

EMERGENTISM, IRREDUCIBILITY, AND DOWNWARD CAUSATION

Achim STEPHAN
University of Osnabrück

Summary

Several theories of emergence will be distinguished. In particular, these are synchronic, diachronic, and weak versions of emergence. While the weaker theories are compatible with property reductionism, synchronic emergentism and strong versions of diachronic emergentism are not. Synchronic emergentism is of particular interest for the discussion of downward causation. For such a theory, a system's property is taken to be emergent if it is irreducible, *i.e.*, if it is not reductively explainable. Furthermore, we have to distinguish two different types of irreducibility with quite different consequences: If, on the one hand, a system's property is irreducible because of the irreducibility of the system's parts' behavior on which the property supervenes, we seem to have a case of "downward causation". This kind of downward causation does not violate the principle of the causal closure of the physical domain. If, on the other hand, a systemic property is irreducible because it is not exhaustively analyzable in terms of its causal role, downward causation is not implied. Rather, it is dubitable how unanalyzable properties might play any causal role at all. Thus, epiphenomenalism seems to be implied. The failure to keep apart the two kinds of irreducibility has muddled recent debate about the emergence of properties considerably.

1. *Introduction*

Some philosophers—for example, J. Kim and T. O'Connor—associate emergentism quite closely with the idea of downward causation. While O'Connor takes "downward causation" to be one of the characteristic features of emergent properties, Kim emphasizes that it is an entirely natural and plausible claim that emergents have their own distinctive causal powers if you believe in emergent properties at all (cf. Kim 1999, 19; 1996, 229; and O'Connor 1994, 98). J. Searle, too, develops a

“more adventurous” notion of emergence, namely “emergent2”, which also entails downward causation, although he tells us that he cannot think of anything that *is* emergent2 (cf. 1992, 112).

In opposition to these views, I prefer to soften the connection between emergence and downward causation by stressing that the notion of *irreducibility* is at the core of all stronger versions of emergence. When discussing irreducible properties we may, *then*, ask whether or not they have distinctive (downward directed) causal powers, and if so, what consequences this might have.

At the outset, however, I examine several notions of emergence with different strength, for it is still highly controversial by what criteria emergent properties should be distinguished from non-emergent properties. I will distinguish between synchronic, diachronic, and weak theories of emergence, and it will become clear that the weaker versions of emergence are compatible with property reductionism, while synchronic emergentism and strong versions of diachronic emergentism are not.

2. *Weak, Diachronic, and Synchronic Emergentism*

The different varieties of emergentism are covered more or less by three theories deserving particular interest: *synchronic* emergentism, *diachronic* emergentism, and a *weak* version of emergentism. For synchronic emergentism the relationship between a system’s property and the system’s microstructure, *i.e.*, the arrangement and the properties of the system’s parts, is in the center of interest. For such a theory, a property of a system is taken to be emergent only if it is *irreducible* or, what I take to be the same, if it is *not* reductively explainable. In contrast, diachronic emergentism is mainly interested in the *predictability* of novel properties. For such a theory, those properties are emergent that could not have been predicted in principle before their first instantiation. These two stronger versions of emergentism are not independent of each other, since irreducible properties are *eo ipso* unpredictable in principle before their first appearance. Hence, synchronically emergent properties are diachronically emergent, too, but not *vice versa*.

Both stronger versions of emergentism are based on the same “weak theory”, which at the present pervades emergentist theorizing in various

approaches to cognitive science, *e.g.*, connectionism, artificial life, and theories of self-organization. Its three basic features—the thesis of *physical monism*, the thesis of *systemic* (or collective) *properties*, and the thesis of *synchronic determinism*—are compatible with reductionist approaches without further ado. The stronger versions of emergentism can be obtained from *weak emergentism* by adding further theses.

2.1 *Weak Emergentism*

The first feature of contemporary theories of emergence, the thesis of *physical monism*, is a thesis about the nature of systems that have emergent properties (or structures). The thesis says that the bearers of emergent properties are made up of material parts only. It denies that there are any supernatural components responsible for a system's having emergent properties. Thus, all substance-dualistic positions are rejected; for they base properties such as being alive or having a mental state on supernatural bearers such as an *entelechy* or a *res cogitans*, respectively. Instead, it is claimed that living beings and cognitive systems consist of the same basic entities that make up inanimate nature: it is nothing but specific sequences of highly complex physico-chemical states that realize their vital behavior or their mental states.

