient* for an entity to be considered as part of the universe. This is equivalent to say that, if an entity is not part of the universe, then it will not

[37] This, of course, is a similar argument to the considerations Chalmers present in his 'second strand' regarding the epistemic consequences of denying the logical character of m-supervenience. (see the longer quote of Chalmers of p.74)

[38] Notice that an apparent difficulty still remains. For, if we suppose an entity with causal efficacy in a different world, the existence of that entity could also not be known (in this world) in spite of its having causal powers. So, it would seem that having causal powers is also not a necessary condition for being part of a world. But notice that the entity would still be part of the world where it had causal powers. While, by the contrary, the only way we could show that that would also happen to physical objects, states, etc, would be to assume that they had causal powers in the worlds they end up (otherwise they existence wouldn't even be noticed in any world whatever). Notice that nothing in our argument depends on the use of other worlds, all that is necessary for us is to suppose there could be such a thing as a physical entity with no causal powers to see that it is not the physical *qua* physical, that allows an entity to be consider to exist, but only its causal powers. Of course, we can either consider this as a *reductio* of the idea that physical entities can exist without causal powers, or that, if this can be conceived, then there are possibly physical entities that cannot even be know to exist. Anyway we will argue later on, that it would be disadvantageous to suppose that the concept of a physical entity amounts to the same thing as the concept of a causal entity.

have causal powers (in the universe).[39] So, what we have to find out is if there is any contradiction in thinking that an entity might have causal powers in the universe without being part of it.

However, the situation here is severely complex, and instead of getting to a definite solution what we will achieve are several alternative ways to look at the problem. The problem here, is that the particle we are trying to conceive would have effects that are undoubtedly part of the universe, but the causes of these effects would be outside of it.

Now, the first question we must place is that if this is conceivable at all. This seems easy to answer. For instance, imagine a story in which some people discover that the world they had been living in is just a computer simulation. They gather then the ability to get in and out of the simulation using their real bodies and apparatus.[40] Now, when they enter the simulated world, they are still capable of having causal effects upon it, those effects are not determined by the simulation itself. And so there is a clear sense in which they are part of the (simulated) world (they can be seen and the consequences of their actions are as real as anything else) but there is also a clear sense in which they are not part of it (in the sense that their actions are determined by something which is not part of the simulation). In the particular case of "The Matrix" we could easily check out (although only in principle) if the origin of behaviour was determined by the universe or not. In principle, the actions chosen by the characters were not in large part determined by facts of the simulated world and so they would violate the causal closure of the simulation. And, in this respect, the film reminds us of cartesian dualism where minds also imprint on bodies actions that are not determined by previous events. In both cases we have entities that do not belong to space and time but capable of acting on space and time.

---

[39] This is a particular form of the contraposition argument: $(A \rightarrow B) \leftrightarrow (\sim B \rightarrow \sim A)$.

[40] In fact this is the plot, with a few changes, of the science fiction movie "The Matrix". This kind of problems has also some similarities with a text of Chalmers available at the internet that we will present later on.

However, there is a simple way of imagining that the entity could be considered a part of the world without having any causal powers on it. For instance, imagine a sequel to "The Matrix" where the characters could enter the simulation but just to see what was going on. Then, they could get back to the real world and tell all about it, but they could in no way alter what was going on. It seems indubitable that this would be a case where, in some loose sense, they would still be part (temporarily) of the simulation, since the events there would have causal effects in their other world. But they would be, in principle, unobservable from the inside of the simulation. This could be considered a case of epiphenomenalism with a twist (a dualist twist).

Now it's time to make a break. It seems that, if at all possible, the conception of an entity that is (in some sense) outside of the universe but has causal efficacy on it, is strongly connected with dualism, or a with epiphenomenalism with a twist, or with some other form of dualism. So, either we are disposed to consider these possibilities seriously, or we must accept that having causal powers is sufficient to consider something as part of the universe. In this section we will consider that the claim of sufficient condition is established for our purposes. In the next section, we will try to show what would be the consequences of accepting the alternative.

Perhaps the alternative would be to conflate the two properties of 'being a physical entity' and 'having causal powers'. However, as we will argue next, this is undesirable for it would allow physicalism to be indistinguishable from, for instance, cartesian dualism. What we would like to remark now is that – while the notion of 'physical' is still left open for debate – having causal powers in the universe seems to be *necessary and sufficient* conditions for being considered a part of the universe. For relational properties are just another way of stating causal powers, and so, to accept that objects are entirely defined by their relational properties is to accept that they are entirely defined by their causal powers. Since this is a rather new proposal we will in this paper call mereological supervenience seen from this perspective **e-supervenience** (i.e. elementary supervenience).

3.. So, with most of the hard work done, we just need to recapitulate our main idea. If e-supervenience is the thesis that the same parts generate the same composites, then e-supervenience is just a logical truism, it is the same as saying that the same causes have the same effects. This happens because we cannot conceive of a cause which is not part of an object, and also because we cannot conceive of a part of an object that has no causal effects upon it. So, when we are speaking about parts of an object, we are in fact speaking about all the causes that make the object to have certain properties and not others. This clearly shows that e-supervenience is a logical principle. But, as we have remarked already, being a logical principle has certain costs. Remarkably, e-supervenience has to be compatible with every kind of empirical evidence at hand. So, our notion of e-supervenience stands between Chalmers' and Kim's version. It is based on entailment and not covariance (following Chalmers) but it does not assert a form of dependency based on what Kim would describe as a metaphysical assumption, or belief. By the contrary, it is a logical principle that can be useful to clarify other forms of supervenience.

Perhaps one of the most useful consequences we can easily seen is in regard to the supervenience of natural laws over physical properties. In the standard formulation of supervenience we have a situation where we start with a bunch a particles, each with spatial, momentum, energy, time, and other such properties. This gives us the initial state of the system. But, it is usually argued, *besides* the initial state we can have rules of transformation that give us the next states of the system. Another way to say this is that, *if we add the laws of nature*, (which are considered to be included in a different description of the description of the elementary level) we will then be able (at least in principle) to get the state of the system at a later time. One of the consequences of the standard perspective is that the laws of nature do not supervene on the basic description of the system, since they are independent of the basic description.

However, if we accept Chalmers (and Russell's) point that all there is to the description of a particle is its set of relational properties, we might wonder how are we supposed to describe any physical system (even very small systems) at the elementary

level without describing at the same time all the laws of nature implied.[41] And in fact it seems we would have to include all laws of nature even for relatively small micro-physical systems (imagine a system composed by an atom and a photon, for instance). Anyway, if entities cannot in fact be described unless by describing their relational properties, then there seems to be no other way that, not only do laws of nature in fact supervene on a description of the elementary entities, but they would in fact be necessarily included in the description!

> "One can see this [the non-supervenience of natural laws] by noting the logical possibility of a world physically indiscernible from ours over its entire spatiotemporal history, but with different laws. For example, it might be a law of that world that whenever two hundred tons of pure gold are assembled in a vacuum, it will transmute into lead. Otherwise its laws are identical, with minor modifications where necessary." (86)

We might get a little surprised by the outlandish example and the 'minor modifications where necessary' could also put our minds to work for a big while,[42] but the

[41] There is a difficulty here, what is the set of laws implied by the description of an atom? If we want to give a complete description of an electron, then this description must be such that it is effectively able to predict how it will behave in every conceivable completely specified situation. If there was a situation, completely specified, except for the behaviour of the electron, that would mean that, to make the situation predictable, we would have to add some properties to the definition of the electron (so it would be incomplete). So a description of every elementary particle (it would seem that every elementary particle known is present almost everywhere) would have to be able to yield predictions relative to every possible system. And, so, it would seem that all the laws of nature would be derivable just for an elementary description of a glass of water (which seems absurd). This is a problem we didn't sorted out.

[42] We should perhaps notice *in passim* that the strength of this example is that it clearly leads us into an outlandish world of fantasy. We should do better to say that whenever Otto said 'charity' a strange stone would appear from nowhere and hit the heads of the three closest ladies! How in fact are we supposed to believe that completely stable molecules of a metal would change into other molecules? Where would the remaining protons and neutrons go? What is the amount of energy that would be needed to make this transformation? And how would it possibly start? And how could the difference in the exterior parts of the metal be

crucial point has nothing to do with this. The point is, the two universes, whatever the period in history, could have never an equal description at the level of the subvenient description. Although they might be macroscopically similar (but this would be incredible almost inconceivable – see our previous footnote) they could never be indiscernible at the level of the elementary particles for a reason Chalmers himself mentions (as we will see in the next section): a description of the elementary particles is constituted by concepts that entail their possible modes of intercourse with other particles. So, since obviously the atoms of gold and / or lead would have quite different properties, this means that the electrons and / or quarks would also have to had quite different descriptions (which would allow for such differences in the atoms to occur). This means that not only would the two universes be different, they would be different in the description of every single atom!

