

SUPERVENIENCE, PHYSICALISM AND EMERGENCE

Richard Campbell

Physicalists appeal to supervenience as a way of expressing their view that everything in the world is 'nothing but' physical. But what does this mean? In general, the accounts offered of supervenience follow one of two routes: either some global definition, or some form of micro-reduction. Both are inadequate. I will consider each in turn, and then argue a) that the germ of truth in the idea of supervenience is more adequately expressed in an ontology that admits genuine emergence; and b) the denial of micro-reduction requires an ontology that gives priority to the relational characteristics of fields in process.

'Supervenience' has become a trendy term in contemporary philosophy because it seems to offer a way of maintaining that everything in the world is 'nothing but' physical whilst allowing for the richness of non-physical phenomena in the world. In recent times, the issue of the status of properties and causal powers that do not clearly and obviously result from basic physical constituents has been addressed by philosophers largely in the context of 'the mind-body problem'. If psychological phenomena can be accommodated within a physicalist metaphysics, all would be over bar the shouting.

But how to do that has proved a problem. Earlier in the 20th century Rudolph Carnap had boldly proclaimed a program to *translate* psychological descriptions into physical language, but even ardent physicalists now concede that that program is hopeless. Another way of expressing the dependence relation would be to say that all non-physical properties and powers can be *explained* in terms of physical properties and powers. But the prospects of doing that in all but a few cases also appear rather dim; in the absence of plausible explanatory mechanisms most physicalists now regard the demand for such explanations as too strong.

A more promising approach in circulation since the 1960s took off from the fact that even if mental descriptions cannot be *translated* into physical descriptions nevertheless it does not follow that the former cannot *refer* to physical states. To cite one of the favourite examples of this new version of physicalism, water and H₂O are said to be identical even though those terms are not synonymous. Likewise, it was proposed, mental phenomena can be identical with states and processes in the brain, even though psychological descriptions have meanings different from physical descriptions. Translations were no longer necessary; co-reference is all that a physicalist ontology needs.

Still, such a 'contingent identity' thesis seemed to lack something. How is it that non-physical phenomena are somehow *dependent* upon their physical base? Here the idea of *supervenience* began to look attractive. For it seemed to provide an intelligible answer. Accordingly, the concept of supervenience was borrowed from moral philosophy, and invested with metaphysical significance. (Or more accurately, it was borrowed back, but with a reversal in its sense – C. Lloyd Morgan, one of the school of 'emergentists' in Britain earlier in the 20th century, used the term to mean the return action of some emergent entity upon the lower level events from which it

arose.¹) The standard example cited to explain the moralist's use of the concept is that paradigm of a good man, St Francis. It is not logically possible, it was held, that there be another man placed in exactly the same circumstances, and who behaved in exactly the same way, as St Francis, but who differed from him only in the respect that the former was *not* a good man. That is, no difference in moral properties can properly be ascribed unless there is a corresponding difference in descriptive, or non-moral properties. This is another way of saying that people are called 'good' *in virtue of* the properties that describe how they are. While no moral property can be *deduced* from purely descriptive properties – these philosophers were not disposed to challenge the Humean principle that “ought” cannot be derived from “is” – properties like goodness were said to *supervene* on non-moral properties. Likewise, physicalists seized on the idea that non-physical properties and powers ‘supervene’ upon physical properties and powers.

If that physicalist hunch is to be at plausible, the definition of supervenience will have to render precisely how some non-physical phenomenon supervenes upon its physical base. In general, the accounts offered follow one of two routes: either all the non-physical properties in the world supervene upon the world's physical properties taken as a whole ('global supervenience'), or some form of micro-reduction, which posits the properties of each non-physical phenomenon as supervening upon *its* physical base. Let me consider each in turn.

The attraction of global supervenience is that it *seems* to provide for apparently emergent phenomena in a way consistent with physicalism. And it allows that non-physical phenomena might arise from interactions between various physical states without having to trace each phenomenon down to its micro-constituents. Therefore, it can accommodate the fact that such phenomena persist even when there is considerable turnover in their physical constituents.

