Emergence and Efficacy

Imagine the day when physics is complete. A theory is in place which unifies all the forces of nature in one self-consistent and empirically verified set of absolutely basic principles. There are some who see this day as perhaps not too distant (e.g. Hawking 1988, Weinberg 1992, Horgan 1996). Of course, the mere possession of this *theory* of everything will not give us the ability to provide a complete *explanation* of everything: every event, process, occurrence and structure. Most things will be too remote from the basic theory to admit of explanation in its terms; even relatively small and simple systems will be far too complex to be intelligibly described in the final theory.

But seeing as our imagined theory is fully developed and mathematically complete it will enable us to set up detailed computer simulations of physical systems. The range of practicable simulations will in fact be subject to the same constraints facing the explanatory use of the theory; the modelling of even very simple systems will require impossibly large amounts of computational resources. Nonetheless, possession of a computational implementation of our final theory would be immensely useful. Real versions of something very like my imaginary scenario now exist and are already fruitful. For example, there are computer models of quantum chromodynamics that can compute the theoretically predicted masses of various quark constituted sub-atomic particles (see Weingarten 1996). The looming problem of computational intractability is all too evident, for realizing these calculations required the development of special mathematical techniques, the assembling of a dedicated, parallel supercomputer specially designed for the necessary sorts of calculations (a computer capable of 11 billion arithmetical operations per second) and roughly a *year* of continuous computing. Weingarten reports that a special 2-year calculation revealed the existence of a previously unrecognized particle, whose

existence could be verified by examining past records from particle accelerator experiments. Modelling the *interactions* of particles would be a much more challenging task, suggesting to the imagination computational projects analogous to the construction of medieval cathedrals, involving thousands of workers for many decades.[1]

I want to introduce a thought experiment that flatly ignores the inevitably insuperable problems of computational reality. Imagine a computer model of the final physical theory which has no computational limits: imagine that detailed specifications of the basic physical configuration of any system, at any time, in terms appropriate for the final theory, are available, so that if the configuration of any physical system is specified as input then the output configuration of the system, for any later time, can be calculated (and appropriately displayed) in a reasonable time. Now, there is nothing incoherent in the idea of an absolutely perfect simulation. In fact, we have some of them in physics already. The Kerr equations for rotating black holes are (if the GTR is true, absolutely perfect models of these strange objects. The famous astrophysicist and mathematician Subrahmanyan Chandrasekhar confessed that 'in my entire scientific life ... the most shattering experience has been the realization that an exact solution of Einstein's equations of general relativity, discovered by the New Zealand

[1] Would it ever make sense to start such a project? Not if computer technology progresses sufficiently quickly. Suppose the original length of the computation is *n* years and technology advances so quickly that after *d* years have passed the computation would take less than *n* - *d* years. Then it would never make sense to start the computation! Of course there are non-computer technical constraints on the time required for such computations and presumably the pace of progress in computer technology must eventually slow down rather than continue its heretofore exponential acceleration. For some *n*, the computations make sense, as evidenced by the real world examples given above, but the problem of this note is equally well illustrated (see Weingarten 1996). Computers of the 1980s would have taken about 100 years to perform the reported computations. It was not worth starting.

mathematician Roy Kerr, provides the *absolutely exact representation* of untold numbers of massive black holes that populate the universe' (as quoted in Begelman and Rees 199?, p. 188).

However, we are not so lucky with the rest of the world, and so even within our dream certain approximations in the input configurations will have to be allowed. We cannot input field values for every point of space-time and it is conceivable that some configurations require an infinite amount of information for their specification if, to give one example, certain parameters take on irrational values which never cancel out during calculation. Let us therefore imagine that we can input specifications of whatever precision we like, to allow for modelling the system for whatever time we like, to whatever level of accuracy we desire. Even though it is physically impossible, I think the idea of such a computer program is perfectly well defined and, I will try to show, can starkly illuminate a problem about the efficacy of emergent properties (which is a much more serious problem than might appear from this bloodless preliminary description).[2]

Emergentism is not a single doctrine but rather a system of related views characterizable through a variety of distinctions.[3] It is common to distinguish a supposedly 'benign' or acceptable emergentism from an unacceptably non-naturalistic 'radical' emergentism. Within

---

[2] Could we define physicalism in terms of this imaginary computer implementation of final physics? We might try something like this: physicalism is the doctrine that everything that occurs/exists in the actual world would have its exact counterpart in a final physics computer simulation of the world, or that the simulation would be, in some appropriate sense, indistinguishable from the actual world. Such a formulation has the advantage of automatically including what Hellman and Thompson (1975) call the principle of physical exhaustion. But it obviously requires a clearer specification.

