JAEGWON KIM

MAKING SENSE OF EMERGENCE *

I

It has been about a century and half since the ideas that we now associate with emergentism began taking shape.[1] At the core of these ideas was the thought that as systems acquire increasingly higher degrees of organizational complexity they begin to exhibit novel properties that in some sense transcend the properties of their constituent parts, and behave in ways that cannot be predicted on the basis of the laws governing simpler systems. It is now standard to trace the birth of emergentism back to John Stuart Mill and his distinction between "heteropathic" and "homopathic" laws,[2] although few of us would be surprised to learn that the same or similar ideas had been entertained by our earlier philosophical forebears.[3] Academic philosophers – like Samuel Alexander and C.D. Broad in Britain, A.O. Lovejoy and Roy Wood Sellars in the United States – played an important role in developing the concept of emergence and the attendant doctrines of emergentism, but it is interesting to note that the fundamental idea seems to have had a special appeal to scientists and those outside professional philosophy. These include the British biologist C. Lloyd Morgan, a leading theoretician of the emergentist movement early in this century, and, more recently, the noted neurophysiologist Roger W. Sperry.

In spite of its obvious and direct relevance to some of the central issues in the philosophy and methodology of science, however, emergentism failed to become a visible part of the Problematik of the mainstream philosophy of science. The main reason for this, I

* For any re-use of this article the author should be contacted directly at the following address: Brown University, Department of Philosophy, Gerard House, 54 College Street, Providence, RI 02912-9021, USA.

Philosophical Studies **95:** 3–36, 1999.

believe, is that philosophy of science during much of the middle half of this century, from the 1930s to the ′60s – at least, in the analytic tradition – was shaped by the positivist and hyper-empiricist view of science that dominated the Anglo-American philosophy at the time. Influential philosophers of science during this period – for example, Carl Hempel and Ernest Nagel[4] – claimed that the classic idea of emergence was confused and incoherent, often likening it to neo-vitalism, and what they saw as the only salvageable part of the emergence concept – the part that they could state in their own postivist/formalist idiom – usually turned out to be largely trivial, something that could be of little interest for serious philosophical purposes.

But the idea of emergence refused to die, continuing to attract a small but steady stream of advocates from both the philosophical and the scientific ranks, and it now appears to be making a strong comeback. This turn of events is not surprising, given the nearly total collapse of positivistic reductionism and the ideal of unified science which was well underway by the early ′70s. The lowly fortunes of reductionism have continued to this day, providing a fertile soil for the reemergence of emergentism. Classic emergentists like Morgan and Alexander thought of themselves as occupying a moderate intermediate position between the extremes of "mechanistic" reductionism on one hand and explicit dualisms like Cartesianism and neo-vitalism on the other. For them everything that exists is constituted by matter, or basic material particles, there being no "insertion" of alien entities or forces from the outside. It is only that complex systems aggregated out of these material particles begin to exhibit genuinely novel properties that are irreducible to, and neither predictable nor explainable in terms of, the properties of their constituents. It is evident that emergentism is a form of what is now standardly called "nonreductive materialism", a doctrine that aspires to position itself as a compromise between physicalist reductionism and all-out dualisms. It is no wonder then that we now see an increasing, and unapologetic, use of expressions like "emergent property", "emergent phenomenon", and "emergent law", substantially in the sense intended by the classic emergentists, not only in philosophical writings but in primary scientific literature as well.[5]

Does this mean that emergentism has returned – as an ontological doctrine about how the phenomena of this world are organized into autonomous emergent levels and as a metascientific thesis about the relationship between basic physics and the special sciences? I think the answer is a definite yes. The fading away of reductionism and the enthronement of nonreductive materialism as the new orthodoxy simply *amount to* the resurgence of emergentism – not all of its sometimes quaint and quirky ideas but its core ontological and methodological doctrines. The return of emergentism is seldom noticed, and much less openly celebrated; it is clear, however, that the fortunes of reductionism correlate inversely with those of emergentism (*modulo* the rejection of substantival dualism). It is no undue exaggeration to say that we have been under the reign of emergentism since the early 1970s.

I have argued elsewhere[6] against nonreductive materialism, urging that this halfway house is an inherently unstable position, and that it threatens to collapse into either reductionism or more serious forms of dualism. But in this paper I am not primarily concerned with the truth or tenability of emergentism or nonreductive materialism; rather, my main concern is with making sense of the idea of emergence – the idea that certain properties of complex systems are emergent while others are not. Even if we succeed with the conceptual task of giving a coherent sense to emergence, it is another question whether any particular group of properties is emergent – for example, whether intentional or qualitative mental properties are emergent relative to neural/biological properties, or whether biological properties are emergent relative to physicochemical properties – or indeed whether there are any emergent properties at all.

In trying to make emergence intelligible, it is useful to divide the ideas usually associated with the concept into two groups. One group of ideas are manifest in the statement that emergent properties are "novel" and "unpredictable" from knowledge of their lower-level bases, and that they are not "explainable" or "mechanistically reducible" in terms of their underlying properties. The second group of ideas I have in mind comprises the specific emergentist doctrines concerning emergent properties, and, in particular, claims about the causal powers of the emergents. Prominent among them is the claim that the emergents bring into the world new causal powers of

their own, and, in particular, that they have powers to influence and control the direction of the lower-level processes from which they emerge. This is a fundamental tenet of emergentism, not only in the classic emergentism of Samuel Alexander, Lloyd Morgan, and others but also in its various modern versions. Emergentists often contrast their position with epiphenomenalism, dismissing the latter with open scorn. On their view, emergents have causal/explanatory powers in their own right, introducing novel, and hitherto unknown, causal structures into the world.

In this paper I will adopt the following strategy: I am going to take the first group of ideas as constitutive of the idea of an emergent property, and try to give a unified account of emergence on the basis of a model of reduction that, although its basic ideas are far from new, is significantly different from the classic Nagelian model of reduction that has formed the background of debates in this area. I will then consider the doctrines that I take to constitute emergentism, focusing on the claims about the causal powers of emergent properties, especially the idea of "downward causation".

II

The concepts of explanation, prediction, and reduction figure prominently at several critical junctures in the development of the doctrine of emergence. Most importantly, the concept of explanation is invoked in the claim that emergent phenomena or properties, unlike those that are merely "resultant", are not *explainable*, or *reductively explainable*, on the basis of their "basal conditions", the lower-level conditions out of which they emerge. This is frequently coupled with the claim that emergent phenomena are *not predictable* even from the most complete and exhaustive knowledge of their emergence base. I believe that emergentists took the two claims to be equivalent, or at least as forming a single package.

Let us assume that every material object has a unique complete microstructural description: that is, any physical system can be exhaustively described in terms of (i) the basic particles that constitute it (this assumes classic atomism, which the early emergentists accepted); (ii) all the intrinsic properties of these particles; and (iii) the relations that configure these particles into a structure

(with "substantial unity", as some emergentists would say). Such a description will give us the total "relatedness" of basal constituents; it also gives us what we may call the *total microstructural (or micro-based) property* of the system – that is, a macro-property (macro since it belongs to the system as a whole) constituted by the system's basic micro-constituents, their intrinsic properties, and the relations that structure them into a system with unity and stability as a substance.[7]

I would expect most emergentists to accept mereological supervenience, in the following form:

[Mereological supervenience] Systems with an identical total microstructural property have all other properties in common.[8] Equivalently, all properties of a physical system supervene on, or are determined by, its total microstructural property.

It is a central claim of classic emergentism that among these properties supervenient on a system's total microstructural property, some have the special character of being "emergent", while the rest are only "resultant". What is the basis for this distinction? Lloyd Morgan says this:

> The concept of emergence was dealt with (to go no further back) by J.S. Mill . . . The word 'emergent', as contrasted with 'resultant', was suggested by G.H. Lewes . . . Both adduce examples from chemistry and from physiology; both deal with properties; both distinguish those properties (a) which are additive and subtractive only, and predictable, from those (b) which are new and unpredictable.[9]

There is no need to interpret the talk of "additivity" and "subtractability" literally; I believe these terms were used to indicate that resultant properties are simply and straightforwardly calculated and predicted from the base properties. But obviously ease and simplicity of calculation as such is of no relevance here; predictability is not lost or diminished if calculationally complex mathematical/logical procedures must be used.[10] I believe that predictability is the key idea here, and that an appropriate notion of predictability must be explained in terms that are independent of addition or subtraction, or the simplicity of mathematical operations.