- (i) *Physical monism*. Entities existing or coming into existence consist solely of material parts. Hence, properties, dispositions, behavior, or structures classified as emergent are instantiated by systems consisting exclusively of physical entities.

While the first thesis puts the discussion of emergent properties and structures within the framework of a physicalistic naturalism, the second thesis delimits the type of properties that are possible candidates for emergents. It is based on the assumption that *general* properties of complex systems fall into two different classes,¹ namely properties which some of a systems' parts also have, and properties that none of a system's parts have. Examples of the first class are properties such as being extended and having a velocity. Examples of the properties in the second class are walking, reproducing, breathing, or having a sensation

1. General properties are properties of a general type, such as having a weight; they are not specific properties, such as having a weight of 85,4 kilogram.

of pain. These properties are called *systemic* (or collective) properties.

- (ii) *Systemic properties*. Emergent properties are systemic properties. A property is a systemic property if and only if a system possesses it, but no part of the system possesses it.

It should be uncontroversial that systems with systemic properties exist. Those who would deny their existence would have to claim that *all* of a system's properties are instantiated already by some of the system's parts. Countless examples refute such a claim.

While the first thesis restricts the type of parts out of which systems having emergent properties may be built up, and while the second thesis characterizes in more detail the type of properties that might be emergent, the third thesis specifies the type of relationship that holds between a system's micro-structure and its emergent properties as a relationship of *synchronic determination*:

- (iii) *Synchronic determination*. A system's intrinsic properties and dispositions depend nomologically on its micro-structure, that is to say, on its parts' properties and their arrangement. There can be no difference in the (intrinsic) systemic properties without there being some differences in the properties of the system's parts or in their arrangement.

In recent debate, the thesis of *synchronic determination* is sometimes stated in a weaker version, namely as the thesis of *mereological supervenience*, which claims that a system's intrinsic properties supervene on its parts' properties and their arrangement. Then, too, there is no difference in the systemic properties without differences in the parts' properties or their arrangement. The thesis of mereological supervenience, however, is weaker than the thesis of synchronic determination, since it does not imply, strictly speaking, the *dependence* of the system's properties on its micro-structure, it only entails their *covariance*.²

2. There is an ambiguity in the concept of supervenience. On the one hand, the notion of supervenience is associated with three different features: covariance, dependency, and non-reducibility. On the other hand, the various definitions of supervenience give us, *strictly speaking*, only weak, global, or strong *covariance*. Since covariance does not entail dependency, supervenience, *strictly speaking*, does not entail dependency either. For the last

Anyone who denies the thesis of synchronic determination either has to admit properties that are not bound to the properties and the arrangement of its bearer's parts, or she has to suppose that some other factors, *e.g.*, non-natural entities or forces, are responsible for the different dispositions of systems that are identical in their microstructure. She would have to admit, for example, that there may exist objects that have the same parts in the same arrangement as diamonds, but which lack the diamond's hardness, *i.e.*, they may have hardness 2 instead of hardness 10 on the Mohs-scale. This seems to be implausible. Equally unthinkable is that there may exist two micro-identical organisms, one of which is viable and the other not. In the case of mental phenomena, opinions may be more controversial; but one thing seems to be clear: anyone who believes, *e.g.*, that two creatures identical in micro-structure could be such that one is colorblind while the other is not, does not hold a physicalist position. Similar considerations hold for propositional attitudes only as long as one does not subscribe to externalism, that is to say, if one does not claim that, *e.g.*, the content of a belief depends essentially on the nature of the referents of the believer's thoughts and concepts.