[3] The classical doctrine of radical mental emergence is associated with such authors as C. D. Broad 1925 and C. Lloyd Morgan 1923. A modern review of this material would be welcome (for a start see Kim 1993, ch. 8, and for more on emergence see the collection edited by Beckermann, Flohr and Kim 1992).

benign emergentism further divisions are possible. For example, John Searle (and also Paul Churchland 1985, note 6) recognizes an emergence of 'systems features' (or, in Churchland's phrase, 'network properties'), which are properties which a system has but no proper part of the system possesses (Searle's examples are particular weights of systems and their shapes), as distinct from (but encompassing) what he calls 'causally emergent system features', which are not merely a matter of physical composition but rather 'have to be explained in terms of the causal interactions among the elements' (examples: liquidity, transparency; see Searle 1992, pp. 111 ff.)[4]. Searle then explicates the notion of *radical* emergence as further demanding that the emergent property have its own causal powers which 'cannot be explained by the causal interactions of' the parts of the system.

Now, I think emergence is and should be a metaphysical (or perhaps it would be better to say *ontological*) doctrine and so I regard the widespread reliance on the notion of explanation in the explication of emergentism as misplaced, a more or less subtle *confusion* of metaphysics with epistemology which is distressingly rampant in discussions of these problems. The fact that we cannot *explain* the causal powers of a property in terms of the basic constituents of a system might reflect no more than cognitive weakness, or perhaps an overwhelming complexity of the system at issue, rather than a metaphysical fact of radical emergence. But our computer model thought experiment makes everything clear. A property, F, of a system, S, is radically causally emergent just in case S's behaviour diverges from the behaviour of our computer model of S where it is the possession of F that causes the divergence. Otherwise, assuming, naturally, that F

[4] Searle's example of weight really belongs in the category of causally emergent features, since the binding energy between the constituents – a matter of their causal interaction – affects the mass of the total system.

is not itself a basic physical property – and so explicitly mentioned in the final theory – F is a benignly emergent property.

It is worth mentioning here that the indeterminism which we will likely find in the final physics can easily be accommodated by considering ensembles of identically prepared systems, and measuring the tell-tale divergences of emergentism against the behaviour of the ensemble. One can envisage the emergent properties revealing themselves in at least two distinct ways. First, the divergent behaviour might be different from the behaviour of *all* the members of the ensemble. Second, we could have a merely probabilistic case for emergence if the behaviour of the real system was consistently like that of only a tiny fraction of our ensemble. That is, persistently improbable behaviour would also be a mark of radically emergent properties at work. Thinking about indeterminism also suggests a new kind of emergentism: the emergence of indeterminism. We can imagine emergent properties that introduce indeterminate behaviour into a system which the final physics claims to be deterministic, or, as in the second case above, emergent properties could skew the probabilities which the final physics assigns to a system's possible behaviour (sometimes free will is thought to introduce such indeterminacies). It is, however, rather more probable that *determinism* is an emergent property of systems constituted from fundamentally indeterministic elements. It is very likely that macroscopic systems will behave in essentially deterministic ways, that is, that the probabilities of their various possible macroscopic state transitions will approach 1 and 0[5]. Our imaginary computer model will naturally take all this into account.

---

[5] Many deep issues lurk in this region having to do with the emergence of the apparently classical macroworld from the non-classical, quantum microworld; see Hughes 1989.

Note also how this model nicely finesses another issue: the notorious oddity of quantum part-whole relations. In quantum mechanics, systems have causal properties which do not always smoothly reduce to the properties of their parts. For example, systems which are in superpositions of possible states are behaviourally distinct from systems which are in mixtures of these states, and individual systems can become 'entangled' and thus form a new unified system with distinctive properties (to see how strange this can get consult Albert 1992 or, for a particular example, Seager 1996). But the theory itself tells us how these superpositions and entanglements arise through particle creation, annihilation and interaction as well as how they subsequently behave, and our computer program would naturally – and by assumption successfully – include such features in its simulations. This means that we shall not *have* to regard quantum mechanics as a theory propounding the *radical* emergence of quantum mechanical properties (e.g. the properties of superpositions). On the other hand, there may be some merit in the notion that quantum mechanics is precisely the scientific appropriation and regimentation of the idea of radical emergence. This is an intriguing idea which I won't explore here but I will remark that quantum mechanics is, due to its linearity, at best a very non-general theory of emergence.