In any case, resultant properties are to be those that are predictable from a system's total microstructural property, but emergent

properties are those that are not so predictable. Morgan's (b) above introduces the idea of "newness", or "novelty", an idea often invoked by the emergentists. Is he using "new" and "unpredictable" here as expressing more or less the same idea, or is he implying, or at least hinting, that emergent properties are unpredictable *because* they are new and novel properties? I believe that "new" as used by the emergentists has two dimensions: an emergent property is new because it is unpredictable, and this is its epistemological sense; and, second, it has a metaphysical sense, namely that an emergent property brings with it new causal powers, powers that did not exist before its emergence. We will discuss the causal issue in the latter part of this paper.

In speaking of predictability, it is important to distinguish between *inductive predictability* and *theoretical predictability*, a distinction that I believe the emergentists were clearly aware of. Even emergent properties are inductively predictable: Having observed that an emergent property, *E*, emerged whenever any system instantiated a microstructural property *M*, we may predict that this particular system will instantiate *E* at *t*, given our knowledge or belief that it will instantiate, *M*, at *t*.[11] More generally, on the basis of such empirical data we may have a well-confirmed "emergence law" to the effect that whenever a system instantiates basal condition *M* it instantiates an emergent, *E*. What is being denied by emergentists is the theoretical predictability of *E* on the basis of *M*: we may know all that can be known about *M* – in particular, laws that govern the entities, properties and relations constitutive of *M* – but this knowledge does not suffice to yield a prediction of *E*. This unpredictability may be the result of our not even having the *concept* of *E*, this concept lying entirely outside the concepts in which our theory of *M* is couched. In cases where *E* is a phenomenal property of experiences (a "quale"), we may have no idea what *E* is like before we experience it.[12] But this isn't the only barrier to predictability. It may be that we know what *E* is like – we have already experienced *E* – but we may be powerless to predict whether or not *E* – or whether *E* rather than another emergent E^* – will emerge when a complex is formed with a novel microstructure M^* that is similar to *M* is some significant respects. In such a case the emergence law "Whenever a system instantiates *M*, it instantiates

E" would have to be taken as a primitive, stating a brute correlation between *M* and *E*.

It is clear that we can inductively predict – in fact, we do this all the time – the occurrences of conscious states in the sense just explained, but, if the emergentists were right about anything, they were probably right about the phenomenal properties of conscious experience: these properties appear not to be theoretically predictable on the basis of a complete knowledge of the neurophysiology of the brain. This is reflected in the following apparent difference between phenomenal properties and other mental properties (including cognitive/intentional properties): We can imagine designing and constructing novel physical systems that will instantiate certain cognitive capacities and functions (e.g., perception, information processing, inference and reasoning, and using information to guide behavior) – arguably, we have already designed and fabricated such devices in robots and other computer-driven mechanisms. But it is difficult to imagine our designing novel devices and structures that will have phenomenal experiences; I don't think we have any idea where to begin. The only way we can hope to manufacture a mechanism with phenomenal consciousness is to produce an appropriate physical duplicate of a system that is known to be conscious. Notice that this involves inductive prediction, whereas theoretical prediction is what is needed to design new physical devices with consciousness. The emergentists were wrong in thinking that sundry chemical and biological properties were emergent;[13] but this was an understandable mistake given the state of the sciences before the advent of solid-state physics and molecular biology. The interest of the ideas underlying the emergentist's distinction between the two kinds of properties need not be diminished by the choice of wrong examples.

III

As was noted at the start of our discussion, another idea that is closely related to the claimed unpredictability of emergents is the doctrine that the emergence of emergent properties cannot be *explained* on the basis of the underlying processes, and that emergent properties are not *reducible* to the basal conditions from which

they emerge. These two claims can be combined into one: Emergent properties are not *reductively explainable* in terms of the underlying processes. Some may wish to distinguish the issue of reduction from that of reductive explanation;[14] we will address this issue later. I will now turn to the task of describing a model of reduction that connects and makes sense of these three ideas, namely that emergent properties are *not predictable* from their basal conditions, that they are *not explainable* in terms of them, and that they are *not reducible* to them.

Let me begin with an example – an idealized, admittedly somewhat simplistic example. To reduce the gene to the DNA molecule, we must first prime the target property, by giving it a *functional* interpretation – that is, by construing it in terms of the causal work it is to perform. Briefly, the property of being a gene is the property of having some property (or being a mechanism) that performs a certain causal function, namely that of transmitting phenotypic characteristics from parents to offsprings. As it turns out, it is the DNA molecule that fills this causal specification ("causal role"), and we have a theory that explains just how the DNA molecule is able to perform this causal work. When all of this is in, we are entitled to the claim that the gene has been reduced to the DNA molecule.

We can now formulate a general model to accommodate reductions of this form. Let **B** be the domain of properties (also phenomena, facts, etc., if you wish) serving as the reduction base – for us, these contain the basal conditions for our emergent properties. The reduction of property E to **B** involves three steps:[15]

Step 1: E must be *functionalized* – that is, E must be construed, or reconstrued, as a property defined by its causal/nomic relations to other properties, specifically properties in the reduction base **B**.

We can think of a functional definition of E over domain **B** as typically taking the following (simplified) form:

Having $E =_{\text{def}}$ Having some property P in **B** such that (i) $C_1, \ldots, C_n$[16] cause P to be instantiated, and (ii) P causes $F_1, \ldots, F_m$ to be instantiated.

(We allow either (i) or (ii) to be empty.) The main point to notice is that the functionalization of E makes E nonintrinsic and rela-

tional – relational with respect to other properties in **B**. *E*'s being instantiated is for a certain property *P* to be instantiated, with this instantiation bearing causal/nomic relations to the instantiations of a specified set of properties in the base domain. We may call any property *P* in **B** that satisfies the causal specification (i) and (ii) a "realizer" or "implementer" of *E*. Clearly, multiple realizers for *E* are allowed on this account; so multiply realizable properties fall within the scope of the present model of reduction. A functionalization of property *E* in the present sense is to be taken as establishing a conceptual/definitional connection for *E* and the selected causal role. An important part of this procedure is to decide how much of what we know (or believe) about *E*'s nomic/causal involvement should be taken as *defining*, or *constitutive of*, *E* and how much will be left out. We should keep in mind that such conceptual decisions can be and often are based on empirical knowledge, knowledge of the causal/nomic relations in which *E* is embedded, and can be constrained by theoretical desiderata of various sorts, and that in practice the boundary between what's conceptual and what isn't is certain to be a vague and shifting one.

Step 2: Find realizers of *E* in **B**. If the reduction, or reductive explanation, of a particular instance of *E* in a given system is wanted, find the particular realizing property *P* in virtue of which *E* is instantiated on this occasion in this system; similarly, for classes of systems belonging to the same species or structure types.

This of course is a scientifically significant part of the reductive procedure; it took many years of scientific research to identify the DNA as a realizer of the gene.

Step 3: Find a theory (at the level of **B**) that explains how realizers of *E* perform the causal task that is constitutive of *E* (i.e., the causal role specified in Step 1). Such a theory may also explain other significant causal/nomic relations in which *E* plays a role.

We presumably have a story at the microbiological level about how DNA molecules manage to code and transmit genetic information. When temperature, for gases, is reduced to mean translational kinetic energy of molecules (another over-simplified stock

example[17]), we have a theory that explains the myriad causal/nomic relations in which temperature plays a role. Steps 2 and 3 can be expected to be part of the same scientific research: ascertaining realizers of *E* will almost certainly involve theories about causal/nomic interrelations among lower-level properties in the base domain.

