PHI840: Intuitions and Conceptual Analysis

Week Four: Physicalism and Entailments

1. Overview

So far we've talked mainly about why conceptual analysis is either wrong-headed or useless. For the next few weeks we'll be looking more at the positive roles for conceptual analysis. In particular, we will be looking at Jackson's argument that conceptual analysis plays an essential role in any 'serious' metaphysics. (As we'll see, 'serious' here is a loaded, and technical, term. Whether one should want metaphysics to be 'serious' in this sense is possibly an open question, one that this choice of term seems to skate over.) Quickly, the way conceptual analysis, and thus intuitions about possible cases, gets into the picture, is as follows:

(1) We can tell the complete story about the world by just listing the properties of some privileged parts of the world, e.g. the position and velocity of atoms in the void, or whatever is in true physics, *etc*.

(2) This does not mean that there are no truths about parts of the world not explicitly mentioned in this story. The Syracuse football team is not an atom in the void, but saying that there are just atoms in the void does not imply that the sentence, "The Syracuse football team won on Saturday" suffers from failure of reference.

(3) (1) and (2) are only consistent if we can *analyse* the terms referring to underprivileged parts of the world in terms of the privileged parts.

(4) Intuitions about possible cases play an important role, at least evidential but possibly constitutive, in determining the truth of the analyses in (3).

For the next two weeks, we'll be focussing mainly on (1), (2) and (3), though the importance of (4) should be always kept in mind.

2. Primitive Languages and Enhanced Languages

Part of the motivation for (1) is obviously the success of physics. But another, important, part of the motivation is the fact that parallels to (1) for specific parts of the world are clearly true.

> Imagine a grid of a million tiny spots – pixels – each of which can be made light or dark. When some are light and some are dark, they form a picture, replete with interesting gestalt properties. The case evokes reductionist comments. Yes, the picture really does exist. Yes, it really does have those gestalt properties. However, the picture and the properties reduce to the arrangement of the light and dark pixels. They are nothing over and above the pixels. They make nothing true that is not made true already by the pixels. They could go unmentioned in an inventory of what there is without thereby rendering that inventory incomplete. And so on.
>
> Such comments seem to me obviously right. The picture reduces to the pixels. And that is because the picture supervenes on the pixels: there could be no difference in the picture and its properties without some difference in the arrangement of light and dark pixels. Further, the supervenience is asymmetric: not just any difference in the pixels would matter to the gestalt properties of the picture. And it is supervenience of the large upon the small and many. In such a case, say I, supervenience is reduction. And the materialist supervenience of mind and all else upon the arrangement of atoms in the void – or whatever replaces atoms in the void in true physics – is another such case. (**Lewis**, 294).

To see whether this argument by analogy might succeed, let's try and isolate the crucial features of the case. That is, let's try and see (a little formally) how it might be that such reductionism is possible. Imagine we have two languages for describing a certain set of facts, call them *L* and *L*+. As the names suggest, *L*+ has all the vocabulary of *L*, and more. Nevertheless, it might be the case that what's true in *L*+ supervenes, in a very strong sense, on what's true in *L*. Put another way, there cannot be any variation in what is true in *L*+ without variation in what is true in *L*. As everyone knows, modal claims like this are normally hopelessly vague, just what does 'cannot' mean in general, but hopefully in particular cases we can be clearer.

Example One: Heights

Jackson uses this example quite a bit, starting with the reference on page 5. Let *L* have enough resources to say what everyone's height in inches is. (So, formally, it contains names for every individual and every real number, and a two-place predicate *H* such that *Hxy* is true iff *y* is the height of *x*.) *L*+ contains all that and the predicate 'taller than'. Now, once we know what everyone's height is, we know who is taller than whom. That is, which sentences in *L* are true determines which sentences in *L*+ are true. In fact, for each sentence of *L*+ which isn't a sentence of *L* (that is, for every sentence of the form "*a* is taller than *b*") the truths of *L* either entail that it is true, or entail that it is false. It is *logically* impossible to have variation in *L*+ without variation in *L*.[1]

Example Two: *Positions and Velocities*

This is a bit more controversial (meaning ‘probably false’). *L* has enough resources to state the position of each object at every time. *L*+ has all that plus resources to state the velocity of every object at every time. Now if you think that velocity just means change of position over time, then all the truths in *L*+ will be determined by the truths in *L*. The reason this is controversial is that there are alternative theories of velocity which are possibly (if not probably) right, defining velocity in terms of instantaneous velocity vectors. On this picture it is possible for a particle which never changes position to have a positive velocity for an instant. More importantly, for us, it means that just knowing the position of a particle at all moments of its life does not give us enough information to determine its velocity at all times. So the truths of *L*+ tell us something more about the world than the truths of *L*.