To see better how the thought experiment works, let's examine a classic case from a naive point of view that might briefly suggest we have a case of radical emergence. The mechanical model of gases allows us to deduce the perfect gas law relating pressure (P), temperature (T) and volume (V): PV = RT. If we plugged a specification of the initial state of a gas – recall that we are assuming that we can, somehow, obtain appropriate descriptions of these states – into a computer simulation based on the idealizations of the simplest mechanical model (which is, for all its conceptual simplicity, no less computationally intractable than our envisaged

computational implementation of final physics) the computer would generate output states that provide a pretty fair representation of the gas and which would, naturally, obey the perfect gas law. But under a variety of conditions the actual gas's behaviour would exhibit greater or lesser divergences from the behaviour of the model gas. So we might – rather too quickly – deduce that there are radically emergent properties of gases at work. *Or* we could deduce that the theoretical structure of our model is inaccurate. Of course, the latter is correct. A *better* model of the parts of the gas (almost) eliminates the divergences. For example, a better model admits that there are some interactions between the microscopic parts of a gas, and that the parts must take up a little of the volume of the gas. An intuitive appreciation of the micro-structures involved in the model led van der Waals to incorporate these features in a modified gas law: $(P + a/V^2)(V - b) = RT$. So, as is common in the development of scientific theories, the appearance of a possible radically emergent property disappears with a refined analysis of the components of the system and their interactions. Pressure and temperature remain as emergent properties of the gas: they are not properties deployed in the computer model or the basic theory, but they are obviously benignly emergent. They can be successfully applied to the gas as represented by the computer model, with no aspect of the gas's behaviour 'left over'. That is, the computer generated simulated gas 'acts' just like it had a temperature and exerted pressure without us having to write these into our program.

Any case of apparently radical emergence can (and usually should) be taken as evidence of weakness in the underlying theory. Interesting methodological issues are raised by the problem of determining under what conditions we should agree that radical emergence has been empirically discovered. An extremely bizarre possibility can be imagined in which radical

emergence is true, but invisible because we are clever enough to develop a theory which, while it is actually false, manages to generate all the empirical effects of the emergent properties from the theory's hypothesized micro properties, entities and processes. According to such a fantasy, it might be that the extremely complex devices employed by high energy physicists are revealing radically emergent features (of complexes such as particle accelerators, detectors, computers, etc.), which serve only to spur theorists' efforts at modifying their fundamental theories to incorporate them (theorists possess, after all, a notorious ability to account for any empirical result)[6]. But I am – for now – assuming that the final theory is *true*.

That old whipping horse, vitalism, can be regarded as involving radical emergence. Another distinction looms here which should be briefly mentioned. It seems possible to develop a theory of 'substance emergence' as well as the more typical property emergence[7]. The former posits the appearance, if and whenever certain definite material configurations arise, of a new

---

[6] One story goes like this: an experimentalist rushes to show the latest graph to a theorist, who proclaims: 'this is easy to explain' and launches into a theoretical assessment. The experimentalist breaks in with 'I'm sorry, this graph is upside down'. The theoritician replies: 'Ah, this is even *easier* to explain'.

[7] What is more, it is possible to add to *property* and *substance* emergence a notion of *law* emergence (in fact this is the sort of emergence that writers like Broad and Morgan may have had in mind). Law emergence would posit that certain complex assemblies of matter exhibit novel behaviour in virtue of obeying genuinely new laws of nature. In terms of our computational thought experiment, the behaviour of such complex assemblies would diverge from the behaviour of identical simulated assemblies because of the new, high-level laws of nature that govern emergent phenomena at the appropriate level of complexity. It may be, however, that law emergence is not a genuinely distinct sort of emergence, if we take the view that it is the properties of things that explains why they obey the laws that they do. If there is a physical description of the conditions under which the new laws emerge and 'take hold' (and presumably there is if they are *laws*), then perhaps we could regard systems which meet these descriptions as possessing an emergent property which causes the divergent behaviour. It is not clear to me that this is an alternative to law emergence or simply a different formulation of it.

substance which then takes an active part in the causal commerce of the world. I suppose that the emergent substance is annihilated upon the destruction of its material base (but who knows). Maybe vitalism ought best to be regarded as a kind of substance emergence (talk of the *élan vital* might suggest this interpretation), but this nicety need not worry us here. Either way, vitalism posits that *living* matter behaves differently than *dead* matter. But *that* is uncontroversial. The proper expression of the radical claim of vitalism is that our final computer model of the physical configuration of, say, a living cell or bacterium, would fail to provide an accurate representation of the cell's behaviour. Unlike simulated gases, computer simulated cells will just not act like their real counterparts. So says vitalism.