Since the point of a global definition of supervenience is to avoid going down the micro-reductive route, it seems the only alternative is to invoke possible worlds. A popular way of expressing this idea goes: if two possible worlds that exemplify natural laws are exactly alike with respect to fundamental physical facts, then they are exactly alike with respect to all other facts.²

This definition of supervenience *sounds* suitably physicalist, since the way it is formulated makes it seem as if all the non-physical truths in our world are determined by the set of all the physical truths. That, no doubt, is the intention. But this use of “determined” is ambiguous. The thesis can be interpreted quite weakly, as meaning no more than that the truth or falsity of all propositions co-varies with the truth or falsity of the set of physical propositions. Or it can be interpreted as making a stronger causal claim, that the physical facts in our world somehow *bring about* all the non-physical facts in it.

The latter, stronger interpretation seems the closer to the physicalist hunch. It accords with an alternative way of expressing the physicalist doctrine, namely, that it

¹ See C. Lloyd Morgan: *Emergent Evolution* (Williams & Norgate, London, 1923). Its more recent use is traced by Jaegwon Kim: 'Supervenience as a Philosophical Concept' in *Metaphilosophy* 21, 1990, pp. 1-27, rep. in *Supervenience and Mind* (Cambridge UP, Cambridge, 1993).

² See, for example, E. LePore and B. Loewer: 'More on Making Mind Matter', *Philosophical Topics*, 17, 1989, pp. 175-191.

is the properties of the ultimate physical constituents of things – the ‘basic particulars’ – that causally determine the properties of everything else. But many physicalists would find this way of defining their position too strong, because it builds requirements about explaining supervenience – or explaining how higher-level properties are ‘micro-based’ – into the definition of physicalism itself.

Yet if all the proffered definition is claiming is co-variance, it is too weak. It can happily be accepted by a non-physicalist. For if the truth or falsity of all *non-physical* propositions in our world co-vary with the truth or falsity of the set of *physical* propositions in any duplicate world, that could be endorsed by a certain kind of idealist. I have in mind someone who holds the converse position to physicalism, namely, that it is the non-physical facts that are ‘basic’ and ‘determinative’. Such an idealist would presumably hold that all the so-called physical facts are somehow determined by the non-physical facts, by some process of ‘downward causation’.³ This imagined non-physicalist position would satisfy the definition of supervenience, understood in the weaker sense. Yet such a position is surely incompatible with physicalism. So, this way of articulating supervenience does not suffice to define physicalism.

Perhaps we were wrong to interpret the definition as positing co-variance; after all, it posits only a one-way implication: that if a possible world is a physical duplicate of some world, then it is a duplicate of that world in all respects. True, but nevertheless the definition does not *exclude* the position of our imagined idealist. For, to mimic the language of the definition we are considering, such an idealist affirms that if a possible world is a non-physical duplicate of some world, then it is a duplicate of that world in all respects. And that thesis is *compatible* with (although of course it does not *satisfy*) the purported definition of physicalism. But surely physicalism is incompatible with such an idealism. Therefore, that definition will not do as an articulation of the physicalist hunch, since it admits of an interpretation that fails to exclude one of its major metaphysical rivals.

Maybe this definition could be saved by switching the primary notion from ‘determination’ to that of entailment. Commitment to entailment is not commitment to any particular explanation or justification of that entailment; the reasons why some sentence entails another surely vary. So, a physicalist might say that supervenient physicalism posits the entailment of non-physical truths by physical truths, but it is not committed to any particular explanation of that entailment (or even to there being an explanation of it). In this way, a physicalist can maintain that supervenience tells us that the physical truths entail the non-physical truths, in the sense that if the first are true in any possible world that is a physical duplicate of our world, the second are also. That sounds much stronger than co-variation; indeed, it renders supervenience as a logically necessary relation – not as a causal one. Entailment in this sense seems therefore to fall nicely between co-variance and causal efficacy, which were involved in our weak and stronger interpretations of the proposed definition of supervenient physicalism.