---

necessary to process the transformation (the vacuum condition)? All the scientifically minded questions we might put on this example shows just one thing, not only it is implausible, it is inconceivable, at least if outside the context of a kids story where there can happily be *ad hoc* laws applicable only to two tons of gold..

Less outlandish examples can be found for universes with the same (or slightly different) laws but with different facts. For instance, we could imagine a similar universe to ours but where black holes did not exist. However, the important point is that the laws would need to be very similar or identical. In general, a small difference in the laws of nature will produce a grand difference in events and not the other way around. We know for instance that having two tons of uranium 235 is an impossibility in our world (we'd get a thermonuclear explosion for certain!) but this is fully derivable from the elementary laws that describe the behaviour of atomic nucleus. To change the minimum size necessary to turn a uranium 235 pile into an instantaneous nuclear explosion, we would have to change the number of atomic radioactive decays per second, or the probability that a certain neutron causes the fission of a nearby atom, etc. To change any of this would have gigantic implications for the rest of the universe. That is why, in the history of science, when we are trying to fit in a phenomenon that is not explainable by known laws, we have often to change the entire structure of our laws. So it is completely implausible to suppose that natural laws could change just with 'minor modifications'. Probably very small changes in laws would generate an entire different universe. (Notice that this is a subject where a proof can be found with the help of computer simulations, we could change one of the basic values of one of the 4 elementary forces, or we could change slightly the value of a particle and see how it ends!)

Chalmers example does not suggest how there can be an alternative to this account. So, it still seems that laws of nature do in fact logically supervene on the lowest-level description of reality.[43] Notice, however, that, although the description of the initial state of the universe seems to logically entail the description of all the natural laws involved, the opposite does not happen. To be exact, the description of the natural laws does not, of course, suffice to describe any particular state of the universe. And this happens, because the same laws apply (by definition) to any particular state of a universe, which is just another way to say that every particular state of a universe entails its natural laws.

1.5. A revised definition of physicalism

In this concluding section we will try to provide an alternative definition of physicalism. We should say, once more, that these two last sections are highly speculative and that they are introduced more to stimulate the debate rather than to demonstrate our particular point of view. For this, many other topics and authors would necessarily have to be discussed.

We should start by referring that common formulations of physicalism are highly uninformative. For instance, in the *Oxford Dictionary of Philosophy* physicalism is defined as:

> "physicalism [in the widest sense of the term] is the thesis ... that whatever exist or occurs is ultimately constituted out of physical entities." (p.617)

[43] Notice that this doesn't seem to affect any of the humean arguments that show that we have no reason to suppose that the future will be like the past, or the resulting humean account of causation. The argument starts by assuming (dogmatically, we might say) the validity of the natural laws. On this basis we argue that the description of elementary particles is not something other than the description of natural laws. Of course, an example based on Hume would be easy to give, we could just suppose that in a world exactly like ours the sun would not rise tomorrow. But this implies putting at stake the possibility of correctly characterizing a physical particle.

David Lewis, when defining physicalism with the help of supervenience, gets at a similar formulation:

> "Minimal Materialism is a supervenience thesis: no difference without physical difference. That is: any two possibilities that are alike physically are just alike *simpliciter*." (Lewis, 1988: 507)[44]

To say that physicalism (or materialism) is an assertion of the physical character of everything that exists is, at least, uninformative. For we would like to say that there are some situations in which physicalism is wrong. For instance, a world where cartesian dualism applies, is a world where physicalism does not apply and *vice-versa*. So, we can see that there must be a) some criterion for distinguishing between physical and non-physical entities; and b) non-physical entities might conceivable exist. If any one of these assumptions is false then physicalism is not really an informative theory about the way the world is (it tells us nothing about the world).

To introduce our topic we will need to make a temporary detour. We can start in the following way: suppose that a physical entity, like an electron, for instance, were to be replaced by a non-material entity that would behave in exactly the same way. How could we know that the electron was not a material entity? We think it is easy to see that there is no answer to this question. Because, the essence of third person knowledge, is that only behaviours of objects can be known. But this amount to saying that the electron (or any other object) is defined by nothing more than a collection of potential intercourses with other objects. And, as we will immediately see, it does also mean that objects would be nothing more than aggregates of relational properties. In this respect we think that Russell's and Chalmers' position is incontrovertible:

> "physical theory only characterizes its basic entities *relationally,* in terms of their causal and other relations to other entities. … [All] that a specification of mass ultimately comes to is a propensity to be accelerated in certain ways by forces, and so on. Each entity is characterized by

[44] Chalmers presents a similar (for our purposes) definition: "materialism is true if all the positive facts about the world are globally logically supervenient on the physical facts." (41)

> its relation to other entities, and these entities are characterized by their relation to other entities, and so on for ever …. The picture of the physical world that this yields is that of a giant causal flux, but the picture tells us nothing about what all this causation *relates.* Reference to the proton is fixed as the thing that causes interactions of a certain kind, that combines in certain ways with other entities and so on; but what is the thing that is doing the causing and combining? As Russell (1927) notes, this is a matter about which physical theory is silent. …
>
> Intuitively, it is more reasonable to suppose that the basic entities that all this causation relates have some … *intrinsic* properties … But physics can at best fix reference to those [intrinsic] properties by virtue of their extrinsic relations" (153)

We think this 'relational' view of the fundamental objects of physics (the 'parts' with which m-supervenience is concerned) is a direct consequence of the way we get to know objects. Indeed, science is the attempt to make objects 'kick back'. (This is also sometimes seen as a criterion for saying that an entity objectively exist.) All we can know is how objects behave. Once we know that a proton behaves in a certain way there is nothing else to know. If something would behave exactly like a proton on what grounds could we possibly deny that it was different from other protons? We can have no reason, because all we can know of objects is the way they behave. So, if physicalism says something further. For instance if physicalism says that an electron, *besides* behaving in this and that way, also is a physical entity, then it is not really saying nothing at all, at least nothing that can be comprehended from a rational point of view. For, if being a physical entity, is not a property that has behavioural consequences, then what could it be? [45]

[45] This is perhaps the only occasion for us to notice that doctrines like materialism or dualism, when stripped off of their religious and political underpinnings, have not that much to say. What we are attempting to do in these two sections is to give an alternative account of physicalism that is independent of religious problems (like those involving the existence of souls, free will, sin, etc). The same should be done for formulations of scientific theories alternative to physicalism, but we will not attempt that here.

But notice that, if the guys from the Matrix got in the simulation, they would not behave according to the rules of the simulation. In the same way, interactionism between mind and body seems to presuppose that the mind causes behaviour in the physical world following rules that are not determined by the physical domain. And this gets us to where we started (in the first section). It seems that physicalism is an empirical theory, because it would be refuted if entities from outside the world could have powers inside the world.

However, this is false! Or at least, it is not true without a further qualification. Suppose that in some strange world, entities with a situation similar to the situation describes by "The Matrix" or cartesian dualism, would in fact exist and exert their causal powers in one world being entirely determined by another. Now, if their causal powers were arbitrary, if they corresponded to 'free will', or something equally unpredictable, then their existence could be asserted by empirical observation; we would see their causal effects, which would be unexplainable in any other way.

But suppose that interacting minds, instead of mysterious and unpredictable entities, would be something on which a theory, lets say a mathematical theory, could be made. Predictions could be made, tested and repeated, and the results were no less reliable then in the rest of science. Or suppose that we would indeed find evidence of downwards causation, in a way that the laws that govern it are completely amenable to scientific discourse. On what grounds would we be able to say that such entities were not 'physical'? To see things more at close hand, suppose that they could change the electronic orbits at will, then we could see their presence whenever an electron suddenly jumped to another orbit with no apparent reason.

But now suppose, by the contrary, that these alien, independent, entities, would be governed by lawful rules. In that case they would change the orbit of the electron but they would also always change it in exactly the same manner given the exactly the same circumstances. Now, the circumstances that dictated the change in orbit were determined not only by what was going on on our world, but also by what was going on in the alien world. But since we are supposing that that world was lawfully fixed, the changes of the orbits of the electrons would have a perfect covariation with

changes in conditions of this world, although they would in fact depended on both worlds.

Now, if the modification was always constant and predictable, we could make a general rule for electron jumping. But, how would we explain this rule? Obviously, we would express a covariation between some environmental circumstances and the jumping of the electron. This would in turn be explained by some property that would attribute a function to the electron such that, when certain conditions were present, other states of the system would ensue. And all these relational properties would be systematised by attributing to the electron certain intrinsic properties from which the extrinsic, relational properties could be derived. And therefore, we would get at a physical theory of electronic orbits exactly alike another one in which the events were completely determined by only one world. So, if dualistic spirits, gnomes and fairies would be all around in our world, we couldn't notice it even in the smallest bit, unless they acted in a non-lawful manner.