I don't think there is any evidence that *life* is a radical emergent property (let alone an emergent substance). But given the absence of our final computer program what justifies amplifying the evidence into an outright denial of vitalism? It is what I call the 'physical resolution'[8] of living systems into basic physical parts which take an entirely normal place within the developing picture of the physical world. Though not conclusive, it is highly suggestive that no strange, inexplicable processes have been found in living cells and that a great many of the basic mechanisms of life have been explained in molecular terms. This is far from proof of course – we have examined only a tiny fraction of the mechanisms of life and those the most elementary, and, after all, the emergentist no less than the physicalist positively expects that all systems will resolve into ordinary material sub-systems, ultimately into the basic physical building blocks. It is *only* their complex combination in living systems that reveals the emergent properties. So perhaps it is a kind of faith that underpins the anti-vitalist, though it is a faith –

[8] For more on this idea and its relation to epistemic/explanatory matters, see Seager 1991.

unlike that of the unredeemed vitalist – which *all* of the still fragmentary evidence we do possess fully supports. Expressed in terms of our computer thought experiment, the faith is that a complete elementary final physics simulation of an entire organism (plus the requisite features of its environment) would successfully replicate the behaviour of the real thing. The faith is supported by the piecemeal successes of resolving a range of very simple biochemical processes into the physical parts which reveals how the non-living properties of these processes produces them (here it is assumed that no radical emergence occurs in the gap between the elementary physical and the biochemical levels – presumably a reasonable assumption, for which we have still more evidence than for our anti-vitalism, and, as always, an excellent working hypothesis). The faith is coupled with a parsimonious desire to understand the most with the least number of hypotheses and so far absolutely no vitalistic assumptions of radical emergence have been warranted.

But reflecting on this example and looking ahead to the case of the mind suggests yet another possible sort of emergence: a form of non-causal emergentism I'll label *epiphenomenal* emergence. The classic doctrine of mental epiphenomenalism *is* a doctrine of emergence, at least so long as it is not a form of panpsychism (I take it that epiphenomenal panpsychism is *almost* equivalent to the Leibnizian doctrine of pre-established harmony or a universal psycho-physical parallelism – it adds only the idea that the physical side of the equation causes the psychical side). Classical epiphenomenalism maintains that certain material configurations bring about instantiations of mental properties which were not exemplified prior to the appearance of that physical configuration and which properties are not to be identified with the physical. Unfortunately (or not depending upon one's viewpoint) these mental properties have no

distinctive, observable causal influences upon behaviour. Non-causal emergentism is invisible.

Perhaps classical epiphenomenalism is not of much interest, but mixing epiphenomenalism with emergentism forces us to ask a very disturbing question: is benign emergentism a kind of epiphenomenalism? Recall again the benign emergence of temperature and pressure. As we know, the pressure of a gas does not *really* cause anything; it has no efficacy of its own. Pressure is the mathematical average of the impact forces of the gas's constituents as they strike the gas's containing boundary and mathematical averages do not cause anything. If one is inclined to think otherwise, consider this example: a demographer might say that wages will go up in the near future since the average family size fell 20 odd years ago (and so now relatively fewer new workers are available). There is not the slightest reason to think that 'average family size' can, let alone, *does* cause things although I think we easily understand the explanation to which such statistical shorthand points[9]. By its very nature, pressure is not one whit less a statistical 'fiction' than is average family size. The ascription of causal efficacy to pressure is only a *façon de parler*, a useful shorthand for the genuine efficacy of the myriad of micro-events that constitute 'pressure-phenomena'. It is entirely correct to use the overworked phrase, and say that pressure is *nothing but* the concerted actions of the countless particles that make up a gas. In terms of our computational thought experiment, it is easy to tell whether some property has any *causal* efficacy of its own: do we need to code that property into the simulation. If not, then the property has no efficacy of its own.

---

[9] One possible snare: the *conscious apprehension* of 'average family size' evidently can cause things but examples like these are – if they are examples of efficacy of any kind – examples of the efficacy of representational states of mind, not of the efficacy of what is represented. Thoughts about unicorns have their effects, but admitting this does not concede any causal powers to unicorns.

Benign emergence arises wherever we can find a descriptive and/or explanatory scheme whose application to some class of systems provides a useful kind of shorthand notation for describing the behaviour of those systems. The behaviour is being orchestrated from below, as it were, in the details of these systems' physical constitution. The joint action of these constituents happens to form patterns which we can codify in terms of the benignly emergent properties deployed in our 'high level' description of the systems (an interesting discussion of this process can be found in Dennett 1991a). Consider, for instance, the coriolis force, which gunnery officers must take into account when computing the trajectory of long-range cannon shells. (A host of other activities require cognizance of the coriolis force as well.) This is a benignly emergent property of the earth, or any other rotating system. But in the context of assessing causal efficacy and the proper physical basis of the world, it is highly misleading to say that the coriolis force causes diversions in a shell's trajectory. At least, if we really thought there was such a force – hence with its own causal efficacy, the world would end up being a much stranger place than we had imagined. Just think of it: rotate a system and a brand new force magically appears out of nowhere, stop the rotation and the force instantly disappears. That is radical emergence with a vengeance. Of course, there is no need to posit such a force. The coriolis phenomena are related to the underlying physical processes in a reasonably simple way – in fact simple enough for us to 'directly' comprehend, but, no matter the complexity, our imaginary computer model of any rotating system would naturally reveal the appearance of a coriolis force.