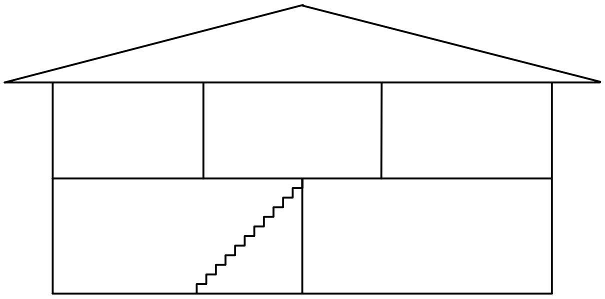
Example Three: *Morse Code and Sound Waves*

Let *L* have enough resources to describe the shape of any particular sound wave, and say that we have a sound wave which is a Morse code message. I think, though I’m not certain about this, that the shape of the wave determines entirely where the dots and dashes are. That is, if *L*+ contains *L* plus enough resources to say where the dots and dashes are, then the facts as stated in *L* determine the facts as stated in *L*+. Of course in this example there should be several other levels of description in which we are interested, such as *L*++ which has resources to say which letters are communicated by which dots and dashes, *L*+++ which has resources to say which words are made of which letters and so on. Now obviously the reduction of dots and dashes to sound waves is a very small step towards a naturalised theory of ‘non-natural’ meaning, it is at least *a* step.

Example Four: *A Doll’s House*

We have a doll’s house made entirely of matchsticks. A cross section of the house looks as follows.

[1] Jackson is assuming here a popular, but wrong, theory of comparative sentences like *Smith is taller than Jones*. The theory is that this should be analysed as *Smith’s height is greater than Jones’s height*. This is wrong for two compelling reasons. First, the analysis if correct should apply to all comparatives, but sometimes we can make comparatives even when there is no salient degree concept nearby. For example, it is true that *Michael Jordan is a better basketball player than I am*. But this does not mean there are degrees of good basketball play corresponding to heights. A similar point applies to ‘smarter than’, ‘smellier than’, and a host of other comparatives. Secondly, the comparative sentence *Smith is taller than Jones* is prior to the degree sentence *Smith is six-foot tall* semantically, and epistemologically, and ontologically, and scientifically, and so on. Precisely how we should analyse degree sentences in terms of comparatives is hard question. An even harder question is how we should analyse comparative sentences (like *Smith is taller than Jones*) in terms of simple sentences like *Smith is tall and Jones is not-tall*, which for similar reasons it looks like we must do. Anyway, Jackson’s false assumption doesn’t lead to him saying anything false (yet) so we’ll let it slide with a footnote for now.



The case should evoke the same kind of reductionist comments that Lewis says attaches to the case of the black and white picture. There are nothing but matchsticks here; the full story about the matchsticks is the full story about the doll's house. And there are five rooms, a staircase and a roof to the doll's house. And neither the rooms, nor the staircase, nor the roof, is a matchstick, though they are all collections of matchsticks. This seems inconsistent, but clearly is not. The story about the doll's house told in terms of matchsticks is *discerning*, it only uses a primitive vocabulary, but it is not *eliminitavist*, it does not deny that there are the rooms and staircase and so on. All it denies is that the rooms and so on are anything over and above the matchsticks.

To link this back to the earlier cases, let *L* contain just enough resources to give the position of every matchstick in the house, and let *L*+ be the full story, with reference to stairs and rooms and roofs. Now in one sense the truths of *L* determine the truths of *L*+, but it is far from obvious that this is the same sense as in the first example. Here, we might want to say that the facts about the matchsticks determine the facts about the rooms, and so on, without saying that this determination is *entailment*.