Notice how this functional conception of reduction differs from the classic Nagel model of intertheoretical reduction[18] – in particular, there is no talk of "bridge laws" or "derivation" of laws. The question whether appropriate bridge laws are available that connect the domain to be reduced with the base domain – more specifically, whether or not there are bridge laws providing for each property to be reduced a nomically coextensive property in the base domain – has been at center stage in debates over reduction and reductionism. However, from the emergentist point of view, the bridge laws, far from being the enablers of reduction (as they are in Nagel reductions), are themselves among the targets of reduction. For it is these bridge laws, laws that state that whenever certain specified basal conditions are present a certain novel property is manifested, that the emergentists were anxious to have explained. Why is it that pain, not itch or tickle, occurs when a certain neural condition (e.g., C-fiber stimulation) holds? Why doesn't pain accompany conditions of a different neural type? Why does *any* phenomenal consciousness occur when these neural conditions are present? These are the kinds of explanatory/reductive questions with which the emergentists were preoccupied. And I think they were right. The "mystery" of consciousness is not dispelled by any reductive procedure that, as in Nagel reduction, takes these bridge laws as brute unexplained primitives.

The philosophical emptiness of Nagel reduction, at least in contexts like mind-body reduction, if it isn't already evident, can be plainly seen from the following fact: a Nagel reduction of the mental to the physical is consistent with, and sometimes even entailed by, many dualist mind-body theories, such as the double-aspect theory, the theory of preestablished harmony, occasionalism, and epiphenomenalism. It is not even excluded by the dualism of mental and physical substances (although Descartes' own interactionist version probably excludes it). This amply shows that the antireductionist argument based on the unavailability of mind-body bridge laws –

most importantly, the multiple realization argument of Putnam and Fodor – is irrelevant to the real issue of mind-body reduction or the possibility of giving a reductive explanation of mentality. Much of the debate over the past two decades about reductionism has been carried on in terms of an inappropriate model of reduction, and now appears largely beside the point for issues of real philosophical significance.

IV

Let us now try to see how the functional model of reduction can meet the explanatory/predictive/ontological demands that reductions of genuine philosophical interest must meet. Let *E* be the property targeted for reduction, where *E* has been functionalized as the property of having some property *P* meeting casual specification *C*.

1. *The Explanatory Question*

Why does this system exhibit *E* at *t*? Because having *E* is, by definition, having a property with causal role *C*, and the system, at *t*, has property *Q*, which fills causal role *C* (and hence realizes *E*). Moreover, we have a theory that explains exactly how *Q* manages to fill *C*.

Why do systems exhibit *E* whenever they instantiate *Q*? Because *E* is a functional property defined by causal role *C*, and *Q* is a realizer of *E* for these systems. And there is a theory that explains how *Q* realizes *E* in these systems.

Suppose that pain could be given a functional definition – something like this: being in pain is being in some state (or instantiating some property) caused by tissue damage and causing winces and groans. Why are you experiencing pain? Because being in pain *is* being in a state caused by tissue damage and causing winces and groans, and you are in neural state *N*, which is one of those states (in you, or in systems like you) that are caused by tissue damage and that cause winces and groans. Why do people experience pain when they are in neural state *N*? Because *N* is implicated in these causal/nomic relations, and being in pain is being in some state with just these causal/nomic relations. It is clear that in this way all our explanatory demands can be met. There is nothing further to be explained about

why pain occurs, or why pain occurs when neural condition *N* is present.

But is this a *reductive* explanation? This question is connected with the question whether, and in what sense, the proposed model is a model of reduction, a question that will be considered below.

It is of course another question whether pain can be functionalized. We will briefly return to this issue later, but our concern here is to give a clear sense to what it is to "reduce" pain.

2. *The Predictive Question*

Will this system exhibit *E* at time *t*? Can we predict this from knowledge of what goes on in the base domain? Yes, because, given the functional definition of *E*, we can in principle identify the realizers of *E* for the system solely on the basis of knowledge of the causal/nomic relations obtaining in the base domain. Given this knowledge of *E*'s realizers for this system, we can predict whether or not the system will, at *t*, instantiate property *E* from our knowledge, or warranted belief, that it will, or will not, instantiate a realizer of *E* at *t*.

Clearly, what enables the ascent from the reduction base to higher properties is the conceptual connections generated by the functionalization of the higher properties. This is in sharp contrast to Nagelian reduction with bridge laws taken as auxiliary premises. These laws are standardly conceived as empirical and contingent, and must be viewed as net additions to our theory about the reduction base, which means that *the base theory so augmented is no longer a theory exclusively about the originally given base domain*. This is why bridge laws only enable inductive predictions, whereas functionalization makes theoretical predictions possible.

These reflections seem to give us an answer to a question we raised earlier – why we seem to lack the ability to design novel physical devices that will exhibit phenomenal consciousness: it is because brute bridge laws may be all we can get to connect phenomenal properties with physical properties, whereas what is required is an ability to make theoretical predictions of qualia solely on the basis of knowledge of the base domain, namely physics, chemistry, biology, and the like. The functionalization of phenomenal experience would give us such an ability.

3. *The Ontological Question*

In what sense is the functional model a model of *reduction*? What does it reduce, and how does it do it? Central to the concept of reduction evidently is the idea that what has been reduced need not be countenanced as an *independent* existent beyond the entities in the reduction base – that if *X* has been reduced to *Y*, *X* is not something "over and above" *Y*. From an ontological point of view, reduction must mean *reduction* – it must result in a simpler, leaner ontology. Reduction is not necessarily elimination: reduction of *X* to *Y* need not do away with *X*, for *X* may be conserved as *Y* (or as part of *Y*). Thus, we can speak of "conservative" reduction (some call this "retentive" reduction), reduction that conserves the reduced entities, as distinguished from "eliminative" reduction, which rids our ontology of reduced entities. Either way we end up with a leaner ontology. Evidently, conservative reduction requires identities, for to conserve *X as Y* means that *X is Y*, whereas eliminative reduction has no need for reductive identities.

Our question, then, is in what ways the model of reduction being recommended here serves the cause of ontological simplification. Two cases may be distinguished: the first concerns instances of property *E*; the second concerns property *E* itself.

First, consider property instances: system *s* has *E*, in virtue of *s*'s instantiating one of its realizers, say *Q*. Now, *s*'s having *E* on this occasion just is its having some property meeting causal specification *C*, and in this particular instance, *s* has *Q*, where *Q* meets specification *C*. Thus, *s*'s having *E* on this occasion is identical with its having *Q* on this occasion. There is no fact of the matter about *s*'s having *E* on this occasion over and above *s*'s having *Q*. Each instance of *E*, therefore, is an instance of one of *E*'s realizers, and all instances of *E* can be partitioned into Q_1-instances, Q_2-instances, ..., where the *Q*'s are *E*'s realizers. Hence, the *E*-instances reduce to the Q_i-instances.

Suppose someone were to object as follow: There is no good reason to identify this instance of *E* with the instance of *Q* in virtue of which *E* is realized on this occasion. Rather, *s*'s having *E* should be identified with *s*'s having some property or other meeting causal specification *C*, and this latter instance is not identical with *s*'s having *Q*. For having some property or other meeting *C* is not the

same property as having *Q*; that is, property *E* ≠ property *Q*. How should we counter this line of argument? I think it will be helpful to consider the causal picture, and ask: What are the *causal powers* of *this instance of E*, namely *s*'s having *E* on this occasion? If *s* has *E* in virtue of *E*'s realizer *Q*, it is difficult to see how we could avoid saying this: the causal powers of this instance of *E* are exactly the causal powers of this instance of *Q*. This is what I have elsewhere called the "causal inheritance principle":

If a functional property *E* is instantiated on a given occasion in virtue of one of its realizers, *Q*, being instantiated, then the causal powers of this instance of *E* are identical with the causal powers of this instance of *Q*.

If this principle is accepted, the *E*-instance and the *Q*-instance have identical causal properties, and this exerts powerful pressure to identify them. What good would it do to count them as different? If they were different, the difference could not even be detected.

This means that on the present picture *E*-instances are conservatively reduced to *Q*-instances, instances of *E*'s realizers. Let us now turn to the reduction of *E*, the property itself. Here we need to come to terms with *E*'s having multiple realizers, Q_1, Q_2, ... There are three possible approaches here.

First, one may choose to defend *E* as a legitimate higher-level property irreducible to its realizers, the *Q*'s. This is the position taken by many functionalists: psychological properties are functional properties defined in terms of input/output correlations, with internal physical/biological properties as realizers, and yet they are irreducible to their realizers, constituting an autonomous domain for the special science of psychology (cognitive science, or whatever).