But, of course, this is just a trivial point, given what we know about the way third person knowledge is possible. Notice that, as we have seen, whatever an entity may be besides the rules of its behaviour, is something that must stay completely hidden from us. If a physical electron were to be replaced by a 'mental' electron, and if the behaviour was the same, then we would have no way of spotting any other difference. So, it seems necessary that, if there is some difference between mental and physical entities, it must be somehow be characterised in terms of behaviour.[46]

[46] In scientific research, questions beyond behaviour seem to be simply meaningless (this was what at the basis of logical positivism, behaviourism and other trends, although perhaps not for the right reasons). For suppose someone would say that there is this invisible hand which is pushing around the planets of the solar system keeping them exactly in their orbits just as if a force of gravity existed. Who would know, or care? In a sense, these are senseless hypothesis, it is not only that we cannot determine their truth value, but that there is no truth value regarding that question, given the way we get to know the world (from a third person perspective). Notice that, whatever the empirical way of determining this kind of questions, we would always end up by achieving only more reactions, structured reactions. If what we pose as question is the origin, not of the behaviour, but of the law itself, then it is not visible what a possible answer to that question would amount to (notice it could not be further behaviour

The answer which we will propose here is quite simple: an entity or part of the world is whatever has causal powers on the world. But a *physical* entity is whatever has causal powers *and* whose behaviour can be adequately predicted by a universal law. The loose point here is the notion of 'adequate prediction'. If determinism was still an available option we could replace the expression by saying that physical is whatever has causal effects and whose behaviour can be completely predicted by universal laws. The advantage of this definition is that, according to it, physicalism would preclude everything that is unexplainable. Which means, if the behaviour of an entity cannot, in principle, be explained by the recourse to natural laws, if it is totally unexplainable, then it would not be physical. And the empirical question that remains is, *either* the world is predictable, intelligible, and there are previous facts which can be invoked to explain present facts; *either* there are mysterious, magical facts that no one knows could explain without supernatural intervention.

If this is to be accepted, physicalism ceases to be a thesis about the ontology of the world but about the lawful character of causality; not what about *what* the world is, but *how* the world behaves. Asserting that there are no such things as ghosts is not a thesis about what implements the laws, but the assertion that the laws of the universe are fully specifiable.

So physicalism, on this reading, is just the assumption that everything in the universe is understandable, that there are no mysteries which are, in principle hidden, from rational explanation. Physicalism is not the thesis that only 'physical' things exist. It is the thesis that the world is intelligible, it is a belief in the rationality of the world, a belief in which the last 300 years of science gave us every motive to confide in.[47]

---

since, besides leading to an infinite regression, would not solve our problem). So, perhaps the most plausible answer would be to say, in Carnap's sense, that the question is simply meaningless.

[47] The connection between dualism and the failure of full-intelligibility can also be found in other authors, for instance in Hofstadter & Dennett (1981: 388) they say: "When antimatter was first postulated by physicists, dualists didn't react with glee and taunts and said 'I told you so!' Why not? Hadn't physicists just supported their claim that the universe had two

1.6. An appendix to the first chapter: some loose strands.

We could not end this chapter without making reference to some of the subjects that would be essential in the full exposition of this subject.

a) *Metaphysical (or a posteriori) necessity.*

Both regarding Chalmers work, but specially in our own text, we have made implicitly assumed that all there is to necessity is logical necessity. Our view might be seen as close to Hume's, with it's old-fashioned distinction between relations of ideas and matters of fact. However, we would have to argue for this exclusive character of logical necessity since, nowadays, it is argued by several authors that when we have, for instance, an identity between two rigid designators, then that identity applies cross worlds. Which means that if Water is H_2O, and 'water' and 'H_2O' are rigid designators, then water is necessarily H_2O (i.e. water is H_2O in all possible worlds).

We didn't feel inclined here to critically analyse this view. Although we understand its importance in current debate where there is an attempt to hold on to the notion that there is some kind of metaphysical (or a posteriori necessity) beyond just simple logical necessity. In this particular point we are completely on Chalmers side. We think Chalmers central idea is that metaphysical necessity is (if anything) just logical necessity. The introduction of metaphysical necessity derives of mixing different meanings of terms like 'water' which sometimes refer to H_2O and other times refer to 'watery stuff' (the stuff we drink, that exists in rivers and lakes, and so on). In this respect we would probably be even a little more radical (in our presentation) than Chalmers is. (So we would say that, if water means H_2O then, it is a logical conse-

---

radically different sorts of stuff in it? The main trouble with antimatter, from the dualist's point of view, was that however exotic it was, it was still amenable to investigation by the methods of the physical sciences. Mind-stuff, on the other hand, was supposed to be off limits to science. But if it is, then we have a guarantee that the mystery will never go away." They end with the sentence "Some people like that idea." Which is unfortunate since the true problem has essentially nothing to do with aesthetics. The problem with this kind of discussions is that people almost invariably mix aesthetic and moral judgements in a discussion that starts and stems from epistemological problems. It's like when we speak of a loved one, it's difficult to avoid extreme passion or hatred to see what's *really* going on.

quence that water is H_2O in all possible worlds; if water means 'watery stuff' then it is a logical consequence that water is watery stuff in all possible worlds. The confusion results from using 'water' indistinctly in the two senses, and then trying to say that there is some kind of (metaphysical or a posteriori) necessary relation between watery stuff and H_2O.) Our position, on this respect, would be much closer to David Lewis who asserts:

> "Kripke (1972) vigorously intuits that some names for mental states, in particular pain, are rigid designators: that is, it's not contingent what their referents are. I myself intuit no such thing, so the non-rigidity imputed by causal-role analyses troubles me not at all." ´(Lewis, 1994:419).

In fact, after exposing Chalmers line of argument (in 1996 and 1995), we would try to attack the notion of rigid designator. We consider this a very important topic and we intend to write something on it in a later occasion.

b) *External properties and wide content.*

The problem of wide content, although not directly relevant to our topic here, is perhaps the problem which has received most attention, whenever we discuss mind-body supervenience. The problem is that, although we have every reason to believe that the causal processes of our behaviour are local, i.e. they are centred on the brain, we also know that there are properties of propositional attitudes which depend on states and events which are enormously eloigned from the brain. So, for instance, if I look at the sky and say 'I'm looking at a star', I'm most probably wrong, since most white points in the sky are not single stars.[48] But notice also that it would be implausible to suppose that the truth value of that proposition would alter my behaviour. And this happens generally (so if we think that the train leaves at 10, we will only arrive at the station a little before, whatever is the true schedule of the train). So it would seem that semantic properties, or perhaps the truth values of propositions, would be causally inefficacious. As Pierre Jacob (1997: 41) puts it:

[48] They are usually binary or ternary solar systems, but of course they can be also galaxies or planets.

> "The semantic property of my belief is a highly relational (extrinsic) property of my brain. The puzzle [...] is to understand how such an extrinsic, non-local property of my brain can be involved in such local processes as the propagation of electrical and chemical signals through nerve fibres whereby nerve muscle contraction and relaxation are controlled."

We will not debate this problem here, we do not have a sufficient knowledge of the literature, and, at least apparently, it seems also tangential to our discussion (since we have always discussed global forms of supervenience, and we also have not debated specifically the causal efficacy of mental or semantic properties). Nevertheless, we should assert that wide-content problems, just like the problem of the aesthetic, moral, legal or monetary values, do not appear in global forms of supervenience.

c) *Supervenience and preemption in the special sciences.*

Perhaps the missing topic that would be the most relevant to this work is the discussion regarding the causal efficacy of entities described by the special sciences. The problem is exactly the same that affects the causal efficacy of the mental. Supervenient properties seem to be preempted by the subvenient properties, since that, for each causal chain we have two competing candidates trying to be at its beginning. So suppose that some state (physical or mental) P supervenes on the subvenient state P*, then although P may plausible generate state Q, P* will also generate state Q*, which is of course the subvenient base of Q. So, we could say that Q was caused by P, but in fact we know that P* caused Q*, and that Q* entails Q. So, we would be conducted either to overdetermination, either to preemption. This is perhaps the topic which we are most anxious to study at the moment. It reveals difficulties in the notion of epiphenomenal properties which are too often associated with mental events, but that in fact seem to generalise all over science.

Abstract of the first chapter: We have analysed the relation between supervenience and physicalism. We have started by noting that covariance is not sufficient for physicalism, and that mereological dependence, although pervasive, seems to be unjustified from a logical point of view. We have then argued that a certain dependence relation can better circumscribe physicalism. A pure logical dependence relation can be found for e-supervenience but it is also not sufficient to

physicalism (it allows for alternatives like dualism). Physicalism seems to involve an extra (non-logical) assumption that restricts physical events to those that can be predicted by a universal law. In that sense physicalism is not (or should not be regarded as) an ontological theory, but is the methodological assumption that the world is capable of being the subject of a fully rational explanation.