In order to be clearer about how this determination claim is non-trivial, let's look at two ways it might fail to be true. First, there may be some parts of the doll's house other than those made of matchsticks. If I added a carport made of plastic to the side of the house, the full story about the matchsticks would no longer be the full story about the house. Secondly, there may be ways for houses to vary independently of the variation in the matchsticks. This is a bit harder to grasp intuitively, but the existence of wide content may help. That the staggered arrangement of matchsticks is a staircase may depend not just on facts about the matchsticks, but on facts about typical staircases around here. I don't want to necessarily endorse this, just note that it is a live possibility. Each of these possible failure of covariation have parallels in the philosophically interesting cases we'll now discuss.

Example Five: *The Physical and the Psychological*

This is where all the action is. Let *L* be a language sufficient to describe everything that we talk about in physics, so all the fields and momentums and masses you can imagine. And let *L*+ be ordinary language, so with reference to tables, chairs and beer mugs, to beliefs, desires and intentions, and to inflations, wars and governments. Jackson's claim is that the facts described in *L* determine the facts described in *L*+, without making it the case that there are no facts which can't be stated in *L*. So it is true that there was hyper-inflation in 1920's Germany, that this inflation altered the behaviour of the German people, and that I believe each of these claims, despite the fact that there are nothing but atoms in the void. Like Jackson, we will not be spending much time on the everyday objects, or the social facts, that are not mentioned explicitly in *L*. Rather, we will focus on the psychological. Jackson thinks that the facts described in *L* not only determine the psychological facts, in some loose sense, but that they entail those facts.

There's two ways that this could fail, corresponding to the two ways my story about the matchsticks could have failed to be the complete story about the doll's house. First, it could be that there are entities which are not discussed by physicists. This would be the case if Cartesian dualism were true. Secondly, it could be that minds can vary independently of the variation in their constituent brains. This, roughly, is the kind of dualism David Chalmers has been promoting for the last few years. (Chalmers does not think that there can be variation in minds independently of variation in brains in this world, but he thinks that in certain other salient worlds, there are objects with identical physiology to ours but with different psychologies. These are the famous zombies.)

Now everyone knows why we believe that the physical determines the psychological, in some loose sense. We don't want to be the type of people who believe in fairies, or the like. But the claim that the physical way the world is *entails* the psychological way the world is strikes many people as preposterous. As you'll have seen at the end of the chapter, there is an argument for this, but not one we're ready to state yet.

Entry by Entailment

Assume that *L* does give a complete account of the world (or some subject-matter). Then, according to Jackson, a sentence in *L*+ is true iff it is entailed by the true sentences in *L*. This, in a nutshell, is the entry by entailment thesis. As Jackson notes, one direction of the biconditional is uncontroversial. Anything entailed by the true sentences in *L* is, of course, true, because entailment, whatever else its faults, is truth-preserving. The controversy is the other direction, whether we can exclude any sentence not entailed by the facts as described in a minimal language.

It should be noted that entailment here is being used in a specific, and perhaps controversial, fashion. Jackson takes entailment to be necessary truth-preservation. So *A* entails *B* iff every world in which *A* is true is a world in which *B* is true. This should be contrasted with a proof-theoretic concept of entailment; where *A* entails *B* iff there is a proper proof of *B* from *A*. (So Jackson is using a *semantic*, not a *syntactic*, notion of entailment.)

Jackson's reliance on the possible worlds framework is probably ineliminable. For one thing, he needs to be able to use analytic premises 'for free', which you can do in the possible-worlds framework, but not the proof-theoretic framework. Secondly, he needs to be able to make sense of entailments from infinitely many premises, for he is not, or perhaps should not, asserting that we can give a finite reduction of the language of *L*+ to the language of *L*. Again, this makes perfect sense in the possible-worlds framework (or at least good enough sense for those with few set-theoretic scruples) but little sense at all in the proof-theoretic framework. While the defence of this framework on pages 10-11 is perhaps a little short on rigour, it is certainly entertaining, and perhaps this is enough.

Even within the possible worlds framework, there is another possible sense of entailment from which Jackson wishes to distinguish from his. We might think that *A* entails *B* iff it is knowable *a priori* that every world in which *A* is true is a world in which *B* is true. So *There's gold in them hills* does not entail *There's a metal with atomic number 79 in them hills* on this picture, even though in every world in which the first is true is a world in which the second is true. The thought is that although *Gold has atomic number 79* is a necessary truth, it is not *a priori*, so it cannot be relied upon in *a priori* reasoning. As it turns out, Jackson thinks that the interesting entailments often are *a priori*, but he wants to leave the argument for that claim for further in the text.