Second, one may choose to identify *E* with the disjunction of its realizers, Q_1 v Q_2 v ...[19] Notice, though, that this identity is not necessary – it does not hold in every possible world – since whether or not a property realizes *E* depends on the laws that prevail at a given world. The reason is that *E* is defined in terms of a causal/nomic condition, and whether something satisfies such conditions depends on the laws that are in force at a given world.

This means that in another world with different laws, *E* may have a wholly distinct set of realizers, and in still others *E* may have no realizers at all. So the identity, $E = Q_1 \vee Q_2 \vee \ldots$ is metaphysically contingent, although nomologically necessary, and "*E*" becomes nonrigid, although it remains nomologically rigid or "semi-rigid" (as we may say). For example, in a world with laws quite different from those prevailing in this world, molecules of another kind, not the DNA molecules, may perform the causal task of coding and transmitting genetic information.[20]

Third, we may give up *E* as a genuine property and only recognize the expression "*E*" or the concept *E*. As it turns out, many different properties are picked out by the concept *E*, depending on the circumstances – the kind of structures involved and the nomological nature of the world under consideration. One could argue that by forming "second-order" functional *expressions* by existentially quantifying over "first-order" properties, we cannot be generating new properties (possibly with new causal powers), but only new ways of indifferently picking out, or grouping, first-order properties, in terms of causal specifications that are of interest to us.[21] As noted, the concept is only nomologically rigid: it picks out the same properties only across worlds that are similar in causal/nomological respects.

Here I will not argue my points in detail. It is clear, however, that the second and third approach effectively reduce the target property *E*: the second is a conservative reduction, retaining *E* as a disjunction of properties in the base domain. In contrast, the third is eliminative: it recommends the elimination of *E* as a property, retaining only the concept *E* (which may play a practically indispensable role in our discourse, both ordinary and scientific). The first approach, as I said, is one that is widely accepted: many philosophers, in spite of (or, in their view, on account of) multiple realization, want to argue that *E* is an irreducible property that nonetheless can be a property playing an important role in a special, "higher-level", science. I believe, however, that this position cannot be sustained. For if the "multiplicity" or "diversity" of realizers means anything, it must mean that these realizers are causally and nomologically diverse. Unless two realizers of E show significant causal/nomological diversity, there is no clear reason why we should count them as two, not one. It

follows then that multiply realizable properties are ipso facto causally and nomologically heterogeneous. This is especially obvious when one reflects on the causal inheritance principle. All this points to the inescapable conclusion that *E*, because of its causal/nomic heterogeneity, is unfit to figure in laws, and is thereby disqualified as a useful scientific property. On this approach, then, one could protect *E* but not as a property with a role in scientific laws and explanations. You could insist on the genuine propertyhood of *E* as much as you like, but the victory would be empty.[22] The conclusion, therefore, has to be this: as a significant scientific property, *E* has been reduced – eliminatively.

What I hope I have shown is this: Functionalization of a property is both necessary and sufficient for reduction (sufficient as a first conceptual step, the rest being scientific research). This accords well with the classic doctrines of emergentism: as I argued, it nicely explains why reducible properties are predictable and explainable, and correlatively it explains why irreducible properties are neither predictable nor explainable on the basis of the underlying processes. I believe this makes good sense of the central ideas that make up the concept of emergence.

However, emergentism may yet be an empty doctrine. For there may not be any emergent properties, all properties being physical properties or else functionalizable and therefore reducible to physical properties. Physical properties include not only basic physical magnitudes and the properties of microparticles but microstructual properties of larger complexes of basic particles. So are there emergent properties? Many scientists have argued that certain "self-organizing" phenomena of organic, living systems are emergent. But it is not clear that these are emergent in our sense of nonfunctionalizability.[23] And, as I said earlier, the classic emergentists were mostly wrong in putting forward examples of chemical and biological properties as emergent. It seems to me that if anything is going to be emergent, the phenomenal properties of consciousness, or "qualia", are the most promising candidates. Here I don't want to rehearse the standard arguments pro and con, but merely affirm, for what it's worth, my own bias toward the pro side: qualia are intrinsic properties if anything is, and to functionalize them is to eliminate them as intrinsic properties.[24]

V

The doctrine of emergence has lately been associated quite closely with the idea of "downward causation". It is not only that emergent properties are to have their own distinctive causal powers but also that they be able to exercise their causal powers "downward" – that is, with respect to processes at lower-levels, levels from which they emerge. The claim that emergents have causal powers is entirely natural and plausible if you believe that there are such properties. For what purpose would it serve to insist on the existence of emergent properties if they were mere epiphenomena with no causal or explanatory relevance?

The very idea of downward causation involves vertical directionality – an "upward" direction and a "downward" direction. This in turn suggests an ordered hierarchy of domains that gives meaning to talk of something being located at a "higher" or "lower" or "the same" position in relation to another item on this hierarchy. As is familiar to everyone, positions on such a hierarchy are usually called "levels", or sometimes "orders". In fact, talk of "levels" – as in "level of description", "level of explanation", "level of organization", "level of complexity", "level of analysis", and the like – has thoroughly penetrated not only writings about science, including of course philosophy of science, but also the primary scientific literature of many fields.

The emergentists of the early 20th century were among the first to articulate what may be called "the layered model" of the world, although a general view of this kind is independent of emergentism and has been espoused by those who are opposed to emergentism.[25] In fact, a model of this kind provides an essential framework needed to formulate the emergentist/reductionist debate. In any case, the layered model takes the natural world as stratified into levels, from lower to higher, from the basic to the constructed and evolved, from the simple to the more complex. All objects and phenomena have each a unique place in this ordered hierarchy. Most early emergentists, such as Samuel Alexander and C. Lloyd Morgan, viewed this hierarchy to have evolved historically: In the beginning there were only basic physical particles, or just a spacetime framework (as Alexander maintained), and these have evolved into increasingly more complex structures – atoms, molecules, unicellular organ-

isms, multicellular organisms, organisms with consciousness and mentality, and so on. Contemporary interest in emergence and the hierarchical model is focused not on this kind of quasi-scientific and quasi-metaphysical history of the world, but rather on what it says about the synchronic structure of the world – how things and phenomena at different levels hang together in a temporal cross section of the world, or over small time intervals. We want to know whether, and how, the emergentist ideas can help us in understanding the interlevel relationships between items at the adjacent levels on this hierarchy, and ultimately how everything is related to the items at the bottom physical level (if there is such a level).

The layered model gives rise to many interesting questions: for example, how are these levels to be defined and individuated? Is there really a single unique hierarchy of levels that encompasses all of reality or does this need to be contextualized or relativized in certain ways? Does a single ladder-like structure suffice, or is a branching tree-like structure more appropriate? Exactly what ordering relations generate the hierarchical structures? But these questions go well beyond the scope of this paper. Here we will work with a fairly standard, intuitive notion of levels that is shared by most of us.[26] This will not significantly compromise the discussion to follow.

Although, as one would expect, there has been no universal agreement among the emergentists, the central doctrines of emergentism are well known. For our present purposes, we will take them to include the following claims:

1. *Emergence of complex higher-level entities*: Systems with a higher-level of complexity emerge from the coming together of lower-level entities in new structural configurations (the new "relatedness" of these entities).

This claim is by no means unique to emergentism; it is completely at home with universal physical reductionism (what the early emergentists called "mechanism"), the view that all things and phenomena are physical, and are explainable and predictable ultimately in terms of fundamental physical laws. A characteristically emergentist doctrine makes its appearance in the idea that some of the properties of these complex systems, though physically grounded, are

nonphysical, and belong outside the physical domain. The following three propositions unpack this idea.

2. *Emergence of higher-level properties*: All properties of higher-level entities arise out of the properties and relations that characterize their constituent parts. Some properties of these higher, complex systems are "emergent", and the rest merely "resultant".

Instead of the expression "arise out of", such expressions as "supervene on" and "are consequential upon" could have been used. In any case, the idea is that when appropriate lower-level conditions are realized in a higher-level system (that is, the parts that constitute the system come to be configured in a certain relational structure), the system will necessarily exhibit certain higher-level properties, and, moreover, that no higher-level property will appear unless an appropriate set of lower-level conditions is realized. Thus, "arise" and "supervene" are neutral with respect to the emergent/resultant distinction: both emergent and resultant properties of a whole supervene on, or arise out of, its microstructural, or micro-based, properties.