Weak emergentism as sketched so far comprises the minimal conditions for emergent properties. It is the common base of all stronger theories of emergence. Moreover—and this is a reason for distinguishing it as a theory in its own right—it is held not only by some philosophers (*e.g.* M. Bunge and G. Vollmer), but also by cognitive scientists (*e.g.* J. Hopfield, E. Rosch, F. Varela, and D. Rumelhart) in exactly its weak form. The three features of weak emergentism—the thesis of *physical monism*, the thesis of *systemic properties*, and the thesis of *synchronic determination*—are compatible with contemporary reductionist approaches. Some champions of *weak* emergentism credit the compatibility of “emergence” and “reducibility” as one of its merits compared to stronger versions of emergentism.

point, cf. Kim (1990; 1993, 142–149), where he says: “it is best to separate the covariation element from the dependency element in the relation of supervenience. ... property covariation alone, even in the form of ‘strong asymmetric covariance’, does not itself give us dependency; in that sense, dependency is an additional component of supervenience” (*ib.*, 148); cf. also Grimes (1988). However, as I have just noted, dependency is not captured by the various proposals to define supervenience.

2.2 *Diachronic Emergentism*

All diachronic theories of emergence have at bottom a thesis about the occurrence of genuine *novelties*—properties or structures—in evolution. This thesis excludes at the same time all preformationist positions.

- (iv) *Novelty*. In the course of evolution exemplifications of “genuine novelties” occur again and again. Already existing entities form new constellations that produce new structures which may constitute new entities with new properties and behaviors.

However, bare addition of the thesis of novelty does not turn a weak theory of emergence into a strong one, since reductive physicalism remains compatible with such a variant of emergentism. Only the addition of the thesis of *unpredictability*, in principle, of novel properties will lead to stronger forms of *diachronic* emergentism.

A short consideration shows that systemic properties can be unpredictable in principle for two completely different reasons: first, they can be unpredictable because the micro-structure of the system, which will exemplify the property for the first time in evolution, is unpredictable. For, if the micro-structure of a newly emerging system is unpredictable, so are the properties which nomologically depend on it. Second, a property can be unpredictable although a novel system’s micro-structure is predictable. That is the case if the property *itself* is irreducible. For, if systemic properties are irreducible, then they are unpredictable before their first appearance. Obviously, this does not preclude that further occurrences of such properties might be predicted adequately.

Since in the second case the criteria for being unpredictable are identical with those for being irreducible, I postpone the analysis of those properties until the next section and confine myself here to the first case, *unpredictability of structure*. This version of unpredictability gains considerable significance in the teeth of strong interest in dynamical systems and chaotic processes.

The structure of a newly formed system could be unpredictable for two reasons yet again. First, belief in an indeterministic universe would imply that there will be novel, unpredictable structures. However, from an emergentist perspective it is of no interest if a new structure’s appearance is unpredictable only because its coming into being is not determined. Furthermore, most emergentists have claimed, anyway, that

the development of new structures is governed by deterministic laws. However, and this is the second case, even deterministic formings of new structures can be *unpredictable in principle* if they are governed by laws which are attributed to deterministic chaos. Let me explain this.

An essential outcome of the theory of chaos is that there exist—even very simple—mathematical functions, whose own “behavior” cannot be predicted. Only the rise of “experimental mathematics” on highly efficient computers has revealed, for example, the properties of various logistic functions. Their intra-mathematical unpredictability has to do with an aperiodic behavior of these functions, by which marginally different initial values of some variable can lead to radically distinct trajectories of the functions.

A standard example is the logistic function $f(x) = \mu x(1-x)$ for $0 \leq x \leq 1$. For a parameter μ with $0 \leq \mu \leq 4$ the logistic function maps the interval $[0,1]$ onto itself. Of particular interest is how parameter μ influences the long-term behavior of the function when iterated repeatedly. For $0 \leq \mu \leq 1$ the situation is obvious. All initial values of the variable x let the function $f(x)$ approximate the value 0 after sufficiently many iterations; the origin thus is the attractor. For $1 < \mu < 3$ there exists exactly one attractor A of value $A = 1-1/\mu$: the function balances out on a stable value. If μ equals 3, the fixed point of the function is “marginally stable”; convergence is decidedly slow—an indication for fundamental change in the function’s behavior. For larger values the dynamics becomes considerably complex. In the case of $3 < \mu < 1 + \sqrt{6}$ values oscillate between two fixed points. By increasing μ the attractors of period two will become instable, too. We get a cycle of period four (*i.e.*, after four iterations the values of the function approach in each case the four fixed points). At 3.56 the period doubles again and becomes eight, at 3.567 it becomes sixteen, and then we get a quickly rising sequence of periods to 32, 64, 128, etc.—vividly one speaks of cascades. At about 3.58, this sequence comes to an end. The period has doubled itself infinitely many times. Hereafter, predictions do not seem to be possible. Marginally different initial values of x lead to radically different trajectories of the iterated function. Values jump pell-mell, convergence and divergence are not discernible: chaos dominates.