3. Defining Physicalism

A large chunk of chapter one in Jackson's book is a discussion of the definition of physicalism. Now this may seem like a rather arid dispute. It can be fun, and useful, to try to find the meaning of terms in natural language. But why should we care about the meanings of theoretical terms like 'physicalism'. The whole point about theoretical terms is that it is fair to stipulate their meaning, isn't it? One reason for thinking otherwise is that there is some common motivation we are trying to find that is behind physicalism, and the dispute is not over the meaning of the word 'physicalism', but of the precise formulation of this intuitive idea. The intuitive idea seems to come in two parts. First part: physics is complete about what individuals there are, so no spooks. Second part: physics is complete about what properties these individuals have, so no zombies. This is even rougher speaking than Lewis's formulation on page 34of **Lewis**, let's follow Lewis's attempts to be more precise. Note that Lewis uses the word 'materialism' for what we, like the majority, call 'physicalism'.

Lewis's Attempts to Define Physicalism

(M1) Any two possible worlds that are exactly alike in all respects recognised by physics are qualitative duplicates.

This is clearly false, because it claims something like physicalism is true at all possible worlds. This is a fairly implausible thesis. An important assumption, apparently part of physicalism, is that there could be a world just like this one which did contain Cartesian souls. More prosaically, Ockham's Razor is taken to be an epistemological, not an ontological, thesis. In any case, imagine a world in which there are Cartesian souls, and these are the repositories of beliefs, desires and the like. If Jack's belief that *p* is irrelevant to his action, then this world has a twin which is a physical duplicate, in which Jack does not believe not-*p*. Hence (M1) is false. But the impossibility of souls is no part of physicalism, so (M1) doesn't capture what we intended.

(M2) There is no difference, *a fortiori* no mental difference, without some nonmental difference. Any two worlds alike all nonmental respects are duplicates, and in particular do not differ in respect of the mental lives of their inhabitants.

Lewis gives two slightly subtle reasons for denying this, but in fact it fails for the same simple reason that (M1) fails. The two worlds Jack inhabits need differ in no nonmental respect whatsoever, if Jack's belief that *p* has no effect whatsoever on his nonmental respects. So imagine a possible world in which Jack has a soul, and the truthmaker for *Jack believes that p* is a sentence in Jack's language of thought. There seems to be no nonmental difference between this world and the world where a different sentence is inscribed in Jack's (non-physical) belief box. So (M2) fails because of esoteric facts about far away possible worlds. This can't be the definition we intended.

(M3) No two (Physicalistic) worlds differ without differing physically; any two (Physicalistic) worlds that are exactly alike physically are duplicates.

Normally if a definition looks circular, it is circular. And this is no exception.

(M4) Among worlds that conform to the actual laws of nature, no two differ without differing physically; any two such worlds that are exactly alike physically are duplicates.

This will be true if it turns out there are souls whose behaviour is nomically correlated with the physical world. In that case, among worlds conforming to those laws, there would be no difference without physical difference. But intuitively physicalism would be false in such a world. And this is not such an outlandish position; it is a variety of epiphenomenal dualism, as endorsed by (some time-slices of) Jackson and Chalmers.

(M5) Among worlds where no natural properties alien to our world are instantiated, no two differ without differing physically; any two such worlds that are exactly alike physically are duplicates.

The first important point to note about this definition is that it uses the 'nothing extra' characterisation of physicalism that Jackson will adopt in chapter one. The second is that it uses Lewis's notion of a 'natural' property. This is described, loosely, in the early parts of the paper. The important point for us is that the natural properties provide a "minimal basis for describing the world completely." (pg. 12) And not just this world, since it is reasonably clear that the naturalness of a property is an *a priori* matter, the natural properties must provide a minimal basis for describing *any* world completely. (These two projects, providing a *minimal* basis for describing this world and describing a minimal basis for describing *any* world do seem to be contradictory, don't they? I suspect there is some work to do here.) Finally, an alien natural property is a natural property not instantiated at this world.[2] So the thought is that any spirit*y* world is one where there are alien natural properties instantiated, because there must be some natural properties in terms of which the spirits can be characterised instantiated there.

Well, maybe this will work, but it is still a schema. At present we have said nothing about how the physical properties are to be divided from the non-physical. It is at this point that Crane and Mellor's criticisms are directed.