The distinction between properties that are emergent and those that are merely resultant is a central component of emergentism. As we have already seen, it is standard to characterize this distinction in terms of predictability and explainability.

3. *The unpredictability of emergent properties*: Emergent properties are not predictable from exhaustive information concerning their "basal conditions". In contrast, resultant properties are predictable from lower-level information.

4. *The unexplainability/irreducibility of emergent properties*: Emergent properties, unlike those that are merely resultant, are neither explainable nor reducible in terms of their basal conditions.

Earlier in this paper we saw how it is possible to give unity to these claims on the basis of an appropriate model of reduction. More specifically, by identifying emergent properties with irreducible properties, on the functional model of reduction, it is possible to explain why emergent properties are neither explainable nor predictable on the basis of the conditions from which they emerge,

whereas nonemergent (or resultant) properties are so explainable and predictable.

Our present concern, however, lies with the question what emergent properties, after having emerged, can *do* – that is, how they are able to make their special contributions to the ongoing processes of the world. It is obviously very important to the emergentists that emergent properties can be active participants in causal processes involving the systems they characterize. None perhaps understood this better than Samuel Alexander, who made the following pointed comment on epiphenomenalism, the doctrine that mental properties are wholly lacking in causal powers:

> [Epiphenomenalism] supposes something to exist in nature which has nothing to do, no purpose to serve, a species of *noblesse* which depends on the work of its inferiors, but is kept for show and might as well, and undoubtedly would in time be abolished.[27]

We may, therefore, set forth the following as the fifth doctrine of emergentism:

5. *The causal efficacy of the emergents*: Emergent properties have causal powers of their own – novel causal powers irreducible to the causal powers of their basal constituents.

In what ways, then, can emergent properties manifest their causal powers?

This of course is where the idea of "downward causation" enters the scene. But when we view the situation with the layered model in mind, we see that the following three types of inter- or intra-level causation must be recognized: (i) *same-level causation*, (ii) *downward causation*, and (iii) *upward causation*. Same-level causation, as the expression suggests, involves causal relations between two properties at the same level – including cases in which an instantiation of one emergent property causes another emergent property to be instantiated. Downward causation occurs when a higher-level property, which may be an emergent property, causes the instantiation of a lower-level property; similarly, upward causation involves the causation of a higher-level property by a lower-level property. I believe that, for the emergentist,[28] there is good reason to believe that downward causation is fundamental and of crucial importance in understanding causation. For it can be shown that both upward

and same-level causation (except same-level causation at the ultimate bottom level, if there is such a level and if there are causal relations at this level) presupposes the possibility of downward causation.

Here is an argument that shows why this is so.[29] Suppose that a property M, at a certain level L, causes another property M^+, at level $L + 1$. Assume that M^+ emerges, or results, from a property M^* at level L (M^* therefore is on the same level as M). Now we immediately see a tension in this situation when we ask: "What is responsible for this occurrence of M^+? What explains M^+'s instantiation on this occasion?" For in this picture there initially are two competing answers: First, M^+ is there because, *ex hypothesi*, M caused it; second, M^+ is there because its emergence base M^* has been realized. Given its emergence base M^*, M^+ must of necessity be instantiated, no matter what conditions preceded it; M^* alone suffices to guarantee M^+'s occurrence on this occasion, and without M^*, or an appropriate alternative base, M^+ could not have occurred. This apparently puts M's claim to have caused M^+ in jeopardy. I believe that the only coherent description of the situation that respects M's causal claim is this: M causes M^+ *by causing its base condition* M^*. But M's causation of M^* is an instance of same-level causation. This shows that upward causation entails same-level causation; that is, upward causation is possible only if same-level causation is possible.

As an example, consider this: physical/mechanical work on a piece of marble (M) causes the marble to become a beautiful sculpture (M^+). But the beauty of the sculpture emerges from the physical properties (M^* consisting in shape, color, texture, size, etc.) of the marble piece. Notice how natural, and seemingly unavoidable, it is to say that the physical work on the marble caused the beauty of the marble piece *by causing it to have the right physical properties*. This of course is an instance of same-level causation. Another example: a bee sting causes a sharp pain. But pain emerges from a certain neural condition N (say, C-fiber excitation). I believe that we want to say, and must say, that the bee sting caused the pain by causing N (the firing of C-fibers). This again is same-level causation.

An exactly similar argument will show that same-level causation presupposes downward causation. Briefly, this can be shown

as follows: Suppose M causes M^*, where M and M^* are both at level L. But M^* itself arises out of a set of properties M^- at level $L - 1$. When we ponder the question how M^* gets to be instantiated on this occasion, again we come to the conclusion that M caused M^* to be instantiated on this occasion *by causing* M^-, its base condition, to be instantiated. But M's causation of M^- is downward causation. This completes the argument.

A general principle is implicit in the foregoing considerations, and it is this:

To cause any property (except those at the very bottom level) to be instantiated, you must cause the basal conditions from which it arises (either as an emergent or as a resultant).

We may call this "the principle of downward causation".

VI

Even the early emergentists were explicit on the importance they attached to downward causation, although of course it is unlikely that they were influenced by anything like the argument of the preceding section. The following statement by C. Lloyd Morgan is typical:

> Now what emerges at any given level affords an instance of what I speak of as a new kind of relatedness of which there are no instances at lower levels . . . But when some new kind of relatedness is supervenient (say at the level of life), *the way in which the physical events which are involved run their course is different in virtue of its presence – different from what it would have been if life had been absent*.[30]

Compare this with what Roger Sperry says over 50 years later:

> . . . the conscious subjective properties in our present view are interpreted to have causal potency in regulating the course of brain events; that is, the mental forces or properties exert a regulative control influence in brain physiology.[31]

Both Morgan and Sperry are saying that life and consciousness, emergent properties out of physicochemical and neural properties respectively, have a causal influence on the flow of events at the lower levels, levels from which they emerge. That of course is downward causation.

The appearance of an emergent property signals, for the emergentists, a genuine change, a significant evolutionary step, in the history of the world, and this requires emergent properties to be genuine properties with causal powers. They are supposed to represent novel additions to the ontology of the world, and this could be so only if they bring with them *genuinely new* causal powers; that is, they must be capable of making novel causal contributions that go beyond the causal powers of the lower-level basal conditions from which they emerge.

But how do emergent properties exercise their novel causal powers? How is that possible? According to the argument presented in the preceding section, they can do so only by causally influencing events and phenomena at lower-levels – that is, through downward causation. That was what we called the principle of downward causation. But is downward causation possible? The idea of downward causation has struck some thinkers as incoherent, and it is difficult to deny that there is an air of paradox about it: After all, higher-level properties arise out of lower-level conditions, and without the presence of the latter in suitable configurations, the former could not even be there. So how could these higher-level properties causally influence and alter the conditions from which they arise? Is it coherent to suppose that the presence of *X* is entirely responsible for the occurrence of *Y* (so *Y*'s very existence is totally dependent on *X*) and yet *Y* somehow manages to exercise a causal influence on *X*? I believe a train of thought like this is behind the suspicions surrounding the idea of downward causation. But if downward causation is incoherent, that alone will do serious damage to emergentism. For the principle of downward causation directly implies that if emergent properties have no downward causal powers, they can have no causal powers at all, and this means that emergent phenomena would just turn out to be epiphenomena, a prospect that would have severely distressed Alexander, Morgan, and Sperry.