Reinterpreting the physicalist thesis in this way does not, however, circumvent the difficulties. Every genuine case of entailment has *some* available explanation,

³ The term ‘downward causation’ is from the American psychologist D.T. Campbell: ‘Evolutionary Epistemology’, in P.A. Schilpp, ed., *The Philosophy of Karl Popper* (Open Court, LaSalle, Il, 1974). Pp. 413-463.

whereas this physicalist thesis has none. And every genuine logically necessary implication has some independent test of that implication, yet this reinterpreted definition offers none. Secondly, on this definition physicalism would still be compatible with idealism. Compatibility with idealism is just as bad; if the supervenience definition of physicalism results in physicalism being compatible with idealism, then it cannot be a good definition of physicalism. (Of course the point is symmetrical: a supervenience definition of idealism should not allow idealism to be consistent with physicalism.)

These are not the only problems besetting the attempt to define supervenient physicalism solely in global terms. We will not canvass here all the technical objections that have been raised, but one is worth mentioning. The definition (whether interpreted as positing a weak co-variance, a strong causal efficacy, or an entailment relation) permits possible worlds that differ only in the most minute physical detail, e.g., an extra hydrogen atom in some remote location in the universe, to differ drastically in their globally supervenient properties. For all that the definition tells us, one possible world may contain creatures with full sentience, while the other has no mentality at all, yet if physicalism is true, the presence of this extra atom in some remote location of the universe ought not to make any difference to the distribution of mental properties. There has been no shortage of proposals to get around this counter-example – for example, by weakening indiscernibility to similarity, or by defining it over spatio-temporal regions rather than worlds – but the proposed solutions are inevitably *ad hoc*.

Furthermore, global supervenience makes physicalism hostage to the ontological mysteries raised by talk of ‘possible worlds’. And it obscures the fact that physicalism is not a claim about *every* possible world, but only a claim about *our* world, to the effect that its physical nature exhausts *all* its nature. So, if the idea behind supervenience is taken to be a *contingent* thesis about our world, global supervenience has to be reformulated. For these latter reasons Frank Jackson and others have proposed the following definition: any possible world which is a minimal *physical* duplicate of our world is a duplicate *simpliciter* of our world.⁴ This reformulated definition tries to get around the difficulty of worlds differing only by a few atoms by speaking of “minimal duplicate” worlds, where “minimal” is meant to exclude precisely the counter-example of such extra isolated atoms. But that device has the effect of rendering the definition analytically true, which is surely not a congenial consequence.

So if this kind of case cannot be satisfactorily ruled out, the condition stated in the definition is not necessary, since it admits the possibility that physicalism might be true in our world, but the condition not be satisfied. Nor is that condition sufficient, since, as we saw, it fails to rule out a certain kind of idealism. A professed definition of global supervenience that turns out to be is neither sufficient nor necessary is seriously flawed.

⁴ As argued, for example, by Frank Jackson in his *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Clarendon Press, Oxford, 1998) p.12. Jackson is a reductive, or ‘eliminative’ physicalist, but that difference is not relevant at this point. By a ‘minimal physical duplicate’ Jackson means a possible world that is identical in all physical respects to the actual world, but which does not contain anything else, in particular, it does not contain any pure experience that does not interact causally with anything else.

The most serious objection to any such global definition, however, does not turn on such technicalities at all. I have already alluded to it: in the absence of specific psycho-physical correlations, and some knowledge of them, global supervenience offers no explanation of *how* some mental phenomenon supervenes upon specific physical facts. Its claimed virtue is also its weakness. This criticism applies to all the proposed characterisations of global supervenience mentioned above. As Jaegwon Kim, who is inclined to support physicalism, has trenchantly objected:⁵

such a supervenience claim should strike us as a mere article of faith seriously lacking in motivation both evidentially and explanatorily; it would assert as a fact something that is apparently unexplainable and whose evidential status, moreover, is unclear and problematic.

Unless physicalists offer *some* account of how specific non-physical processes come to supervene upon underlying physical states, their central concept does no more than name a mystery, and their commitment to physicalism remains a pure dogma.