Crane and Mellor on Physical Properties

Crane and Mellor think that there is no way to demarcate the physical properties in a way which makes physicalism both non-trivial and true. Now in part this is because they mean something different by physicalism to what we (and most everyone else) means by it. (This really is one case where a difference in use really is a verbal dispute, not a dispute over the way the world is.) For one thing, they say, "chemistry, molecular biology and neurophysiology are … indisputably physical sciences." For another, they imply that physicalism means that there are no "atoms, molecules, tables, trees or tennis rackets, figs or fast food restaurants—or animals or people with minds."

On our account of physicalism, neither of these claims is true. First, it seems we can give a complete account of the world without reference to molecules, cells or neurons, so these need not be part of the physicalistic base. Secondly, as noted in (2), physicalism does not mean that there aren't any fast food restaurants (what a horrible possibility!) but merely that we can tell a complete story about the world without explicit reference to them.

What has probably happened here is that Crane and Mellor have confused the safe, trivial, doctrine of physicalism with the dangerous, radical doctrine of eliminitavism. Eliminitavists, like Stich and the Churchlands, agree with (1), but disagree with part of what is implied by (2). They hold that science shows ordinary talk not merely to be eliminable in principle, but false. According to the physicalist, as we've been conceiving of her, beliefs

[2] Lewis goes into much detail about how alien properties can't be constructed out of other properties by Boolean operations, but this seems redundant, since it seems that natural properties can't be constructed by Boolean operations on other natural properties.

are nothing over and above the pattern of neuron firings science discovers. According to the eliminitavist, the pattern of neuron firings science discovers shows that beliefs are nothing at all, that beliefs aren’t there.

Strictly speaking, this form of eliminitavism should be referred to as eliminitavism about the (folk) psychological. For there are other parts of folk ontology that we might want to eliminate, and the name should perhaps be reserved for the general position. For example, van Inwagen and (recently) Merricks have argued for eliminitavism about medium-sized dry goods, like fast food restaurants. But Stich and the Churchlands appear to have won the name, and I’m generally happy to stick with popular usage.

There’s some interesting questions here which I’m either going to glide over or defer. First, if you happen to agree with Stich, the Churchlands, van Inwagen and Merricks, that some items in folk ontology exist and some don’t, you have to give a reason for the distinction. Crane and Mellor provide some reasons for thinking some reasons usually provided here will fail. Since physicalism as we’ve construed it isn’t eliminitivist about very much at all (well, perhaps phlogiston and witches) this isn’t a problem for physicalism, today’s topic. Secondly, there is an issue about how we can keep talking about cells, beliefs and fast food restaurants if there is nothing but atoms in the void. Much more on this to follow.

Thirdly, there is an issue as to why we buy all of physics, with its multitude of parts, but not the other sciences. The reason, which will be elaborated on below, is that physics with all its parts gives a complete story of the world, but remove one of the parts and this is no longer the case. Should it turn out that one of the parts is removable in this sense, physicalists would be happy to remove it. (And of course reinstate it immediately along with the cells, beliefs and fast food restaurants.)

Bracketing this error, Crane and Mellor raise an interesting challenge for the physicalist. Physical properties cannot refer to those properties in current physics. For that would mean physicalism is almost certainly false, because physics isn’t yet finished. And it cannot refer to those properties in ideal science. For that would make physicalism trivial, and we think it is an interesting, *contingent*, claim about the world. So how can it be defined?

Digression on Supervenience and Indeterminacy

One way of stating the physicalist position is as a supervenience thesis: no variation without physical variation. Crane and Mellor claim that the indeterminacy found in modern physics shows that this thesis is false. The argument is really badly flawed, so we might as well stop to mention this. (The first argument, relying on the ‘intuition’ that we could have intrinsic duplicates standing in exactly the same external relations while having thoughts with distinct content, is so bizarre that I will barely stop to mention it. That is on page 204.)

The argument appears on the bottom of page 205. The reasoning appears to run as follows. By indeterminacy, we can’t say for certain what the physical effects of the existence of a certain physical state will be. So we certainly can’t say what the non-physical effects will be (where non-physical refers to anything not mentioned in the austerely physical story). But this means we can have non-physical variation without physical variation.

The problem is that Crane and Mellor have confused causation with constitution. Atoms in the void don’t (just) cause me to have certain beliefs, they *constitute* my beliefs. And constitution can be determinate even when causation is indeterminate. A simple example should show this.