But we need to analyze whether the kind of intuitive argument in the preceding paragraph against downward causation has any real force. For cases in which higher-level entities and their properties prima facie causally influence lower-level entities and their properties seem legion. The celadon vase on my desk has a mass of 1 kilogram. If it is dropped out the window of my second floor office,

it will crash on the paved sidewalk, causing myriads of molecules of all sorts to violently fly away in every which direction. Even before it hits the ground, it will cut a rapid downward swath, causing all sorts of disturbance among the local air molecules. And these effects are surely micro and lower-level in relation to the fall of an object with a mass of 1 kilogram. Note that we cannot think of this case as one in which the "real" causal process occurs at the micro-level, between the micro-constituents of the vase and the air molecules, for the simple reason that no micro-constituents of the vase, in fact no proper part, of my celadon vase has a mass of 1 kilogram. There is no question that the vase, in virtue of having this mass, has a set of causal powers that none of its micro-constituents have; the causal powers that this property represents cannot be reduced to the causal powers of micro-constituents of its bearers. Of course, emergentists would not consider mass an emergent property; they would say that the mass of an object is a resultant property, a property that is merely "additive or subtractive". But this simple example suffices to show that there need not be anything strange or incoherent in the idea of downward causation as such – the idea that complex systems, in virtue of their macrolevel properties, can cause changes at lower microlevels.

However, the idea of downward causation advocated by some emergentists is stronger and more complex than what is suggested by our example. Here again is Sperry:

> The subjective mental phenomena are conceived to influence and govern the flow of nerve impulse traffic by virtue of their encompassing emergent properties. Individual nerve impulses and other excitatory components of a cerebral activity pattern are simply carried along or shunted this way and that by the prevailing overall dynamics of the whole active process (in principle – just as drops of water are carried along by a local eddy in a stream or the way the molecules and atoms of a wheel are carried along when it rolls down hill, regardless of whether the individual molecules and atoms happen to like it or not).[32]

Sperry has used these and other similar analogies elsewhere; in particular, the rolling wheel seems to have been one of his favorites. What is distinctive about this form of downward causation appears to be this: Some activity or event involving a whole *W* is a cause of, or has a causal influence on, the events involving its *own* micro-constituents. We may call this *reflexive downward causation*, to distinguish it from the more mundane *nonreflexive* kind, involved in

the example of the falling vase above, in which an event involving a whole causes events involving lower-level entities that are not among its constituents.

But downward causation must be viewed in the context of the doctrine that emergent properties arise out of their basal conditions (claim 2. in section V). For Sperry himself recognizes this in his claim that there is also *upward determination* in this situation. The paragraph quoted above from Sperry continues as follows:

> Obviously, it also works the other way around, that is, the conscious properties of cerebral patterns are directly dependent on the action of the component neural elements. Thus, a mutual interdependence is recognized between the sustaining physico-chemical processes and the enveloping conscious qualities. The neurophysiology, in other words, controls the mental effects, and the mental properties in turn control the neurophysiology.[33]

After all, an eddy is there because the individual water molecules constituting it are swirling around in a circular motion in a certain way; in fact, an eddy *is nothing but* these water molecules moving in this particular pattern. Take away the water molecules, and you have taken away the eddy: there cannot be a disembodied eddy still swirling around without any water molecules! Thus, reflexive downward causation is combined with upward determination. When each and every molecule in a puddle of water begins to move in an appropriate way – and only then – will there be an eddy of water. But in spite of this, Sperry says, it remains true that the eddy is moving the molecules around "whether they like it or not".

Thus, reflexive downward causation is combined with upward determination. Schematically, the situation looks like this: a whole, W, has a certain (emergent) property M; W is constituted by parts, $a_1, \ldots, a_n$, and there are properties $P_1, \ldots, P_n$ respectively of $a_1, \ldots, a_n$ and a certain relation R holding for the a_is. The following two claims make explicit what Sperry seems to have in mind (I do not want to rule out other possible interpretations of Sperry):

(i) [Downward causation] W's having property M causes some a_j to have P_j; but

(ii) [Upward determination] each a_i's having P_i and R holding for the a_is together determine W to have M – that is, W's having M depends wholly on (or is wholly constitued by) the a_is having the P_i respectively and being related by R.

The question is whether or not it is possible, or coherent, to hold both (i) and (ii).

VII

As I said, downward causation as such presents us with no special problems; however, what Sperry wants (also there is a hint of this in the quotation from Lloyd Morgan above) is the reflexive variety of downward causation. But how is it possible for the whole to causally affect its constituent parts on which its very existence and nature depend? If causation or determination is transitive, doesn't this ultimately imply a kind of self-causation, or self-determination – an apparent absurdity? It seems to me that there is reason to worry about the coherence of the whole idea.

Let us see if it is possible to make reflective downward causation intelligible. To sharpen the issues we should distinguish two cases:

Case 1. At a certain time *t*, a whole, *W*, has emergent property *M*, where *M* emerges from the following configuration of conditions: *W* has a complete decomposition into parts $a_1, \ldots, a_n$; each a_i has property P_i; and relation *R* holds for the sequence $a_1, \ldots, a_n$. For some a_j, *W*'s having *M* at *t* causes a_j to have P_j at *t*.

Note that the time *t* is fixed throughout, and both the downward causation and upward emergence (or determination) hold for states or conditions occurring at the very same time. We may, therefore, call this "synchronic reflexive downward causation".[34] A whole has a certain emergent property, *M*, at a given time, *t*, and the fact that this property emerges at *t* is dependent on its having a certain micro-configuration at *t*, and this includes a given constituent of it, a_j, having P_j at *t*. That is, unless a_j had P_j at *t*, *W* could not have had its emergent property *M* at *t*. Given this, it makes one feel uncomfortable to be told *also* that a_j is caused to have P_j at that very time, *t*, by the whole's having *M* at *t*.

But what exactly is the source of this metaphysical discomfort? Why does this picture seem in some way circular and incoherent? Moreover, what is it about causal circularity that makes it unacceptable? One possible explanation, something I find plausible myself, is that we tacitly subscribe to a metaphysical principle like the following:

For an object, *x*, to exercise, at time *t*, the causal/determinative powers it has in virtue of having property *P*, *x* must *already* possess *P* at *t*. When *x* is caused to acquire *P* at *t*, it does not already possess *P* at *t* and is not capable of exercising the causal/determinative powers inherent in *P*.

If a name is wanted, we may call this "the causal-power actuality principle". The reader will have noticed that this principle has been stated in terms of an object "acquiring" property *P* at a time. In Case 1 above, we said that the whole, *W*, causes one of its proper parts, a_j, to "have" *P*. If there is real downward causation, from *W*'s having *M* to a_j's having *P*, this "having" must be understood as "acquiring". For if a_j already has P_j at *t*, what role can W's having *M* at *t* play in causing it to have P_j at *t*? Obviously, none.

In any case, it is now easy to see the incoherence involved in Case 1: the assumption that *W*'s having *M* at *t* causes a_j to have P_j at *t* implies, together with the causal-power actuality principle, that a_j does not already have P_j at *t*. This means, again via the causal-power actuality principle, that a_j cannot, at *t*, exercise the causal/determinative power it has in virtue of having P_j, which in turn implies that the assumed emergence base of *W*'s having *M* at *t* has vanished and *W* cannot have *M* at *t*. Case 1, therefore, collapses.

If you are willing to reject the causal-power actuality principle and live with causal circularity (perhaps even celebrate it in the name of "mutual causal interdependence"), then Case 1 could serve as a model of downward causation for you. Speaking for myself, I think there is a good deal of plausibility in the principle that says that for properties to exercise their causal/determinative powers they must actually be possessed by objects at the time; it cannot be that the objects are in the process of acquiring them at that time. So let's try another model.

Case 2. As before, *W* has emergent property *M* at *t*, and a_j has P_j at *t*. We now consider the causal effect of *W*'s having *M* at *t* on a_j at a *later time* $t + \Delta t$. Suppose, then, that *W*'s having *M* at *t* causes a_j to have *Q* at $t + \Delta t$.

This, therefore, is a case of *diachronic* reflexive downward causation. It is still reflexive in that a whole causes one of its micro-constituents to change in a certain way. Notice, however, that the

mysteriousness of causal reflexivity seems to have vanished. The reason is obvious: the time delay between the putative cause and effect removes the potential circularity, and the causal-power actuality principle does not apply. *W*'s having *M* at *t* causes a_j to have *Q* at $t + \Delta t$. But a_j's having *Q* at $t + \Delta t$ is not part of the basal conditions out of which *M* emerges in *W* at *t*; so there can be no problem of circular reciprocal causation/determination. This becomes particularly clear if we consider the four-dimensional (or "time slice") view of persisting things. On this view, *W*'s having *M* at *t* turns out to be *W* at *t* having *M* – that is, the time slice of *W* at *t* having *M*. Let us use "[*x*, *t*]" to denote the time slice of *x* at *t* (if *t* is an instant, [*x*, *t*] is a temporal cross section). Diachronic downward causation, then, comes to this: [*W*, *t*] having *M* causes $[a_j, t + \Delta t]$ to have *Q*, where, of course, $t < t + \Delta t$. The point to notice is that $[a_j, t + \Delta t]$ is *not* a constituent of [*W*, *t*], and this gets rid of the hint of reflexivity present in Case 2.