So let us abandon global supervenience and try the other route. In its place Kim and others advocate a micro-reductive definition of supervenience, formulated as follows:⁶

mental properties supervene on physical properties in the sense that if something instantiates any mental property *M* at time *t*, there is a physical base property *P* such that the thing has *P* at *t*, and anything with *P* at some time *necessarily* has *M* at that time.⁷

This approach has the virtue of appearing to address the issue of how particular ‘local’ phenomena are supposed to supervene upon the properties of their constituents. But for that very reason, it generates difficulties of its own. For it is far from obvious that for everything that instantiates psychological or social properties, there is a set of physical properties that it has at the same time from which those psychological or social properties follow necessarily. Consider the Court of Appeal of the Australian Capital Territory. This might thought of as a social or legal entity, of which many social and legal predicates are true. But what are the physical characterisations by which this same entity could be identified, and in virtue of which the Court those social and legal characterisations necessarily follow? The judges who at any one time comprise this court are drawn from a large pool. Indeed, there might be no particular physical entity (in any natural sense of the term) that is identical to the Court of Appeal, and yet physicalists are committed to saying that the truths or facts concerning the court supervene on physical truths or facts.⁸ (Global supervenient physicalism, on the other hand, imposes no such requirement, which goes to show that it does not imply ‘strong’ supervenience.)

⁵ ‘Supervenience as a Philosophical Concept’, *Metaphilosophy*, p. 27.

⁶ See, for example, Kim: *Mind in a Physical World*, p. 9.

⁷ An alternative proposal to meet the objection that ‘global’ supervenience fails to account for local dependency relations, is to retain it, but *combine* it with ‘weak’ supervenience, that is, with a definition like that quoted except that the word “necessarily” is omitted. See Oron Shagrir: ‘More on Global Supervenience’, *Philosophy and Phenomenological Research*, LIX, 1999, pp. 691-701.

⁸ For the classic presentation of this point, see John Haugeland: ‘Weak Supervenience’, *American Philosophical Quarterly*, 1983.

Furthermore, we should note that defining the supervenience relation along these lines takes it to hold between properties – singular properties. One property is supervenient on another property; a property cannot, according to such definitions, be supervenient upon a relation or configuration or organization. It is also noteworthy that such definitions presuppose a metaphysics that takes *things* and their *properties* as basic. These points will become significant later.

The idea of supervenience, then, is no more than a name for an intellectual puzzle. It posits a dependence of the non-physical upon the physical, but the prospect that it might make that dependence intelligible has proven illusory. It is becoming manifest that supervenience is not a well-defined concept; it does not itself offer an explanatory theory; nor is it a type of dependence relation.

Why, then, did it seem so attractive? I suggest, because it does contain a germ of truth. It does seem that social phenomena in some sense arise from psychological phenomena, that the latter in turn arise from biological phenomena, and so on down a hierarchy of levels until at what seems like a bottom, there is a physical base. And phenomena at each level seem *necessary* for the emergence of phenomena at the next level up. The physicalists' mistake is to think that they are also *sufficient*, that despite the different kinds of description and explanation appropriate to each level, there is nevertheless a contingent *identity* between the entities so described at common spatio-temporal locations. If, on the contrary, a physical base is necessary but not sufficient for the emergence of higher-level phenomena, then the contingent identity thesis is false, and a physicalist metaphysics has to be rejected.

Now, physicalists typically assume that the only alternative to their own position is some variant on a Cartesian ontology, which posited two distinct kinds of substance: mind and matter. (Their setting the dialectic that way just shows that they are really one-legged Cartesians themselves!) But no Cartesian could accept the hierarchical model I have just sketched either, since they could not accept that anything physical is a necessary condition for any mental phenomena. That, however, demonstrates that an ontology which admits emergence is a genuine third alternative. The distinctness of these three alternative metaphysical positions can be clearly demonstrated in the following table, which marks for each position whether a physical base is necessary and/or sufficient for 'higher-level' properties and powers:⁹

	Physical base necessary	Physical base sufficient
Cartesian	X	X
Physicalist	√	√
Emergent	√	X

⁹ Jaegwon Kim has argued in many places that so-called emergent phenomena cannot do any serious causal work since, if they were causally efficacious in bringing about higher level effects they would also have to bring about the physical base of those effects. But then, the physical base of this higher-level cause has its own physical base, which is sufficient for the presence of that higher-level cause. It follows by causal transitivity that the physical base is doing all the causal work. This, however, begs the question. That a physical base of some higher-level phenomenon is sufficient to bring it about is precisely the distinctive physicalist thesis.