Imagine a very small pinball machine, where the ball is just one electron, and there are no flippers (like in the 1920’s machines). Sub-atomic creatures play such machines for fun in their rather short (by our standards) lives. The machine has various walls from which the electron can bounce, and a faint positive charge at the bottom so the electron normally drifts downwards, but its all very indeterminate. Even if, *per impossible*, we knew the position and velocity of the electron at any one time, we would not be able to determine its future position, because of quantum indeterminacy. However if we know the electron's current position, we can tell whether the ball has gone out of play. That is because 'going out of play' isn't caused by the electron being in a certain position, it is constituted by the electron being in that position.

The same holds of more large-scale examples. A soccer game is (probably) an indeterministic system, because the players are (probably) quantum machines. From just a knowledge of the current game setup, we can't, even in principle, work out when the next goal will be scored. However from a complete physical description of some part of the game, we can tell whether goals were scored in that time; because the physical nature of the game doesn't cause certain goals to be scored. Rather, what it is for goals to be scored is for some physical state or other to be instantiated. This relation can be determinate even when movement of the soccer ball is not. *End of Digression.*

Jackson's Definition of Physical Properties

In order to solve the puzzle set by Crane and Mellor, Jackson offers three accounts of what it is for a property to be physical. These are:

(a) Physical properties are the properties needed to give a complete account of paradigmatically non-sentient things;

(b) Physical properties are something like the properties found in current physics; or

(c) Physical properties are those needed to give a complete account of everything below a certain size.

As he notes, (c) does not commit us to the view that the *intrinsic* properties of the very small can be used to provide a complete account of the way the world is. It might be that we need to appeal to the irreducibly relational properties of the very small in order to do that. And physicalists are, apparently, allowed at this stage to be quite agnostic about what size the 'very small' happens to constitute. Without being an expert on quantum mechanics, I'm not sure quite what to say about the various technical problems this approach might face.

Jackson's Definition of Physicalism

With this definition of physical properties, Jackson proceeds to give his definition of physicalism. Like Lewis, he starts with something like (M1), and works on alterations to it. The first alteration is the obvious one of making physicalism just a claim about this world, not about all worlds. So instead of saying something like "All physical duplicates are duplicates," we say, "Any physical duplicate of this world is a duplicate of this world." Lewis is a little more bold – his claim is of the form "Any physical duplicate of a world in *M* is a duplicate of that world," where *M* is a class of worlds which are built of the same basic ingredients as this world. A physical duplicate of a world *w* is a world which is just like *w* in its distribution of physical properties.

The second amendment Jackson makes is a little more interesting. He notes (or stipulates) that it is consistent with physicalism that there is a possible world just like this one except that it contains some non-physical stuff. But that world is a physical duplicate of our world without being a duplicate of it. Hence the claim is merely that *minimal* physical duplicates of this world are duplicates of this world. If you built a world by duplicating this world's physical properties and doing nothing more, you would have built an exact replica of this world, replete with all the psychological features of this world. Given this, the definition of physicalism is as given on page 12.

(B) Any *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world.

There is a problem with this definition if there are *necessary* connections between the physical way the world is and the non-physical way the world is. To illustrate, assume there are Cartesian souls in the world, but it is metaphysically necessary that these go where certain physical objects go. If that were true, (B) would be true but materialism false. I can see no good reason for believing this, so perhaps it isn't a major problem for Jackson. (But then if we never discussed positions I could see no good reason for believing, many courses would be quite short.)

Kantian Physicalism

Pages 15 to 24 are concerned with tidying up loose ends in the definition of physical properties and physicalism, and mostly aren't of concern to us. However one point does seem worth spending a little time on just because it seems to be a qualification of the definition of physical properties.

Probably, it is improper to have different opinions about two objects which enter into *exactly* the same causal relations. This is because our perceptions of the object, no matter how indirect, are determined at most by the causal relations into which the object enters. So if there were two different properties *P* and *P**, such that having *P* had the same causal consequences as having *P**, then we probably should have the same opinions about things with those properties. Anyway, whether we should have different opinions or not, we most certainly would. Jackson thinks this possibility is compatible with physicalism. He dubs it Kantian physicalism. He then makes this rather odd claim about the upshot of Kantian physicalism.