Another extremely important example of presumably benignly emergent phenomena is found in evolutionary biology. The theory of evolution, with its profound use of the concept of natural selection along with the now unravelled genetic basis of inheritance, is an undeniably

deep insight into the nature of the world. And unlike the case of pressure in thermodynamics, there is no simple mathematical relation which connects, say, adaptedness (to environment X) to underlying physical structure. But given our knowledge of biochemistry and the molecular basis of genetics it is very likely that adaptedness (along with all other evolutionary properties) is indeed benignly emergent. If so, we know that adaptedness itself adds nothing to the causal forces of the world. Evolutionary theory is a way of consciously apprehending certain abstract structures in the world which brilliantly highlights certain more or less enduring patterns to be found in the 'dance of the atoms'. *We* need the idea of evolution to understand the world, but the *world* has no need of it; the world – not even the biological parts of it – is not being driven by evolutionary properties. Predator-prey relations will be revealed in the final physics computer model of an environment in precisely the form they take in the world itself; the Hardy-Weinberg law will emerge from the quark/lepton/boson sea of our imaginary simulation with not a jot of evolutionary theory needed in the underlying program (unless of course evolutionary properties are more than just benignly emergent).

The basic problem is that, metaphysically speaking, benignly emergent properties are unnecessary hypotheses, even if they are parts of useful high level explanatory schemes. The world can be described, predicted and understood in terms of any number of interesting and useful explanatory schemes, but a reasonable principle of economy enjoins us not to multiply hypotheses beyond necessity and manifestly, by definition one might say, there is never any need to appeal to benignly emergent properties when we consider the metaphysical underpinnings of the world. They are nothing but artifacts of convenient ways of thinking and speaking. I am not saying that the benignly emergent properties are practically eliminable or that there is any

prospect of doing without these ways of organizing the world around us, or even that it would be abstractly desirable to do away with them. In fact, they are necessary for our *understanding* of the world and indispensable aids in the construction of the picture of the physical world that finally reveals them to be nothing but benignly emergent properties. But I am saying that they are *metaphysically* unnecessary and therefore cannot be counted among the 'driving elements' or forces of the world.

It seems to me interesting that such an outlook would be entirely acceptable, and indeed possessed of a rather austere beauty, were it not for one recalcitrant phenomenon: consciousness. We understand perfectly how things like the coriolis and centrifugal forces, or pressure and temperature, or adaptedness and natural selection 'cause' things in the world through the action of a host of underlying genuine causes whose proper description does not require any appeal to these properties. Again I stress that this is not in any way to dismiss the importance of the high-level schemes. It is a remarkable and fascinating feature of the world that it can support such a wondrous hierarchy of high-level descriptive schemes, sometimes (especially in such cases as thermodynamics, evolutionary theory and, perhaps, intentional psychology) in ways that utterly transcend the details of the systems so described[10]. But there is no discomfort in assigning

---

[10] It is one of the glories of physics that it can often *show* how the elementary transforms itself into complex systems which obey the laws of the high-level theories, as in the case of thermodynamics, but in many other places as well. For example, Newton famously showed how the gravitational force of a myriad of low-mass particles could act as a single high-mass object and it is a set-piece in physics texts how the angular momentum of the parts of a body is combined into the angular momentum of the whole (from which follows the law of the conservation of angular momentum). Nowadays, chaos theory is successfully entering this business. The way that thermodynamical properties emerge may have more relevance to psychology than mere metaphor. There are deep analogies between thermodynamics and the dynamics of neural networks (see for example Churchland and Sejnowski 1992), and if the latter underlie psychology then psychological properties may be surprisingly closely analogous to

the benignly emergent properties a metaphysically second-class status. This is true also for mental properties, so long as we suppose they are not conscious states (and here it helps to consider them as states of beings other than oneself). For these mental states, the analogy with the coriolis force can look quite good (a well developed treatment can be found in Dennett 1987; I do not find the extension of the theory to consciousness in Dennett 1991b so plausible).