Thus, it looks just as if the most exact science of all has led us back to one of the starting points of emergentism. Whereas—after pioneering successes in chemistry and physics—today we do not count properties and dispositions of chemical compounds any more among emergent

phenomena, examinations of deterministic chaos suggest the existence of systems that might develop structures that are unpredictable in principle and thus might show *structure-emergent* behavior.

Of course, one could argue that a Laplacean calculator could predict correctly even chaotic processes. Whether or not this could actually be the case, however, is not settled yet. It depends mainly on the question of what kind of information we allow such a fabulous creature to have. For example, in Alexander's considerations (cf. 1920, ii, 72 f., 328) Laplace's calculator knows several earlier states of the whole world and, in addition, all natural laws that govern changes in the world. He seems to be able to extrapolate from his knowledge of all events that have occurred in the universe so far even the course of chaotic processes. But on what basis could he do that? Since chaotic processes are aperiodic, one can not determine definitely from those processes that have occurred up to a certain time the exact formula which would describe their further course. Even if the further course of the world is governed by deterministic laws, it does not follow from the earlier events and states alone, by *which* laws it is governed. Entirely different continuations seem to be compatible with the earlier course of the world. Therefore, even a Laplacean calculator could fail in his predictions. If one grants, however, that he knows *all* details of earlier world states—up to infinitely many digits—and if one grants that he knows a priori which processes are governed by which *specific* chaotic laws, then, of course, he would be able to predict the forming of structures that are governed by these laws. I will leave it open whether or not it is plausible to ascribe this kind of knowledge to such a fabulous creature. However, we can preclude that foretellers of our mental capacities have these abilities, and suppose that where chaos exists, structures exist that are unpredictable in principle, and that is to say, that there will be *structure emergence* in our sense.

- (v) *Structure-unpredictability*. The rise of novel structures is unpredictable in principle, if their formation is governed by laws of deterministic chaos. Likewise, all novel properties that are instantiated by those structures are unpredictable in principle.

Summing up, it may be said that a *systemic property* is *unpredictable* in principle before its first appearance, if the structure which instantiates it is unpredictable in principle before its first formation. Although

unpredictability of structure always implies unpredictability of properties instantiated by the structure, it does not thereby imply the irreducibility of the properties instantiated by the structure. As far as that goes, unpredictability in principle of systemic properties is entirely compatible with their being reducible to the micro-structure of the system that instantiates them. However, systemic properties are also unpredictable in principle, if they are irreducible. It is this feature that is at the core of synchronic emergentism, the doctrine I am going to examine next.

2.3 Synchronic Emergentism

The notion of *irreducibility* lies at the heart of all strong versions of emergentism. Although we might think of introducing it in contrast to the concept of *reductive explanation* as defined by J. Levine and J. Kim, it is intriguing first to follow C. D. Broad's attempt to explicate a strong (synchronic) notion of emergence. In his book *The Mind and its Place in Nature* we find a passage that nowadays may count as downright classical; it reads:

Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents *A*, *B*, and *C* in a relation *R* to each other; that all wholes composed of constituents of the same kind as *A*, *B*, and *C* in relations of the same kind as *R* have certain characteristic properties; that *A*, *B*, and *C* are capable of occurring in other kinds of complex where the relation is not the same kind as *R*; and that the characteristic properties of the whole *R(A,B,C)* cannot, even in theory, be *deduced* from the most complete knowledge of the properties of *A*, *B*, and *C* in isolation or in other wholes which are not of the form *R(A,B,C)* (1925, 61).