Examples falling under Case 2 are everywhere. I fall from the ladder and break my arm. I walk to the kitchen for a drink of water and ten seconds later, all my limbs and organs have been displaced from my study to the kitchen. Sperry's bird flies into the blue yonder, and all of the bird's cells and molecules, too, have gone yonder. It doesn't seem to me that these cases present us with any special mysteries rooted in self-reflexivity, or that they show emergent causation to be something special and unique. For consider Sperry's bird: for simplicity, think of the bird's five constituent parts, its head, torso, two wings, and the tail. For the bird to move from point p_1 to point p_2 *is* for its five parts (together, undetached) to move from p_1 to p_2. The whole bird is at p_1 at t_1 and moving in a certain direction, and this causes, let us suppose, its tail to be at p_2 at t_2. There is nothing mysterious or incoherent about this. The case – the bird's being at p_1 at t_1 and moving in a certain way – includes its tail's being at p_1 at t_1 and moving in a certain way. But that's all right: we expect an object's state at a given time to be an important causal factor for its state a short time later. And it is clear that Sperry's other examples, such as the water eddy and the rolling wheel, can be similarly accommodated.

We must conclude then that of the two types of reflexive downward causation, the diachronic variety poses no special problems but

perhaps for that reason rather unremarkable as a type of causation, but that the synchronic kind *is* problematic and it is doubtful that it can be given a coherent sense. This may be due to its violation of what I called the causal-power actuality principle, but apart from any recondite metaphysical principle that might be involved, one cannot escape the uneasy feeling that there is something circular and incoherent about this variety of downward causation.

VIII

Emergentists like C. Lloyd Morgan will likely point out that the Sperry-style cases do not really involve downward causation by emergent properties, since the motion of the bird as a whole is the same kind of event as the motion of its constituent parts. The properties implicated in causal relations in these cases are one and the same, namely motion, and this shows that these cases simply are not cases of emergent causation, whether downward or upward. (The same will be said about the example of the falling celadon vase.) It would seem, then, that contrary to what Sperry seems to suggest, emergent downward causation should not simply be identified with causation from properties of the whole to properties of its own parts, that is, reflexive downward causation.

One reason that downward causation is thought interesting and important is that mental-to-physical causation is commonly supposed to be a special case of it, the mental occupying a higher emergent level relative to the physical level. So let us turn to mind-body causation. Here again we may consider two varieties, synchronic reflexive downward causation and its diachronic counterpart. Can my experience of pain at a given time causally influence its basal neural process (C-fiber excitation, say) at the very same time? Here we encounter exactly the same difficulties that we saw in Sperry's examples of the water eddy and the like (taken as cases of synchronic downward causation), and I do not believe that classical emergentists, like Alexander, Morgan, and C.D. Broad, would necessarily have insisted on it. Nor do I see why Sperry himself, as an emergentist, should need it; it isn't at all clear that Sperry's overall position on the mind-body relation requires a commitment to this dubious variety of emergent causation.

This leaves diachronic downward causation as the only player on the scene – up to this point, at any rate. One might say that this is all that the emergentists need – the diachronic causal influence of emergent phenomena on lower-level phenomena. But the problem is that even this apparently unproblematic variety of downward causation is beset with difficulties. On my view, the difficulties boil down to a single argument to be sketched below. The critical question that motivates the argument is this: If an emergent, *M*, emerges from basal condition *P*, why can't *P* displace *M* as a cause of any putative effect of *M*? Why can't *P* do all the work in explaining why any alleged effect of *M* occurred?[35] As you may recall, I earlier argued that any upward causation or same-level causation of effect M^* by cause *M* presupposes *M*'s causation of M^*'s lower-level base, P^* (it is supposed that M^* is a higher-level property with a lower-level base; M^* may or may not be an emergent property). But if this is a case of downward emergent causation, *M* is a higher-level property, and as such it must have an emergent base, *P*. Now we are faced with *P*'s threat to preempt *M*'s status as a cause of P^* (and hence of M^*). For if causation is understood as nomological (law-based) sufficiency, *P*, as *M*'s emergence base, is nomologically sufficient for it, and *M*, as P^*'s cause, is nomologically sufficient for P^*. Hence, *P* is nomologically sufficient for P^* and hence qualifies as its cause. The same conclusion follows if causation is understood in terms of counterfactuals – roughly, as a condition without which the effect would not have occurred. Moreover, it is not possible to view the situation as involving a *causal chain* from *P* to P^* with *M* as an intermediate causal link. The reason is that the emergence relation from *P* to *M* cannot properly be viewed as causal.[36] This appears to make the emergent property *M* otiose and dispensable as a cause of P^*; it seems that we can explain the occurrence of P^* simply in terms of *P*, without invoking *M* at all. If *M* is to be retained as a cause of P^*, or of M^*, a positive argument has to be provided, and we have yet to see one. In my opinion, this simple argument has not so far been overcome by an effective counter-argument.

If higher-level property *M* can be reduced to its lower-level base, *M*'s causal status can be restored. As may be recalled from our earlier discussion, however, if *M* is emergent, this is precisely what cannot be done: emergent properties, by definition, are not reducible

to their lower-level bases. The conclusion, therefore, isn't encouraging to emergentists: If emergent properties exist, they are causally, and hence explanatorily, inert and therefore largely useless for the purpose of causal/explanatory theories.

If these considerations are correct, higher-level properties can serve as causes in downward causal relations only if they are reducible to lower-level properties.[37] The paradox is that if they are so reducible, they are not really "higher-level" any longer. If they are reducible to properties at level *L*, they, too, must belong to *L*. Does this make the idea of downward causation useless? Not necessarily. For example, we may try to salvage downward causation by giving it a *conceptual* interpretation. That is, we interpret the hierarchical levels as levels of concepts and descriptions, or levels within our representational apparatus, rather than levels of properties and phenomena in the world. We can then speak of downward causation when a cause is described in terms of higher-level concepts, or in a higher-level language, higher in relation to the concepts in which its effect is represented. On this approach, then, the same cause may be representable in lower-level concepts and languages as well, and a single causal relation would be describable in different languages. The conceptual approach may not save real downward causation, and it brings with it a host of new questions; however, it may be a good enough way of saving *downward causal explanation*, and perhaps that is all we need or should care about.[38]

NOTES

[1] For helpful historical surveys of emergentism see Brian McLaughlin, "The Rise and Fall of British Emergentism", and Achim Stephan, "Emergence – A Systematic View on Its Historical Facets", both in *Emergence or Reduction*?, ed. A. Beckermann, H. Flohr, and J. Kim (Berlin: De Gruyter, 1993).

[2] In *A System of Logic* (1843), Bk. III, ch. vi.

[3] It appears that Galen (AD 129 – *c*. 200) had a clear statement of the distinction between emergent and nonemergent properties of wholes; see *On the Elements according to Hippocrates*, 1.3, 70.15–74.23. I owe this reference to Victor Caston.

[4] See Hempel's "Studies in the Logic of Explanation" (with Paul Oppenheim), reprinted in his *Aspects of Scientific Explanation* (New York: The Free Press, 1965), and Nagel, *The Structure of Science* (New York: Harcourt, Brace & World, 1961), ch. 11. It is interesting to note that another early positivist philosopher of science, Karl Popper, became in the final stages of his career a strong defender

of emergentism; see John C. Eccles and Karl R. Popper, *The Self and Its Brain* (Berlin & New York: Springer International, 1977).

[5] E.g., John Searle, *The Rediscovery of the Mind* (Cambridge: MIT Press, 1992); Francisco Varella, Evan Thompson, and Elearnor Rosch, *The Embodied Mind* (Cambridge: MIT Press, 1993).