But if consciousness turns out to be benignly emergent in the same way as the coriolis force then we have an extremely serious problem about the efficacy of the mental (see Kim 1993, e.g. ch. 8, for more on explanatory exclusion[11] and efficacy as well as emergentism). If the mental turns out to be benignly emergent in the way that pressure and temperature are then the problem may appear to be less severe. This appearance has given false comfort. For in fact, with regard to their efficacy there seems to be little to choose between the coriolis force and the force of pressure (or any other benignly emergent property): they emerge from the concerted actions of the relevant physical systems' sub-structures in fundamentally the same way. Another example to nail the point down: do centres of mass – say, their positions or the motions of these positions – really cause anything? Surely not; they are, we say, convenient fictions. They are very useful in calculation and as general aids to *understanding*, but they are not themselves efficacious. Now, notice that, formally, the centre of mass and the force of pressure are perfectly analogous. The

---

thermodynamical properties.

[11] Which I think is misleadingly labelled by Kim. It should be termed something like 'efficacy exclusion'. Explanation is a matter of how we *understand* phenomena and for that the high-level schemes of benignly emergent properties are indispensable and do not exclude one another. But efficacy is a matter of what is *driving* the world forward at its metaphysical roots, and there if the efficacy of the elementary parts, in all their conjunctions, suffices to *produce* every phenomenon in the world, then there is just no evidence for their *being* any other efficacious features in the world.

centre of mass is the average position of all the masses. The pressure is the average force of impact of all the particles. Such properties have no efficacy of their own; they are *nothing but* certain, extremely useful, ways of describing the observable effects of the joint action of the constituents of the system to which they apply.

It has often been remarked that if we try to imagine that the efficacy of, say, the painfulness of a consciously experienced pain is to be eliminated in the metaphysically correct view of the world, great discomfort ensues. Nonetheless, it looks very much like the benign emergence of the mental is tantamount to epiphenomenalism. Note that if this argument is correct then both reductive and non-reductive materialism fall victim to it, since they both subscribe to the completeness of physics as expressed in the imaginary computer thought experiment. There has been a lot of work lately on the problem of mental efficacy within the context of non-reductive materialism, with results that remain controversial. The problem of benign emergence I am urging here threatens to make the problem much worse.

Because it is especially clear in, but not unique to, his position, let me illustrate the problem as it arises for Searle. Searle explicitly claims that consciousness is benignly emergent and that it is *not* epiphenomenal (see Searle 1992). He also notes that ‘first-person features are different from third-person features’ (1992, p. 117). I take it that first-person features are the subjective properties of states of consciousness whereas third-person features are the objective properties of the physical world. If we say that pain is not epiphenomenal we are at least saying that pains cause some events. Let’s suppose, to be definite, that a certain pain in subject S causes S to start sweating, trembling and complaining. But now consider the idea that it is not the painfulness which causes these effects but some non-mental features either of the pain or the

subject in general. Since the painfulness is distinct from the 'third-person' features of the pain, this is a possibility (as Searle in effect admits during the development of his thought experiments in 1992, ch. 3). Worse than that, since pain is benignly emergent, this possibility must be actual. Our perfect computer model of S would show S sweating and trembling and making certain sounds without any need to code the 'painfulness' into the program. So even if, in some attenuated sense, the *pain* is not epiphenomenal, the *painfulness*, at least, is epiphenomenal after all.

Contrary to what many think, the situation is not improved by trying to suppose, *pace* Searle, that we can *identify* first- and third-person features. Unless we adopt panpsychism, there is no hope of identifying consciousness with the elementary physical features of the world, so if subjective properties are benignly emergent they are to be identified as high level patterns or configurations of the basic elements and as such they are themselves benignly emergent and hence metaphysically redundant.

Since the conclusion of this argument is likely to be hard to accept, let me make the point another way. Let us try to *distinguish* between benign emergence and epiphenomenalism. The difference we would like to find is that the benignly emergent properties retain a causal efficacy that their epiphenomenal surrogates lack. But how could this efficacy be defined? Epiphenomenal emergentism asserts that there is a lawlike connection between certain (sets of) physical states and their correlated emergent states, so the sorts of counterfactuals that typically attend and reveal causal efficacy remain true under the hypothesis of epiphenomenalism. In the nearest world where the emergent states are different, the underlying physical states will be different as well (none of the relevant physical states which ground the epiphenomenal state will

obtain), and so the target caused event will also not occur[12]. So a counterfactual analysis of efficacy cannot distinguish between benign emergence and epiphenomenalism.

It is evident that epiphenomenal emergence entails a kind of benign emergence. The only thing we might consider unbenign about classical epiphenomenalism is the supposedly non-physical nature of the epiphenomenal properties it posits. This is nothing to worry about: let us simply *define* 'epiphenomenal physical emergentism' (EPE) as the doctrine that the epiphenomenal properties brought into being upon the creation of certain basic physical configurations are, whatever else they might be, physical properties. Of course there is also a more familiar form of epiphenomenalism which asserts that the epiphenomenal properties are non-physical. But I am concerned here with EPE, and in particular with mental EPE – the doctrine that mental properties are (1) epiphenomenal and (2) physical as well as mental properties. I can be prevented from making this definition if it is impossible for a physical property to be epiphenomenal and for this, I would like to see a proof. Any possible proof would seem to require the assumption that all physical properties are causally efficacious, which begs the question.