So the question becomes: what is the most plausible ontology to admit emergence? It will acknowledge that all phenomena of higher level have as constituents physical phenomena, but will renounce micro-reduction. How will that go? Well, the definition of micro-reductive supervenience discussed earlier was framed in terms of something's instantiating mental properties which are supposed to supervene upon that thing's physical base properties. What are these 'things'? One way of interpreting the definition is to take these supposedly basic things to be fundamental particles. But as Mark Bickhard has pointed out, they are abandoned by recent developments in physics. Our best contemporary physics tells us that *there are no fundamental particles*, only processes. Quantum field theory shifts the basic ontology of the universe from micro-particles to quantum fields. What have seemed to be particles are now conceptualised as particle-like processes and interactions resulting from the quantisation of field processes and interactions, and those are no more particles than are the integer number of oscillatory waves in a guitar string. Each of these things is a quantised field process.¹⁰ Indeed, Bickhard argues, "it is processes all the way up and all the way down". And more complex processes manifest genuinely novel properties and powers that are not reducible to those of their constituent processes. I won't rehearse the details of his argument today; I take it that most of you here will be familiar with it.

Now, shifting from a particle/property ontology to a process one does not necessarily refute physicalism. For a physicalist can interpret the word "thing" in the definition in the weakest possible sense, as meaning whatever is referred to by the nouns in physical theory. Physicalists only need maintain that everything that happens in the world is ultimately determined by the properties and powers of some kind of *basic particulars*. But what kind? Some physicalists believe that it does not matter, that the notion of a physical particular might be defined as an object, a concrete event, or whatever.¹¹ That does seem to me what physicalists should say; that is, their basic physical particulars could well be micro-processes. So long as the properties and powers of larger-scale processes are wholly derived from, and dependent upon, micro-processes, physicalism can accommodate all apparently emergent phenomena by claiming that they supervene upon basic processes. This preserves the ontological dependence crucial to any physicalist position.

So shifting from a thing and property ontology to a process-based ontology does not necessarily dispose of physicalism. Nor does examining the concept of supervenience suffice to get at the metaphysically interesting issue. The crucial question, I suggest, is whether downward causation is possible. If it is, then physicalism is false and physicalists' invocation of supervenience, however defined, is beside the point.

Now, it seems that instances of downward causation are all around us. A simple example is blushing. Blushing is a physiological phenomenon standardly occasioned by someone's feeling shame, guilt or embarrassment. Those feelings, in turn, are standardly associated with the interpretation of semantic phenomena. The chains of causation (or of entailment) here do not move 'upwards' from physical

¹⁰ Mark Bickhard, 'Autonomy, Function and Representation', *Communication and Cognition - Artificial Intelligence*, 17 (3-4), 2000, 111-131.

¹¹ So F. Jackson: *From Metaphysics to Ethics*, p. 6, fn. 5.

events, though chemical to biological and eventually psychological and linguistic phenomena. On the contrary, with blushing the causal movement is precisely the reverse. Emotions generated by the interpretation of semantic content cause blood to rush into the capillaries in the cheek and neck, which of course involves a movement of blood-cells, of the chemicals of which those cells are composed, and therefore of the physical quanta which comprise those chemicals.

For a physicalist, the basic problem, despite the manifest phenomena of downward causation, is to imagine how such a downward causal chain could even be possible. Surely, they would say, a sentence or a thought cannot bring about physical movements. Religious devotees might believe that faith can move mountains, but no-one who takes modern science seriously could possibly do so. Accordingly, belief in downward causation is simply mumbo-jumbo! It must be that the emotions mentioned supervene upon, and are manifestations of, certain underlying physical states, and it is the latter which bring about the physical consequences described. Any other account, they would say, is simply incredible.

This is where a process-based ontology comes into its own. My imaginary physicalist's response is understandable if one believes that the physical consists of micro-particles defined in terms of their properties and powers. How on earth could those properties and powers be changed by psychological phenomena? But if rejects such a thing-based ontology, and instead adopts a process-based one, downward causation suddenly ceases to be a metaphysical impossibility. Just as the flow of the current in a river affects the shape and movement of the eddies within it, so any macro-process can affect the micro-processes of which, in one sense, it is constituted. There is nothing mysterious about this. And there is no reason why those macro-processes should not have properties and powers of a kind which 'over and above' those that can be attributed to those micro-processes that are its spatio-temporal parts.