2. Chalmers' 'hard problem'.

Through his book *The Conscious Mind* (in 1996) and also through his highly influential key-note article in the *Journal of Consciousness Studies* (one year before), Chalmers formulation of the 'hard problem' of consciousness has become highly influential and has gained widespread acceptance. In this section we will try to show that the hard problem, as formulated, cannot possibly be solved. In the next sections we will try to show that a reformulated version of the hard problem, and we will propose an alternative version of the hard problem which we will show to be solvable.

In the standard formulation, the 'hard problem' is distinct from the (relatively more) 'easy' problems of discrimination, reportability, integration and control, etc:

> "The easy problems are easy precisely because they concern the explanation of cognitive abilities and functions. To explain a cognitive function, we need only specify a mechanism that can perform the function. … The hard problem, by contrast, is not a problem about how functions are performed. For any given function that we explain, there remains a nontrivial further question: why is the performance of this function associated with conscious experience?" (Chalmers, 1997: 380)

> "Why doesn't all this information-processing go on 'in the dark', free of any inner feel" (Chalmers, 1995: 13).

Notice that if the performance of any particular function would entail conscious experience, the problem could be solved. However, the formulation of the hard problem is accompanied, in Chalmers thought, by the thesis that we can conceive the performance of *every* cognitive function without an associated phenomenal experience. So the solution to the hard problem cannot be a logical one.

The thesis that functional implementation does not entail consciousness is one of the consequences of the thesis that the mental does not logically supervene on the physical, or the thesis of the logical possibility of zombies. So there is an asymmetry between supervenience regarding physical properties and supervenience of the mental over the physical. As we have seen, according to Chalmers every physical process

logically supervenes on the physical since "the physical facts about those processes *entail* the existence of the phenomena." (106) but the same does not happen to mental properties.

The (mere) natural supervenience of the mental over the physical has raised much discussion, it has severe implications (like the epiphenomenal character of all mental properties[49]) and it would take us another chapter just to review the literature on this topic, which is mainly focused on metaphysical necessity.[50] What we will do here is to capture the main argument that Chalmers adduces for the lack of logical supervenience of the mental over the physical.

In the chapter dedicated to the lack of logical supervenience of the mental over the physical (Chalmers 1996: 93-107) Chalmers presents five arguments following the same strategy invoked to show the logical character of m-supervenience. So, like in the previous argument, Chalmers presents arguments based on conceivability, epistemology and analysability. And, as before, we find the arguments in this chapter very inadequate. For instance, regarding the argument based on the conceivability of zombies Chalmers' punch line is just that:

> "the question is whether the notion of a zombie is conceptually coherent. The mere intelligibility of the notion is enough to establish the conclusion." (96)

But of course it isn't clear that the notion is intelligible. Like Dennett (1991) points out, both regarding zombies and regarding the Mary-the-scientist argument,[51] the thought experiment is simply too complex to be imagined. And we also don't know enough about the brain to know which functions processes are correlated with

[49] For the relation between natural supervenience (global or local) and epiphenomenalism see Chalmers (1996: 150).

[50] The most recent paper by Chalmers on this topic is "Does Conceivability Entail Possibility?" (available at http://www.u.arizona.edu/~chalmers/papers/conceivability.html).

[51] Regarding zombies see (Dennett 1991: 281). Dennett uses the same argument in a more detailed form regarding Mary's thought experiment (see pp. 399-406).

conscious experience.[52] So the intelligibility of zombies can depend on two things: it might be that zombies are really a logical possibility, or perhaps it is just our lack of imagination that stops us from seeing the incongruence.[53] In any case, it just won't do to say "I confess that the logical possibility of zombies seems equally obvious to me." (96) The evidence presented to support the other four arguments does not seem substantially different from the evidence presented in this one. In any case, we should perhaps make a short description of the five arguments, three of them are widely known: (1) zombies; (2) inverted spectrum; (4) Mary-the-scientist. The third argument just reasserts without arguing that the low-level facts do not entail facts about consciousness.[54] And the fifth argument assumes that: "For consciousness to be entailed by a set of physical facts one would need some kind of analysis of the notion of consciousness … and there is no such analysis to be had." (104)

All these arguments suffer from the same difficulty of the arguments regarding the logical supervenience of m-supervenience. In both cases what is missing is a rationale that explains all the relevant facts through a well established principle. Our last sec-

[52] Chalmers says about this that "it is enough to imagine the system at a coarse level, and to make sure that we conceive it with appropriate sophisticated mechanisms of perception, categorization," etc. But this is question-begging, because what Chalmers is stating is just that current cognitive models do not entail consciousness (which is trivial but far from enough), while what he needs to argue is that no cognitive model can entail consciousness. For this we need a quite different argument.

[53] We should note that perhaps the greatest evidence for zombies is that 'we are all zombies' (like Dennett says in a different context), but just sometimes. This is something that all of us remark. Most of our cognitive functions, like typing in the keyboard, writing a letter, or choosing our next letters, words, or sentences (in oral speech), are usually made in zombie mode. The problem is that there seems to be specific tasks that we cannot make without consciousness (like learning a completely new task). But Chalmers argument blurs this distinction. More to the point, our experience of zombihood promotes the possibility of what Chalmers calls psychological zombies and not of phenomenal zombies (for the distinction between the two types of zombies Chalmers 1996: 95).

[54] Chalmers only argument here seems to be: "But nothing in this vast causal story could lead one who had not experienced it [consciousness] directly to believe that there should be any *consciousness.* The very idea would be unreasonable, almost mystical perhaps." (102)

tions of the previous chapter were highly tentative and speculative, since we were working without any supporting bibliography. However, in this case, it is Chalmers himself who provides such a rationale, although in a different context of his book.

We can find this argument in the criticism that Chalmers elaborates on 'new-physics materialism' and 'interactionist dualism'. In several passages Chalmers notices that attempts to explain consciousness that involve new, or to be discovered, physical theories, are deemed to fail. This happens because: "the phenomenal component can be coherently subtracted from the causal component."[55] (163). Chalmers gives several examples of new-physics cases. For instance, one might be 'tempted' to use the quantum indeterminacy of quantum mechanics to allow that "consciousness might be responsible for filling the resultant causal gaps, determining which values some physical magnitudes might take within an apparently "probabilistic" distribution" (157), or consciousness might be thought to have a causal role on the world by bringing "about the so-called 'collapse of the wave function.'" (157) In all these cases, and in all possible cases, whatever our physics might be, this cannot provide an answer to the hard problem. Because,

> "even if it [a new physics] explicitly incorporates phenomenal properties, the fact that these properties are phenomenal can play no essential role in the causal / dynamic story; we would be left with a coherent physics even if that aspect were subtracted. Either way, the [structure and] dynamics is all we need to explain causal interactions, and no set of facts about [structure and] dynamics adds up to a fact about phenomenology. A zombie story can therefore still be told." (163)

[55] Chalmers gives a specific example of how this exclusion can be done: "Imagine (with Eccles) that "psychons" in the non-physical mind push around physical processes in the brain, and that psychons are the seat of experience. We can tell a story about the causal relation between psychons and physical processes, and a story about the causal dynamics among psychons, without ever invoking the fact that psychons have phenomenal properties. Just as with physical processes, we can imagine subtracting the phenomenal properties of psychons, yielding a situation in which the causal dynamics are isomorphic. It follows that the fact that psychons are the seat of experience plays no essential role in a causal explanation, and that even in this picture experience is explanatorily irrelevant." (158)

So, now we can see how hard the hard problem really is: it cannot be solved whatever physical theory we might have, whatever possible world we are considering. The only thing we must assure in order for the hard problem to have no solution is just that the observation of the universe is made from an 'outside' perspective, i.e. from a third person perspective. If the observers inside any possible worlds, are able to characterize their environment only in terms of structure and dynamics, then they will necessarily (as long as they are conscious) be facing an insolvable hard problem. But it also seems that there are not many ways in which we can get around the limitations of third person knowledge. For how could we describe objectively any object except in terms of possible (external) relations with other objects? The explanation of phenomenal properties seems to be excluded from the outset.