In fact, I think there are lots of epiphenomenal physical properties even within physics, but they are usually dismissed as 'mathematical artifacts': properties that arise in the mathematical expression of the theory but which don't do anything (and weren't targets of reduction) and so are considered 'unobservable' or 'unphysical', though they might be useful for calculation. But one can't take such a simple stance towards them however since sometimes

[12] This is somewhat crude. For a more precise discussion of efficacy and counterfactuals see Seager 1991, ch. 6.

these ‘artifacts’ can bite, as in, for example, the astonishing Bohm-Aharonov effect in which a particle responds – if that is the right word – to a magnetic field which has value zero (where the particle is), the explanation of which appeals to the efficacy of what was thought to be a mathematical artifact of electro-magnetic theory (the so-called vector potential). Note that our imaginary computer model would catch this effect. Of course, from the point of view I have been urging, pressure, temperature, the coriolis and centrifugal forces, etc. are all epiphenomenal physical properties.

You might want to say that a physical property is a property which can be instantiated *only* by physical objects (assuming we have some idea of what these are apart from their instantiating just any physical property – perhaps objects are physical in virtue of possessing *fundamental* physical features such as mass, charge, momentum, etc.). If we allow that it is *possible* that there are non-physical objects, perhaps in very distant possible worlds, which instantiate mental properties then mental properties are, by this criterion, not physical and mental EPE is ruled out. But then we have also rejected physicalism and adopted classical mental epiphenomenalism, for we are asserting that the actual world has honest to God non-physical properties instantiated within it, which don’t do anything in this world. On this line, if you will not allow me to define mental EPE then you are denying the truth of physicalism.

Some doubts may remain about the converse, about whether benign emergentism implies epiphenomenalism. But I think these doubts stem from the attractive but illicit assimilation of the explanatory with the causal order. I am not claiming that benignly emergent properties are explanatorily epiphenomenal – they are features of useful, utterly indispensable, descriptive schemes, by which we bring order and understanding into our picture of the world. But

metaphysically they don't do anything; the efficacy of the benignly emergent properties is, by definition, an unnecessary hypothesis.

So we have it that benign emergence is equivalent to epiphenomenalism. This places us in an unpleasant dilemma with regard to the mental: either mental properties are epiphenomenal or they are *radically* emergent properties[13]. This is a disturbing conclusion which I am sure will be resisted. One mode of resistance is once again to say something soothing like the following: look, a proper analysis will legitimate the claim that pressure causes things *insofar* as the concerted appropriate actions of the gas's constituents cause things. As I have said above, this is not really objectionable as a *façon de parler*. But it is obviously unsatisfactory with regard to the conscious mind. Consider the relevant gloss on a case of mental causation: consciousness causes things *insofar* as a bunch of non-conscious events cause things. The way the mental property drops out of the picture is only too evident: this is a species of epiphenomenalism.

The problem can be approached from a slightly different angle. The world is made of basic physical constituents all with their basic physical properties, arrayed and interacting in innumerable complex ways, but everything that happens happens because of properties and events at the basic level. The high level descriptions we use to order and understand the world are products of consciousness and they are apparent only to consciousness. The efficacy that high-level properties appear to possess is also only evident to consciousness. The high-level properties are like shadows cast in the mind by the action of the fundamental elements of physical reality upon the conscious mind. Imagine that I try to show that pressure is efficacious as

[13] I note again that panpsychism should be allowed as a third, though I doubt any more welcome, disjunct.

such by getting you to squeeze a balloon – you can *feel* the pressure actively resisting your efforts can't you? But obviously this doesn't tell us anything about pressure as such, but only about the conscious apprehension of the world *as* revealing 'pressure phenomena'. The question should be: does the mathematical average of impact forces on your hand play an essential causal role in producing your apprehension of the balloon's resistance, or can the myriad of impact forces do this by themselves? Obviously, it is the latter question that gets the affirmative answer. Your apprehension is caused – in part of course – by the myriads of impacts whose average force is codified as pressure; this average itself does not intervene in the causal process[14]. Metaphorically speaking, the world does not need to pay any attention to high-level properties, whereas, so to speak, the world is paying attention to the low level properties described in the final physical theory. They, and only they, are driving the world forward. Metaphorically speaking, the world is *running* a perfect simulation of the final physics. Consciousness is, if benignly emergent, a feature of a high-level description, apparent only to consciousness itself. The world does not need to pay, and is not paying, any attention to it; it is, therefore, epiphenomenal.