Let me express this 'over and above' more precisely. One of the distinctive theses of physicalism is that all macro-entities derive their distinctive properties entirely from the properties of the micro-physical entities of which they are comprised. In technical philosophical jargon, physicalism is a 'mereological' doctrine: the properties and relations of wholes are fixed by the properties and relations that characterise their parts. But there is one sense in which a physicalist can readily allow that macro-entities have properties and powers that are 'over and above' those of their constituents, namely those cases where the properties and powers of some whole can be derived from an *aggregation* of those of its parts. For example, having a mass of ten kilograms is a causally efficacious property of a table but of none of its constituents, but the mass of the table can be computed from the masses of its parts. In such cases, the relations between the properties of the whole and those of its parts are linear.

The problem for physicalists is that most properties in the world are not aggregative. *Non-linear* functions are what is crucial to causal emergence. By definition, every instance of non-linearity is an instance of whose causal properties cannot be derived by summation or aggregation of lower-level consequences.¹² In

¹² Any force of the form $(1/r)^n$, where $n > 0$, is non-linear, increasing strongly as n increases. Much traditional physics gets around this by treating sufficiently small variations as locally approximated by linear departures. Thereby, it suppresses all of the global non-linear character of these interactions. (I am indebted to Cliff Hooker for pointing this out to me.)

that sense, every instance of non-linearity is an instance of emergence. William Wimsatt has shown that there are at least four different ways in which some property of a complex stable whole might be an ‘aggregate’ of the properties of its parts.¹³ Correlatively, there is a range of ways in which a property can *fail* to be aggregative. It was common during the dominance of logical positivism to treat reducibility as the universal solvent for conceptual problems in the sciences. Wimsatt’s analysis of the different ways in which a property might result from aggregation has led him to propose that, instead of reducibility, we should take ‘not resulting from aggregation’ as the very *criterion* of a property’s being emergent.

The two crucial differences between emergence and supervenience are *a)* the properties and powers of an emergent entity or process are non-linear, and *b)* emergent entities and processes are capable of exerting downward causation upon their constituents.

One final point. To be *composed of* is not to be *a part of*, as the treatment of part-whole relations in classical mereology supposes; a table that is composed of some kind of stuff (say, wood) does not have wood as *a* part. The logic of processes is like that of stuffs, which are referred to by mass-nouns, and unlike that of things, which are referred to by count-nouns. So, although my body is composed of flesh, blood, bone, etc., those stuffs are not *parts* of my body, in the sense formalized by classical mereology. While blood, for example, can be analysed as consisting of red and white corpuscles, and those corpuscles are cells, it does not follow that my body is *nothing other* than a collection of cells.. The relation of being composed of does not amount to identity. Likewise, it is a mistake to *identify* a body of water with a collection of H₂O molecules, because the properties of water crucially involve the *interactions* which occur between those molecules in very large ensembles of them. Despite the frequency with which this example is cited as the very paradigm of contingent identity, it is far from clear that a single molecule of H₂O is itself water.¹⁴ Likewise, the properties of blood are not determined simply by the properties of the red and white corpuscles that comprise it; the interactive configurations of those cells play a crucial role in determining the properties and powers of the blood that they compose. That is why the transitivity posited by classical mereology does not go through and why the properties generated by configurations do not *reduce* to the relational properties of constituent ‘basic particulars’.

There are further twists and turns in this debate, which there is not time to canvass here. Physicalists do not give up easily. But it is becoming clear that both contemporary physics and logic require switching from an ontology of micro-reduction to an ontology that gives priority to the relational characteristics of fields in process is not only necessary but also yields a distinctively different understanding of ourselves and the world we live in.

¹³ William C. Wimsatt: ‘Forms of Aggregativity’ in *Human Nature and Human Knowledge*, ed. Alan Donagan, Anthony Perovich & Michael Wedin (Reidel, Dordrecht, 1986) pp. 259-291.

¹⁴ As Paul Humphreys has pointed out, “individual molecules of water do not have the all the properties of water – such as liquidity – despite the popular dogma of their identity. Some properties of water are a result of features that occur only in very large ensembles of (interacting) molecules. You can’t stipulate *that* away.” See his ‘Aspects of Emergence’, *Philosophical Topics*, **24**, 1996, p. 68, fn 25.