2.2. Denying the hard problem.

It is perhaps this remarkable characteristic of phenomenal experience that explains why, a few years ago, the most usual answer to any attempt to formulate the hard problem would be just to say that it was a meaningless or unanswerable question. Even today, some authors would still hold that position. For instance Dennett's position, according to Chalmers, is simply the 'denial' that such problem does exist or has any interest whatsoever:

> "The type-A materialist [like Dennett], … denies that there is any phenomenon that needs explaining, over and above explaining the various functions: once we have explained how the functions are performed, we have thereby explained everything. … it is asserted that there is no interesting fact about the mind, conceptually distinct from the functional facts, that needs to be accommodated in out theories." (*idem*)

And this is also the way Dennett sees is one position, both in his reply to Chalmers article (see Chalmers 1995):

> "[we cannot] make progress on the easy questions of consciousness without in the process answering the hard question". (Dennett, 1996: 34)

And in Dennett (1991) and Dennett (1995) we can also find a strong espousal of this perspective. For instance, in Dennett, in his (1995), speaks of 'one author'[56] who has asked if solving all the functional aspects of the mind would tell us if "the mental lights would be out". Dennett's answer is the one expected:

> "there is no work for such a clarification or definition [of mental lights] to do. For suppose that we have indeed answered all the other questions about the mind of some creature, and now some philosophers claim that we still don't know the answer to that all-important question, Is the mental light on – yes or no? Why would either answer be important? [57] We are owed an answer to *this* question, before we need to take their question seriously." (Dennett, 1995: 210)

However, it is clear that the answer would have important consequences: the defeatist position on the hard problem means that the attribution of mental states to any physical organisms cannot be the object of any rational justification. But this is precisely what Dennett wants us to conclude. He wants to convince us that the attribution of consciousness to others must remain forever hypothetical, a matter of definition or convention, or, more precisely, a matter of pure faith:

[56] It seems clear that Dennett should be referring to Chalmers. We should remind that *The Conscious Mind*, by Chalmers, was only published in 1996, whereas Dennett ended writing his *Kinds of Minds* in the end of 1995, so he could not make reference explicitly to Chalmers position (although Chalmers 1995 article appeared in the JCS during 1995). We should also add, perhaps, that the 'inner light' question had already appeared before in Hofstadter & Dennett (1981: 10), although there the authors just say that "this question looks unanswerable" (10).

[57] Dennett's answer is the one expected but the justification for the answer is quite idyosincratic. Most authors would say that it would be useful to have a theory that explained why consciousness is associated with certain functions. Their dismissal of the question as nonsense would be simply because there is no possible answer to it.

"Intentional systems are, by definition, all and only those entities whose behaviour is predictable / explicable from the intentional stance." (Dennett, 1995: 45)[58]

"but are these simple organisms [consciously] seeking or just [mechanically] «seeking»? We don't need to answer that question." (Dennett, 1995: 42)

2.3. Having 'faith' in consciousness.

From the previous result it would seem that the association of conscious states to a system will always involve an arbitrary 'leap of faith' (Chalmers 1996b) – in fact, it would seem that, whatever is our current cosmology, we would have only *evidence* to believe in the existence of our own mental states. In this section we will argue, however, that it is not at all clear why 'faith', 'belief' or 'interpretation' should be given so much credit[59] in the attribution of consciousness.[60]

The notion that the attribution of conscious states is a matter of interpretation, that it is somehow in the eye of the beholder, is very widespread, as we have already seen. And, accepting the argument we have analysed in section 2.1. it might seem that it is an inescapable fact. However, we will try to show here that we really won't have to suppose any act of faith to attribute consciousness to others. Perhaps we are true be-

[58] See also p.128. Notice that, according to Dennett, intentional systems include chess playing computers but also thermostats: see Dennett's (1981)

[59] Perhaps we should do better to say 'belief', but as we will see, the choice was not random. For instance Chalmers (1996b) talks about an inevitable 'leap of faith'in what regards the acceptance of bridging principles: "In a sense, in relying on these principles we are taking a leap into the epistemological unknown. Because we don't measure consciousness directly, we have to make something of a leap of faith. It may not be a big leap, but nevertheless it suggests that everyone doing this sort of work is engaged in philosophical reasoning. Of course one can always choose to stay on solid ground, talking about the empirical results in a neutral way; but the price of doing so is that one gains no particular insight into consciousness."

[60] Chalmers declares for instance: "[the bridging principle] is not itself experimentally testable, at least from the third-person viewpoint; instead it acts as a kind of prior background assumption. … [they are not] experimental conclusions. … These principles effectively precede any experimental results." (237)

lievers in some other respects, but, or so we will argue, attributing consciousness to others is not one of them. To see this more clearly, it is perhaps noteworthy to make a small list of scientific disciplines that are studying conscious processes today:[61]

> **Amnesia**: In cases of amnesia the ability to learn / train behaviour is maintained even without awareness of the past learning process or of the behaviour effectively learned. The evolution of learning is comparable to non-amnesic subjects. This distinction between ability and awareness is also present in other memory impairments like prosopagnosia in which the inability to recognise subjects at a conscious level is contradicted by the involuntary responses at the body level (for example by differences in skin conductance) when in the presence of familiar faces.
>
> **Acquired Dyslexia**: A subject unable to distinguish meaningful from non-sense words on a conscious level may distinguish them correctly (in suitable conditions) if asked to discriminate them by chance. The same applies to more specific tasks: "[using forced-choice responding] he could say whether the written name of a country belonged to the inside or outside of Europe, whether the name of a person was that of an author or politician, or whether the name of an object was living or non-living" (Weiskrantz, 1997: 27.). This dissociation also occurs in different kinds of aphasia, regarding both semantic and syntactic content (for references see Weiskrantz, 1997: 27-29)
>
> **Blindsight**: Ability to behave as if the external stimuli was detected but with no (visual or any kind of) awareness of the stimuli (in fact this ability to react appropriately was so good that the unawareness passed unnoticed in the first studies of monkeys in which the V1 was ablated in monkeys). More recently, it was found out that the identification of affective expressions could also be made without awareness, which was dubbed affective blindsight.
>
> **Blind Touch** (or numbsense): ability to react to tactile stimuli of which the subject is not aware of the stimuli. For instance a subject is able to 'guess' correctly the location of a tactile stimulus on a numb arm.
>
> **Deaf Hearing**: not much explored. Other literature shows that familiars of the patients sometime describe involuntary reactions to sounds in otherwise complete cortical deaf patients, but these cases were not experimentally studied.
>
> **Unilateral neglect**: Although the subjects deny any awareness of objects in their neglected field, they can nevertheless guess correctly about some of their properties.
>
> **Anasognosia**: In all the previous cases the "subject may not 'know' it, but some of part of the brain does" (Weiskrantz, 1997, p.26). In this case, there is still a dissociation between function and awareness, although we observe the inverse relation: the subject consciously maintains he continues to possess an ability that he no longer maintains (for example in blindness or paralysis).

Now, if Chalmers (and Dennett's) perspective is correct, if there is really a leap of faith here, then there must be several ways of interpreting these experiments. Like Chalmers says (1996b):

[61] This list is based on Weiskrantz (1997) and was previously presented (with some modifications) during an oral exposition about blindsight to professor S. Thorpe. (We have also taken off the references so as to not overcharge the "References" section of this paper.)

> "we don't have a consciousness meter, and there seem to be principled reasons why we can't have one. Consciousness just isn't the sort of thing that can be measured directly. So: What do we do without a consciousness meter? How can the search go forward? How does all this experimental research proceed?
>
> I think the answer is this: we get there through principles of *interpretation*. These are principles by which we interpret physical systems to judge whether or not they have consciousness. We might call these *pre-experimental bridging principles*. These are the criteria that we bring to bear in looking at systems to say (a) whether or not they are conscious now, and (b) what information they are conscious of, and what information they are not. We can't reach in directly and grab those experiences and "transpersonalize" them into our own, so we rely on external criteria instead." [62]

But both the detailed exposition of these experiments and the precise way in which consciousness entered the experimental field (in the case of blindsight, for instance), would show that precisely the opposite happened. That is, people like Weiskrantz started by interpreting the experimental results in any way they could to avoid having to deal with consciousness, but they simply were unable to do so. When a subject denies having any conscious experience of a visual scene, but nevertheless can guess accurately several of its properties, there is simply no way of stating the results in a coherent fashion without recurring to consciousness and the distinction between conscious and unconscious perception, etc. If these last decades of neuroscience research have proved something was that we simply cannot make sense of the data without involving consciousness. It was not a choice, it was an imposition; consciousness, in

[62] For an account of the relations between bridge laws and NCC see Chalmers (1996: 238-9)

the XX century science, has entered forcefully, through the back door, and with no help from the researchers.[63]

Notice that there is one sense in which consciousness *could* be denied. If we supposed that in every scientific experiment regarding blindsight (for instance) all the patients would, with no apparent reason, systematically deceive the experimenter. This is, of course, a logical possibility, but it is also irrelevant. It is the same kind of possibility as presuming that the sun will not rise tomorrow, that all laws of nature will change after this work is delivered, that solipsism is true, or that only two persons and a kangaroo are really conscious in the entire universe. Even if these hypothesis are all logical consistent with our experience of the world, but all of them presuppose a variant of systematic deceiving, they are also incredibly implausible, and they carry with them an immense cost in explanatory capability. In practice, they simply are not worth considering. Denying consciousness, in the context of current neuroscience, is not a reasonable option, it would carry to the same epistemic desert as solipsism or the belief that the world is going to end tomorrow. It is a possibility, but not a possibility that we can accept.[64]

---

[63] For a detailed exposition of how consciousness has become a major field of research in the last decades see Weiskrantz (1997), for a more general but original and informative survey on neuroscience see McCrone (1999).