Or else it is radically emergent (hopefully – I guess – a case of radical *physical* emergence). This would mean that the physical world itself is the home of radical emergence – and I mean here, as above, radical causal emergence, not explanatory emergence, the 'disunity of science', theoretical non-reductionism, or some other weaker notion by which we might seek to salve our consciences. I think this would be a stunning reversal of three hundred years of

[14] This is almost evident from the way that talk of pressure will break down in conditions of extremely rarified gases. While we reduce the number of particles to very small numbers, pressure can be held constant in the sense that we can imagine increasing the velocity of the remaining particles of the gas, but with sufficient rarefaction we will no longer observe pressure-like phenomena.

scientific progress, and require a complete rethinking of the metaphysical basis of naturalism and the very nature of physical theory. For example, radical emergentism would seem to raise serious problems for some basic conservation laws (notably the conservation of energy) which form a part of the basic physical theory, which are evident at all levels of description and which are all 'implemented' in the most fundamental physical processes. If this is the cost of a scientific approach to consciousness, it is a very high cost indeed.

It remains possible to say that although the metaphysical picture of physical resolution suggested by the approaching final physics has no alternative, we are simply incapable of understanding how consciousness can be fitted into the picture. If this is the correct response to the dilemma then the argument of this paper is another step in the development of 'mysterianism' about consciousness[15]. Although it is seemingly evident that nature has managed to combine consistently a genuinely efficacious consciousness with a world that causally resolves itself into a system of non-conscious elementary units, how this has been accomplished is as deep a mystery as the production of consciousness itself.[16]

William Seager
University of Toronto at Scarborough

---

[15] The term 'mysterianism' is Owen Flanagan's dismissive label for the views of Colin McGinn 1989 and certain aspects of Thomas Nagel's position on subjectivity (see Nagel 1974).

[16] Another possible response, but one I will not explore here, is the wholesale denial of the kind of scientific realism which underpins the whole argument given above. A non-realist approach to science similar to that of Bas van Fraassen reduces all scientific theorizing to mere model constructing (see van Fraassen 1980); the realm of efficacy can then be in the surface phenomena where we found it in the first place. However, a problem analogous to the one urged above will emerge in the efforts to make the maximally unified model of the physical world. We may have to be content with a radically disunified science.

References

Albert, David (1992). *Quantum Mechanics and Experience*, Cambridge, MA: Harvard University Press.

Beckermann, A., H. Flohr and J. Kim (eds.) (1992). *Emergence or Reduction: Essays on the Prospects of Nonreductive Physicalism*, Berlin: W. de Gruyter.

Broad, C. D. (1925). *The Mind and Its Place in Nature*, London: Paul, Trench, Trubner.

Churchland, Patricia and Terrence Sejnowski (1992). *The Computational Brain: Models and Methods on the Frontier of Computational Neuroscience*, Cambridge, MA: MIT Press.

Dennett, Daniel (1987). *The Intentional Stance*, Cambridge, MA: MIT Press.

Dennett, Daniel (1991a). 'Real Patterns,' *The Journal of Philosophy*, 89, pp. 27-51.

Dennett, Daniel (1991b). *Consciousness Explained*, Boston: Little, Brown and Co.

Hawking, Stephen (1988). *A Brief History of Time*, New York: Bantam Books.

Hellman, Geoffrey and Frank Thompson (1975). 'Physicalist Materialism', *Nous*, 11, pp. 309-45.

Horgan, John (1996). *The End of Science: Facing the Limits of Knowledge in the Twilight of the Scientific Age*, New York: Addison-Wesley.

Hughes, R. I. G. (1989). *The Interpretation of Quantum Mechanics*, Cambridge, MA: Harvard University Press.

Kim, Jaegwon (1993). *Supervenience and Mind*, Cambridge: Cambridge University Press.

McGinn, Colin (1989). 'Can We Solve the Mind-Body Problem', *Mind*, 98, 391, pp. 349-66.

Morgan, Conwy Lloyd (1923). *Emergent Evolution*, London: Routledge and Kegan Paul).

Nagel, Thomas (1974). 'What is it Like to be a Bat?', *Philosophical Review*, 83, pp. 435-50. Reprinted in Nagel's *Mortal Questions*, Cambridge: Cambridge University Press, 1979.

Seager, William (1991). *Metaphysics of Consciousness*, London: Routledge.

Seager, William (1996). 'A Note on the Quantum Eraser', *Philosophy of Science*, 63, 1, pp. 81-90.

Searle, John (1992). *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Weinberg, Steven (1992). *Dreams of a Final Theory: The Search for the Fundamental Laws of Nature*, New York: Pantheon Books.

Weingarten, Donald (1996). 'Quarks by Computer', *Scientific American*, 274, 2, pp. 116-120.