> If Kantian physicalism is true, some minimal physical duplicates of our world differ markedly from our world in intrinsic nature, but not in ways that the inhabitants of those worlds know about.

Now any world which differs markedly from our world in intrinsic nature is not a duplicate of this world. So it seems like Jackson has contradicted his claim that Kantian physicalism is a kind of physicalism. Something has to be given up; what is the best option? I think that will depend on which of Jackson's three proposed definitions of physical property we accept.

If we accept the first or third definition of physical property, then this supposed consequence of Kantian physicalism will not really hold. The properties needed to give the full story about the non-sentient, or the very small, will include the unobservable intrinsic properties which ground the causal properties. Hence any physical duplicate of our world will duplicate these unobservable properties. If we accept the second definition of physical property, then the conditional mentioned here will be true. But now it is false that Kantian physicalism is a kind of physicalism, *on Jackson's definition*. The whole point about Kantian physicalism is that, if it is true, we could produce a world which was just like this one in terms of the causal relations between entities, but which was quite different in terms of the intrinsic properties of individuals. This is clearly inconsistent with Jackson's definition of physicalism; I leave it to you to judge whether this means Kantian physicalism isn't really a kind of physicalism, or whether Jackson's definition of physicalism is incorrect.

Historical Digression

Jackson was at one time an *opponent* of physicalism. He put forward a famous thought-experiment designed to show that physicalism is false, concerning Mary, a scientist who has learned all there is to learn about colour vision while trapped in a black-and-white room.[3] At some stage in the late 1980's or early 1990's he decided that this position was untenable, and converted to being a physicalist. The reasons for his change of heart aren't entirely clear, but some of them are given in the attached note from *Mind, Method and Conditionals*, a collection of papers published in 1998.

4. *The Entry by Entailment Thesis*

Jackson's definition of physicalism is kind of interesting for those interested in that particular debate. But what is much more interesting is the reliance on supervenience theses to capture the physicalist's claim, and the implication of this role for supervenience theses for the entry by entailment thesis.

Given the possible-worlds account of entailment that Jackson is using, the path from (B) to the claim that the physical way the world is entails the psychological way the world is runs very smoothly. Jackson uses Φ and Ψ for long sentences which tell, respectively, the complete physical story of the world and the complete psychological story of the world.[4] It can be quickly shown that given physicalism, Φ entails Ψ.

(i) Assume Φ is true at some world *w*.
(ii) By the definition of Φ, *w* is a minimal physical duplicate of this world.
(iii) By the assumption of physicalism, *w* is a duplicate of this world.
(iv) Any proposition true at this world is true at duplicates of this world.
(v) So since Ψ is true at this world, and *w* is a duplicate of this world, Ψ is true at *w*.

Hence given physicalism, any world in which Φ is true is a world in which Ψ is true. And by the definition of entailment, this means Φ entails Ψ. The only assumption used in the argument was physicalism, so this means physicalism implies that the physical story about the world entails the psychological story about the world.

[3] "Epiphenomenal qualia" *Philosophical Quarterly*, 32 (1982), 127-36; "What Mary Didn't Know" *Journal of Philosophy* 83 (1986), 291-5.

[4] *Technical Digression*. I don't think the idea of a sentence with uncountably many parts makes a great deal of sense, but the rephrasing necessary to avoid this problem is rather trivial. Instead of having Φ and Ψ name sentences with uncountably many parts, let them name sets with uncountably many elements, something which does make sense, where each of the elements is a sentence about, respectively, the physical and the psychological way the world is. (As noted on page 26, Φ will have to include a "That's all folks," sentence, which poses troubles of its own.) We now need a concept of entailment between sets, but this isn't too hard to build. Say that a set *A* entails another set *B* iff every possible world where every element of *A* is true is a world where every element of *B* is true. Now Jackson's argument can go through almost as is, without the questionable reliance on uncountably long sentences.

What is really spectacular about this result is how easily it generalises to other cases. Any time we say (a) the story about the world told in terms of *X* is complete, and (b) we cash out that notion of completeness in terms of supervenience theses, we are committed to there being an entailment from the full story of the world told in terms of *X* to any true sentence about the world. If we had thought that there would be no true sentences that could be stated using language outside *X*, this would be a trivial result. But if we think that metaphysics, or physics, or social deconstruction, or can reveal a complete picture of the world without showing all rival pictures to be *false*, Jackson's argument will go through.