[64] Chalmers response would probably be that neuroscientists could simply study the verbal reports of subjects suffering from blindsight making reference to cognitive abilities only (what Chalmers idiosyncratically calls 'awareness' in opposition to consciousness (1996: 28-9). But this is simply not possible to merge with current scientific practice. In blindsight experiments, for instance, a standard procedure is to show a specific patient suffering from blindsight if he consciously sees a moving spot of light. From previous experiments we know that that particular patient (usually a patient called GY) is able to guess correctly some results (for instance the orientation of the moving spot) and, when the speed of the spot passes a certain limit, the subject is able to report that he is not only guessing but knowing the correct answer (i.e. he knows he is giving the correct answer), or, with even greater speeds, he says he is able to see the moving spot. And we can ask the patient to distinguish between experiments in which he just guesses (correctly, most of the times), knows or sees the spot. Now, to interpret this experiment in terms of 'awareness' (in Chalmers sense) would be as difficult as interpreting our wife's reactions at breakfast as cases of 'awareness'. We would be back to

If this is roughly true then the sense in which attributing consciousness to others is a question of faith is the same in which believing that the world has natural laws is a question of faith, or that tomorrow my toaster will not suddenly get alive and pursue me throughout the house is a question of faith. The point is, we have no reasonable choice in either of these options.

What we have tried to show so far is that, in current neuroscience experiments, if we find that certain physical states are associated with human verbal reports (from different subjects) that systematically describe conscious experiences, then we simply cannot avoid the inference that they are associated with conscious experiences. (This would amount to a claim as implausible as solipsism.) This is not a matter of belief but can be seen as a general methodological principle of always choosing the simplest explanation (in this case it is simpler not to think that every subject is systematically deceiving the investigator). In any case, we simply don't have a reasonable option.

2.4. Defining consciousness.

In this section we will argue that, although conscious experiences are ineffable, there is, in principle, no reason why an objective definition of consciousness cannot be found. We will also argue that the hard problem would remain equally unsolved whatever the definition found. This, we think, will provide a more accurate characterization of the hard problem, and will allow for the presentation, in the next section, of a revised version of the hard problem.

It seems undeniable that there is no consensual definition of consciousness, at least for now.[65] Chalmers, for instance, says that "Trying to define consciousness is just

---

solipsism once again, which, of course, is not an option. It was precisely the fact that researchers could not interpret their results in terms of 'awareness' that has pushed 'consciousness' into the neuroscience field. It wasn't a question of philosophical speculation. By the contrary, on this respect, philosophy is still trying to accompany neuroscience results.

[65] We think this is a trivial remark, Chalmers (1996) starts is introduction by noticing the absence of a suitable definition of consciousness. The divergences between the authors are found in the consequences they are prepared to infer from the ineffability of qualia (from the falsity of materialism (like Nagel 1974 and Jackson 1979) to the denial of conscious experience (see for instance Dennett, 1991: 411, 406).

fruitless." (4) Although he presents a bunch of expressions that are intended to suggest the phenomenal character of experiences. For instance, Chalmers says that "a being is conscious if there is *something it is like* to being that being" (4), and he also speaks about the 'qualitative feel' of experience.

The difficulty in giving a definition of phenomenal experiences in terms of more basic constituents reveals what is generally called the ineffability of mental states. For instance, a person that has not tried a specific experience (like the flavour of anchovies, or someone who was born an achromat) can only understand what is like to have that experience, either by analogies with her past experiences, either by passing through them. Phenomenal experiences are something that cannot be conveyed by language.

However, the ineffable character of experience is not a direct demonstration of the impossibility of a definition, it just shows that the definition, if possible, is not sufficient for conveying a phenomenal experience. In the same way, a definition of energy in terms of mass multiplied by the square of *c* does not convey, or needs to convey, the concept of 'energy', 'mass' or 'time',[66] but it is nevertheless a correct definition.

In natural science a definition of an object can be framed in terms of necessary and sufficient conditions. In general a definition of X is just the conditions which are necessary and sufficient for X to occur. But in the case of conscious experience things get more complicated. Not because it is impossible to find the necessary and sufficient conditions for consciousness to appear but because consciousness seems to involve something extra to the mere establishment of causal relations. Chalmers espouses this point clearly:

> "Although conscious states may play various causal roles, they are not *defined* by their causal roles. Rather, what makes them conscious is that

[66] Einstein's formula is a mere mathematical relation between three symbols, and it would be the same relation if, instead of energy, mass and gravity we would have only x, y and z. To understand its full significance we must associate those symbols with their real meanings (which of course is very difficult).

they have a certain phenomenal feel, and this feel is not something that can be functionally defined away." (105)

We will get back to this quote later on, but what needs to be pointed out now is that this does not mean that we cannot find both the necessary and sufficient conditions for the presence of consciousness. Indeed this is the objective that is at the heart of the search for the neural correlates of consciousness (NCC) and the search for NCC is the centre of Chalmers constructive proposal to approach the hard problem. What Chalmers is denying is that consciousness can be *defined* in terms of NCC.[67]

Another way of putting the subject is by distinguishing between correlation (i.e. finding the necessary and sufficient conditions for consciousness) and explanation (which is connected with a definition of consciousness). Finding a correlation will not explain why the particular correlation holds; at best, it will open the doors to such explanation. To see this suppose we would find that:

a system has physical property X if and only if it has conscious experiences. [68]

---

[67] This, of course, amounts to Chalmers adoption of the dualism of properties according to which consciousness states are fundamental or irreducible to other kinds of explanations.

[68] This could be found in numerous ways. For instance, we know that most cognitive functions are not associated with consciousness. Suppose that whenever we found physical property X in the brain during a cognitive process we would find invariably that it was associated with consciousness, whereas all the processes that developed without property X would go on in the dark (in the same way, for instance that we recognise written words, speech, that we learn how to focus and direct our eyes, to cognitively distinguish an object in a complex scene, etc). Suppose further that we would have a way of disrupting the property X and maintaining the cognitive process, we would then find that processes that usually demand consciousness are, after the disruption, performed without consciousness. By doing this we would find that the person would continue to do the cognitive process but in a kind of zombie state. Or suppose, for instance, that Hameroff's and Penrose's ORCH model of consciousness would prove correct. Then we would find out that, by disrupting the superposition states of microtubules we would also disrupt consciousness, although some automatic cognitive functions might continue to be operational. We could also produce strange consciousness effects, for instance, we could conceivably connect the brain to some kind of apparatus that would enlarge the superposition states to new devices that would enhace conscious experience to new cognitive states, or we could perhaps find a way to enlarge the coherence state to enlarge

Now, as we can remember the hard problem was not defined as the problem of *finding the correlations* but of *explaining the correlations* between physical and mental states, that is, to understand

> "why is the performance of this function associated with conscious experience?" (5)

However, there are two problems here. The first is that, in large part, the need to find an explanation is based on the attempt to avoid the weight that faith, or believe, seems to have in the attribution of consciousness to physical systems. But we have seen that the attribution of consciousness is forced upon us. We simply do not have a choice in the vast majority of cases (at any rate regarding normal human beings). The second is that it is not clear if such an explanation can be found at all (no one has conceived a possible explanation so far).

But there is a more important difficulty. It seems that we are demanding an explanation for consciousness that we cannot find for anything else. For instance, suppose we would have said: OK, the electron can be defined by the performance of certain functions given certain conditions. But now we must ask a further, non-trivial question: how do we explain this association between the electron and these functions. In other words, why does the electron also exist, why not just the functions? This is indeed a very hard problem! But also a senseless one, or at least, a question for which we have no hope of finding a reply.

So, it might *seem* that the hard problem is posing a ridiculous question. And in other passages Chalmers can be read the same way, for instance reconsider this passage:

---

to areas where they are not normally present in the brain so that we could experience, for instance, processes related to th sound processing units, etc. There would be of course a tremendous impact on technology, we could perhaps build computers that used quantum superposition which would allow us to better understand the specific function of consciousness, etc.

> "Although conscious states may play various causal roles, they are not *defined* by their causal roles. Rather, what makes them conscious is that they have a certain phenomenal feel, and this feel is not something that can be functionally defined away.
>
> To see how unsatisfactory these analyses are, note how they trivialize the problem of explaining consciousness. … To analyse consciousness in terms of some functional notion is either to change the subject or to define away the problem. One might as well define "world peace" as "a ham sandwich". Achieving world peace becomes much easier, but it is a hollow achievement." (105)

Compare this to the following passage where 'conscious states' was replaced by 'existing things':

> "Although existing things may play various causal roles, they are not *defined* by their causal roles. Rather, what makes them existing is that they have a certain existential character, and this character is not something that can be functionally defined away.
>
> To see how unsatisfactory these analyses are, note how they trivialize the problem of explaining existence. … To analyse existence in terms of some functional notion is either to change the subject or to define away the problem. One might as well define "world peace" as "a ham sandwich". Achieving world peace becomes much easier, but it is a hollow achievement." (105)

It would be possible to draw a complete parallelism between the question of explaining the existence of consciousness and the more general problem of explaining existence. And it would also be possible to do that in relation to the definition of space (we know that space is associated with these properties, but we cannot explain why!) or time, causality, etc. In general, we get a kind of 'hard problem' each time we get outside the limits of any possible experience, that is to say, each time we consider questions whose answers cannot be given in terms of relations with other objects.