[6] In particular, "The Myth of Nonreductive Materialism", reprinted in *Supervenience and Mind* (Cambridge: Cambridge University Press, 1993).

[7] Some will complain that this picture is inextricably wedded to the now defunct prequantum classical particle physics; that may be, but it is the picture the British emergentists worked with. Moreover, it is an open question, I believe, whether anything of substance would change if the issues were set in a quantum-mechanical framework.

[8] Obviously extrinsic/relational/historical properties (e.g., being 50 miles to the south of Boston) must be excluded, and the statement is to be understood to apply only to the intrinsic properties of systems. There is also a tacit assumption that the intrinsic properties of a system determine its causal powers.

[9] C. Llyod Morgan, *Emergent Evolution* (London: Williams and Norgate, 1923), pp. 2–3.

[10] What I believe to be more appropriate here is William C. Wimsatt's notion of aggregativity defined in terms of certain invariance conditions; see his "Emergence as Non-Aggregativity and the Biases of Reductionisms" (forthcoming in *Natural Contradictions: Perspectives on Ecology and Change*, ed. P. J. Taylor and Jrjo Haila). However, I bypass these considerations here in favor of a simpler and more straightforwrad notion of predictability. See also Paul Humphreys' interesting paper, "How Properties Emerge" (forthcoming in *Philosophy of Science*). I cannot discuss here Humphrey's interesting proposals, but I believe everything of any significance I say here is consistent with his views.

[11] Cf. Morgan: "Lewes says that the nature of emergent characters can only be learnt by experience of their occurrence; hence they are unpredictable before the event", *Emergent Evolution*, p. 5.

[12] See, e.g., what Michael Tye calls "perspectival subjectivity", in his *Ten Problems of Consciousness* (Cambridge: MIT Press, 1995). And of course the situation here reminds one of Frank Jackson's much discussed case of the blind superneurophysiologist Mary.

[13] As noted by McLaughlin in "The Rise and Fall of British Emergentism".

[14] For example, David Chalmers, *The Conscious Mind* (New York: Oxford University Press, 1996), p. 43.

[15] The fundamental ideas for this view of reduction are present in David Armstrong's *A Materialist Theory of Mind* (New York: Humanities Press, 1964), and David Lewis's "An Argument for the identity Theory", *Journal of Philosophy* 67 (1970): 203–211. However, neither Armstrong nor Lewis, to my knowledge, explicitly associate these ideas directly with models of reduction. The idea of functional analysis of mental terms or properties is of course the heart of the functionalist approach to mentality; it is interesting, therefore, to note that most functionalists have regarded their approach as essentially antireductionist. For

similar views on reduction see Robert Van Gulick, "Nonreductive Materialism and the Nature of Intertheoretic Constraint", in *Emergence or Reduction?*, ed. A Beckermann, H. Flohr, and J. Kim; Joseph Levine, "On Leaving Out What It Is Like", in *Consciousness*, ed. Martin Davies and Glyn W. Humphreys (Oxford: Blackwell, 1993). See also Chalmers' discussion of "reductive explanation", in *The Conscious Mind*, ch. 2. I discuss these issues in greater detail in *Mind in a Physical World* (forthcoming).

[16] For brevity we will often speak of a property causing another property – what is meant of course is that an instantiation of a property causes another property to be instantiated.

[17] See Lawrence Sklar, *Physics and Chance* (Cambridge: Cambridge University Press, 1993).

[18] See Ernest Nagel, *The Structure of Science* (New York: Harcourt, Brace, and World, 1961).

[19] In "Reduction with Autonomy" (forthcoming in *Philosophical Perspectives*, 1997) Lousie Antony and Joseph Levine advance interesting arguments against the disjunctive approach.

[20] This point is valid whether or not *E* has single or multiple realizers in the actual world. A property may have a single realizer here but multiple realizers in other worlds, and vice versa.

[21] For more details on this approach see my "The Mind-Body Problem: Taking Stock After 40 Years", forthcoming in *Philosophical Perspectives*, 1997, and *Mind in a Physical World* (forthcoming).

[22] For more details see my "Multiple Realization and the Metaphysics of Reduction", reprinted in *Supervenience and Mind*.

[23] This point is argued by David Chalmers; see his *The Conscious Mind*, p. 129.

[24] More details and an overview of the philosophical terrain involved, see Chalmers, *The Conscious Mind*, ch. 3. Two early papers arguing this point are Joseph Levine, "Materialism and Qualia: the Explanatory Gap", *Pacific Philosophical Quarterly* 64 (1983): 354–61, and Frank Jackson, "Epiphenomenal Qualia", *Philosophical Quarterly* 32 (1982): 127–36.

[25] See, e.g., Paul Oppenheim and Hilary Putnam, "Unity of Science as a Working Hypothesis", in *Minnesota Studies in Philosophy of Science*, vol. 2, ed. Hervert Feigl, Michael Scriven, and Grover Maxwell (Minneapolis: University of Minnesota Press, 1958). As the title of the paper suggests, Oppenheim and Putnam advocate a strong physical reductionism, a doctrine that is diametrically opposed to emergentism.

[26] For an informative discussion of the issues in this area see William C. Wimsatt, "Reductionism, Levels of Organization, and the Mind-Body Problem", in *Consciousness and the Brain*, ed. Gordon G. Globus, Grover Maxwell, and Irwin Savodnik (New York & London: Plenum Press, 1976), and "The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets", *Canadian Journal of Philosophy*, Supplementary Volume 20 (1994): 207–274.

[27] *Space, Time, and Deity*, vol. 2 (London: Macmillan, 1927), p. 8.

[28] As I have argued elsewhere, this holds for certain positions other than emer-

gentism, e.g., the view that higher properties supervene on lower properties, and the view that higher properties are realized by lower properties.

[29] I first presented this arguement in "'Downward Causation' in Emergentism and Nonreductive Physicalism", in *Emergence or Reduction*?

[30] *Emergent Evolution* (London: Williams & Norgate, 1927), pp. 15–16. Emphasis added.

[31] "Mental Phenomena as Causal Determinants in Brain Function", in *Consciousness and the Brain*, ed. Globus, Maxwell, and Savodnik, p. 165.

[32] "A Modified Concept of Consciousness", *Psychological Review* 76 (1969): 532–536.

[33] Ibid.

[34] This case, therefore, involves the controversial idea of simultaneous causation (where a cause and its effect occur at the same time). However, this is a general metaphysical issue, and in the present context it will be unproductive to focus on this aspect of the situation.

[35] I raised this question earlier in "'Downward Causation' in Emergentism and Nonreductive Physicalims", in *Emergence or Reduction*? The argument can be generalized to the supervenience and realization views of the mind-body relation. For more details see my "The Nonreductivist's Troubles with Mental Causation" (reprinted in *Supervenience and Mind*) and *Mind in a Physical World*. See also Timothy O'Connor, "Emergent Properties", *American Philosophical Quarterly* 31 (1994): 91–104, for an attempt to counter the argument.

[36] C. Lloyd Morgan explicitly denies that emergence is a form of causation, in *Emergent Evolution*, p. 28.

[37] Here I must enter some caveats. As the reader may recall, I earlier said that there is no special problem of downward causation, citing such examples as my celadon creashing on the pavement of the sidewalk. Cases like this are not the cases of downward causation that most emergentists have in mind, for like Sperry's example of the flying bird they don't seem to involve genuine "higher-level" properties. In general, complex systems obviously can bring new causal powers into the world, powers that cannot be identified with causal powers of more basic, simpler systems. Among them are the causal powers of microstructural, or micro-based, properties of a complex system. Note that these properties are not themselves emergent properties; rather, they form the basal conditions from which further properties emerge (for example, that consciousness is not itself a microstructural property of an organism, although it may emerge from one). If all this sounds too complicated, you could regard the argument in the text to be restricted to consciousness and other standard examples of emergent properties. For further discussion, see my *Mind in a Physical World*.

[38] This paper is largely based on the following two papers of mine: "Explanation, Prediction, and Reduction in Emergentism", forthcoming in *Intellectica*, and "Making Sense of Downward Causation", forthcoming in a volume of essays on emergence and downward causation, ed. Peter Boegh Andersen et al.

Brown University