We won't pursue this topic further here though, for reasons of time and space. We will just consider that Chalmers is asking something for which an answer cannot, in principle be given. We can see now how connected this point is with our previous chapter in which we were referring the kind of problems that third person perspectives on the world imposes. According to what we have saw, third person perspectives do not tell us what a thing is, but only how it behaves. A physical theory characterizes all its entities externally, that is, all the properties of elementary particles are properties regarding the set of possible interactions with other particles. To describe what a particle *is* beyond this set of relational properties is just to make a question for which no possible solution can be given – see *supra* our footnote[46].So, Chalmers is in fact asking us to explain why consciousness exists, as he says

> "The first and most central problem is the very existence of consciousness. Why does consciousness experience exist?" (5)

But this problem was not solved for any other entity in the world. We also do not know why electrons exist, or why gravity exists, or, as a matter of fact, why the universe exists. These are perhaps important problems, but why should they apply more to consciousness than to any other entity in the universe? If there is a hard problem, then we might think that it applies equally well to every entity in the universe (including the universe itself). But in fact, this interpretation mischaracterizes the problem, as we will see in the next section.

2.5. Angels and minds.

There is a strong parallel between the angel-world problem and the problem of consciousness. The angel problem arises:

> "from the logical possibility of a world physically identical to ours, but with additional non-physical stuff that is not present in our own world: angels, ectoplasm, and ghosts, for example. There is a *conceivable* world just like ours except that it has some extra angels hovering in a non-physical realm, made of ectoplasm." (39)

What is interesting to us in the angel-world problem is that the existence of angels, although a logical possibility, would constitute a mystery. They would not depend on elementary facts (they would not be composed by them), and they would not be able to act on the physical realm. So their existence, although logically possible, would remain utterly unexplainable.

Now, it seems clear (although Chalmers does not make this comparison) that mental properties are on a same footing then angel properties. Both of them do not logically supervene on the physical and both of them (taking into account the causal closure of the physical) have no causal powers on the world (although, of course, only mental states are epiphenomenal– since only mental states would be caused by the physical).

It is due to this strange circumstance that angels and mental properties exist in the world without being part of the world (at least according to our definition of being part of). It is this special characteristic of mental properties (their causal irrelevance) that turns the hard problem more relevant in the case of mental properties than in the case of physical properties. In fact, it is difficult to ask why is the electron associated with certain functions[69] but it wouldn't be difficult, or awkward, to ask why is this p-electron[70] associated with the functions of the electron. As Dennett points out:

> "Consider, for instance, the hypothesis that there are fourteen epiphenomenal gremlins in each cylinder of an internal combustion engine. These gremlins have no mass, no energy, no physical properties; they do not make the engine run smoother or rougher, faster or slower. There is *and could be* no empirical evidence of their presence, and no

[69] If a function is a causal relation between certain inputs and outputs, then, arguably, an electron can be defined in terms of functions. But our argument does not depend on this claim.

[70] A p-electron, in this text, is an epiphenomenal entity, that is associated with every electron. We know, somehow, that it has the head of a white, smiling horse, but the rest of the body seems just like a siren. It is always drinking chocolate milk, for some reason. However, its existence or inexistence does not make any causal difference and, therefore, it cannot be observed by third person methods.

> empirical way in principle of distinguishing this hypothesis from its rivals: there are twelve or thirteen or fifteen … gremlins. By what principle does one defend one's wholesale dismissal of such nonsense? A verificationist principle, or just plain common sense?" (Dennett, 1991: 403-4)

Gremlins, just like the angel world, or p-electrons, can and must be swiftly dismissed, or else, they must be explained independently of their causal powers. But consciousness cannot be equally dismissed, since it is the very centre of our epistemic world, and, of course, the fact that it cannot be dismissed and does not have any causal role to justify its existence creates the hard problem.

As we can see, from this perspective, Chalmers hard problem regains a perfect sense and importance. But it is a sense and importance that is the result of an epiphenomenalist standpoint. If we get an interactionist[71] account of consciousness then the hard problem will no longer make sense. At least it won't make more sense than when it is applied to any other (physical or mental) entity. In an interactionist picture of consciousness the hard problem simply dissolves into the general problem of finding a reason for the explanation of any entity in the universe.

So the criticism that Chalmers makes to 'new-physics' theories of consciousness simply does not apply. It is true that any explanation in which references to consciousness are made can be replaced by a coherent explanation in which all the phenomenal aspects are subtracted and only the causal relations – dynamics and structure – remain. However, this would imply the kind of irrational interpretation, that is common to views like solipsism and others, as we saw. The reason is simple: suppose we would find that consciousness would appear whenever a certain function (like the orchestrated reduction of a global superposition state in the brain) was present. This would be consistent with verbal reports, with our experience, etc. Now, the only way to deny

[71] We should perhaps say why interactionism, in our view, does not imply dualism. In fact there are at least three different perspectives that allow for interactionism: a special form of emergent properties based on quantum indeterminacy and chaotic dynamics. A dualist view based on the reduction of the wave packet (see Stapp (1999) for bibliography. And parallelism based on Everett's interpretation of quantum mechanics.

this connection would be to suppose some form of systematic deceiving procedure through which all relevant data would be systematically manipulated. In this sense, and only in this sense could we doubt that this function was both necessary and specific for consciousness. So, to argue that the phenomenal component could be subtracted maintaining a coherent physics is, at least, difficult to maintain, since this subtracting would imply a solipsistic monstrous theory that would imply some form or other of systematic deceiving events.

On the other hand, suppose we would say: but this does not explain *why* this specific function is associated with consciousness. We could just say that there is nothing more to consciousness then executing this particular function. (In the case of the ORCH model, we would say there is nothing more to consciousness than just choosing between available alternatives.) Of course a problem would still remain, why wouldn't this function go on in the dark? Here, it seems that Chalmers strategy of considering consciousness an irreducible entity would suffice to shed light on this. We could simply assume (as Chalmers does) that phenomenal properties are simply irreducible to more simple properties and that, by some reason, phenomenal properties have an advantage in making good choices (that is, they should have some evolutionary advantage). In any case their association with the particular function that we'll find consciousness to have will have less probabilities of being unexplainable than if consciousness doesn't play any role whatsoever in the world. If consciousness is part of the world, it should gain her place.

2.6. Functional isomorphs versus *supervenience and time.*

One of the most central arguments in Chalmers (1996) regards the natural supervenience of mental properties over physical properties. Chalmers argues against the natural possibility of inverted qualia, absent qualia, etc. He does that using an argument that is based on functional isomorphs. If the argument is correct it provides evidence in favour of a functionalist account of consciousness. Counterarguments to

Chalmers thought experiment are based on the idea that silicon or computational based components could not reproduce the functions of a biological living brain.[72]

Our argument, by the contrary, is that, although functional isomorphs behave always alike, not all physical systems can be duplicated in a way to make their replicas behave alike, more precisely, only physical systems in which quantum indeterminacy does not play a role can have functional isomorphs.[73] If the brain is one such a system or not is something that current empirical investigation cannot display. To make it short, I'll just give examples of systems that can and cannot be functionally replicated.

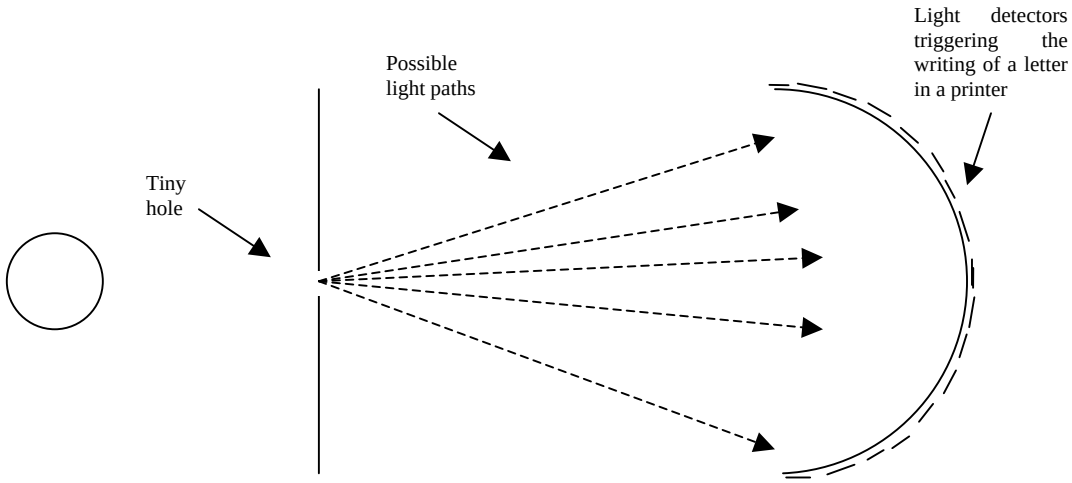
Lets call systems A, systems that can be replicated. Examples of such systems are Turing Machines, digital computers, chaotic systems based on classical dynamics, and, in general, every system that does not involve quantum randomness.

Systems B cannot be functionally replicated. We only know examples of such systems when they depend on quantum processes. Examples of these systems are the Schrödinger cat though-experiment, a quantum random generator, or other quantum triggered devices that depend on quantum randomness. In all these cases, exact replicas of the system will behave differently according to the degrees of freedom available in the system. As a specific example of a B system, I will give one that I've called the 'word random generator'.

---

[72] For counterarguments to functional isomorphs see Lowe (1995), Velmans (1995), Libet (1996), and Hardcastle (1996).

[73] We won't dwell here with the macrophysical effects of quantum indeterminacy. The level where the quantum superposition effects is thought to terminate is a subject not well understood and depends heavily on the particular interpretation one holds of quantum mechanics. However we should notice that Chalmers accepts that quantum superposition effects (when understood in the context of Everett's interpretation) apply to persons; for instance he says that: Take the mind M_1 that I remember being around this time yesterday. Today there will be a large number of minds descending from that mind, in different "branches" of the superposition. My mind M_2 is only one of them. I might as well ask: Why did I end up here, rather than in any one of the other branches?" (353). See also Albert (1992: 75).



A word random generator generates letters using quantum indeterminacy. Although it is possible to create replicas of such a system, no two replicas will behave systematically alike.

Imagine that we get some duplicates of this machine. Suppose each of them has 30 light detectors (connected with letters, punctuation marks and other symbols), that for each photon emission there is a detector that goes off, and that the probability of triggering a given detector is alike for each detector (1/30). Now imagine something like a billion duplicates of this machine. This would mean that after 5 trials we would get a mean of 4 systems with the same result. Now, the fact that any two systems have behaved exactly alike can tell us nothing about their future resemblance. Even if any two systems would have exactly the same behaviour after 30 trials we could not say that the probability that they will resemble in the future is increased over systems that have never behaved alike. And, after a sufficient number of times, we would with all probability found out that no system had produced exactly the same sequence of letters.

Now, we should notice that the existence of functional isomorphs presupposes that the future state of the system supervenes (either locally or globally) on its past state. This is the only way to ensure that identical systems (in identical environments) would evolve in the same way. If two indistinguishable systems and environments would evolve differently its fairly easy to conclude that they would not be functional isomorphs (or, at any rate, we could not guarantee that they would be functional isomorphs in the relevant sense). Now in the case of the random word generator we can easily see that future states do not supervene on past states, because indistinguishable past states (as far as we can see) lead to different future states.

Apparently we could ‘save’ supervenience by adding the constraint that it was not a particular future state that supervene on the past state, but a precise set of possible future states. And, in this case, every identical past state would have an identical set of possible future states. However this does not work for two reasons: first the future never shows a set of possible states but only one observable state. So, even if possible outcomes could supervene on past states of the system, actual outcomes would continue not to supervene (and, plausibly, its the actual states, and not the possible states, that are relevant). On the other hand, introducing supervenience over sets of states strongly weakens the dependence relation involved in supervenience. Suppose we were to say that a set of mental states supervenes on a single physical state. This formulation would obviously be capable of accommodating dualism, so the dependence relation between the mental and the physical would no longer be assured.

A second remark regards the present. Although it seems clear that past events supervene on other more eloigned past events, it is not clear at all that present events supervene on past events. The reason can be found by analogy with future events. We have dismissed supervenience regarding future events because the past allowed for too much in relation to possible future states of the system. But it seems the same reasoning can be applied to present events, because the immediate past only dictates a certain wave function for present events. There is always a certain minimum amount of indeterminacy in present events.

The analogy with our example is simple. When we consider two machines with the same sequence of letters (for instance ‘QWE’) we know that their future state is described by a wave function that allows for 30 possible future states. However, in the meantime between the launch of the photon and its detection, there is a state which has exactly the same theoretical description (the same wave function) but that conduces to different results. But it seems clear that, if the present moment stands before the arrival to the detectors, then the description of the state will be the same (the same wave function), while as, if the present is the moment when the detection is made, then this present moment does not supervene on the past description of the system.

Abstract of the second chapter: In this chapter we have tried to show that Chalmers' hard problem derives from his epiphenomenalism regarding mental properties. We have also tried to argue that only by granting causal powers to mental states can we hope to dissolve the hard problem. More specifically it will be possible to define consciousness only in terms of necessary and sufficient conditions. In the last section we showed that future events do not supervene on past events, which has important consequences for Chalmers' argument based on the concept of functional isomorphs (and is also necessary to argue for an inter-actionist picture of mind and body).

References:

Albert, D. (1992), *Quantum Mechanics and Experience,* Harvard.

Chalmers, D. (1995b), *The Components of Content*, (available in the Internet through Chalmers home page).

Chalmers, D. (1996b), "On the Search for the Neural Correlate of Consciousness", available in the Internet, at Chalmers site.

Chalmers, D., (1995), "Facing up to the Problem of Consciousness", Shear (1997)

Chalmers, D., (1996), *The Conscious Mind: in search of a fundamental theory,* Oxford.

Chalmers, D., (1997) "Moving Forward on the Problem of Consciousness", Shear (1997)

Davidson, D. (1987), "Knowing one's own mind", in *Proceedings and Address of the American Philosophical Association,* 60, 441-458.

Dennett, D. (1981), "True Believers: The Intentional Stance and Why it Works", in *The Intentional Stance.*

Dennett, D. (1996), *Kinds of Minds: Towards an Understanding of Consciousness,* Phoenix.

Dennett, D., (1991), *Consciousness Explained,* Penguin.

Dennett, D., (1996), "Facing Backwards on the Problem of consciousness", Shear (1997)

Deutsche, D. (1997), *The Fabric of Reality: The Science of Parallel Universes and Its Implications.* Allen Lane.

Hardcastle V., (1996) "The Why of Consciousness", Shear (1997)

Hofstadter & Dennett (1981), *The Mind's I – Fantasies and Reflections on Self and Soul*, Bantam New Age Books.

Jackson, F. (1982), Epiphenomenal Qualia, in *Mind and Cognition,* Blackwell.

Jacob, P. (1997), *What minds can do: Intentionality in a non-intentional world,* Cambridge.

Kim, J.J. (1984a), “Concepts of supervenience”, in Kim (1993), *Supervenience and Mind*.

Kim, J.J. (1984b), “Epiphenomenal and supervenient causation”, in Kim (1993), *Supervenience and Mind.*

Kim, J.J. (1987), “Supervenience as a philosophical concept”, in Kim (1993), *Supervenience and Mind.*

Kim, J.J. (1993), *Supervenience and Mind: selected philosophical essays.* Cambridge.

Kim, J.J. (1994), “Supervenience”, in S. Guttenplan (ed.), *A Companion to the Philosophy of Mind,* Blackwell.

Kim, J.J. (1998), *Mind in a Physical World,* The MIT Press.

Lewis, D. (1988), “What experience teaches”, in Lycan, (1990).

Lewis, D. (1994), “Lewis, D.”, in S. Guttenplan, *A Companion to the Philosophy of Mind,* Blackwell.

Libet, (1996), “Solutions to Hard Problem of Consciousness”, Shear (1997)

Lowe, E., (1995), “There are no Easy Problems of Consciousness”, Shear (1997)

Lycan, W. (1990), *Mind and Cognition,* Blackwell.

McCrone, J., (1999), *Going Inside,* Faber and Faber.

Nagel, T. 1974. “What is it like to be a bat?”, in *Mortal Questions.* Cambridge.

Putnam, H., “The meaning of ‘meaning’”, in Minnesota Studies in the Philosophy of Science 7: 131-193.

Russell, J. (1927), *The Analysis of Matter,* Kegan Paul.

Searle, J. (1992), *The Rediscovery of the Mind,* MIT Press.

Shear, J. (1997), *Explaining Consciousness: The Hard Problem,* Bradford.

Stapp, H., (1999), “Attention, Inattention and Will in Quantum Physics”, *in JCS.*

Velmans, M., (1995), “The Relation of Consciousness to the Material World”, Shear (1997)

Weiskrantz, (1997), *Consciousness Lost and Found.*