According to Broad's definition, a systemic property, which is supposed to be nomologically dependent on its system's micro-structure (by the thesis of synchronic determination), is called *irreducible* and therefore *emergent*, if and only if it cannot be deduced from the arrangement of its system's parts and the properties they have "in isolation" or in other (more simple) systems.³

3. Properties that might be attributed to a part "in isolation" are, according to Broad, properties that depend essentially on the micro-structure of the part, while external factors, such as the part's arrangement and its neighboring parts, can be seen as almost irrelevant for the part's having these properties (cf. 1919, 112 f.).

Although, *prima facie*, it looks as if Broad's proposal gives us a clear and distinct explication of what it is for a systemic property to be irreducible (or non-deducible), a further look reveals that two different kinds of irreducibility are concealed that have quite different consequences. As we will see, one type of irreducibility seems to imply *downward causation* while the other seems to imply *epiphenomenalism*. The failure to keep apart the two kinds of irreducibility has muddled the recent debate about the emergence of properties.

To make things clearer, I shall first discuss when a systemic property is *reducible*. For this to be the case, two conditions must be fulfilled: The first is that from the behavior of the system's parts alone it must follow that the system has some property *P*. The second condition demands that the behavior the system's parts exhibit when they are part of the system follows from the behavior they show in isolation or in systems simpler than the system in question. If both conditions are fulfilled, the behavior of the system's parts in *other* contexts reveals what systemic properties the actual system has. That is to say, those properties are reducible. Since both conditions are independent of each other, two totally different possibilities for the occurrence of *irreducible* systemic properties will result: (a) a systemic property *P* of a system *S* is *irreducible*, if it does *not* follow, even in principle, from the behavior of the system's parts that *S* has property *P*; and (b) a systemic property *P* of a system *S* is *irreducible*, if it does *not* follow, even in principle, from the behavior of the system's parts in constellations simpler than *S* how they will behave in *S*.

Thus, a necessary requirement for a systemic property to be reducible is that its being instantiated has to follow from the behavior of its bearer's parts. In other words: From the behavior of the system's parts it must follow that the system has all characteristic features that are essential for having the systemic property. Broad, for example, takes this condition, which is enclosed in the first criterion for reducibility, to be always fulfilled in the case of the characteristic properties of chemical compounds and viable organisms. Their properties might be irreducible only by violating the second criterion, which means that from the behavior of the system's parts in other (simpler) systems it might not follow how they will behave in the actual system.

In contrast, he claims that the irreducibility of secondary qualities and phenomenal qualities results already from a violation of the first condition, since they are neither adequately characterizable by the

macroscopic nor by the microscopic behavior of the systems' parts, not even in principle. For, when we say that a certain object is red or a chemical substance has the smell of liquid ammonia, we do not mean that the corresponding system's parts *behave* or *move* in a certain way. No progress in the sciences could change this state of affairs in any way.⁴ Broad has illustrated the fundamental distinction between (behaviorally) *analyzable* and *unanalyzable* properties by pointing to characteristic properties of organisms and secondary qualities, respectively.

If secondary and phenomenal qualities are not analyzable,⁵ even in principle, then there is no prospect that an increase of scientific knowledge will close the gap between physical processes and secondary qualities or between physiological processes and phenomenal states of consciousness (qualia), respectively.

We can now specify more exactly the feature of irreducibility which is central for *synchronic* emergence. Its first variant is based on the behavioral unanalyzability of systemic properties. It reads:

- (vi.a) *Unanalyzability*. Systemic properties which are not behaviorally analyzable—be it micro- or macroscopically—are (necessarily) irreducible.

However, even if secondary and phenomenal qualities belong to the class of unanalyzable properties, it does not follow that the specific behavior of the system's parts upon which those qualities supervene is itself not deducible from the behavior those parts show isolated or in other (simpler) systems. The irreducibility which results from a violation of the first criterion of reducibility does not imply, by itself, a violation of the second criterion of reducibility.

On the other hand, however, even analyzable systemic properties can be irreducible and therefore emergent. This is the case if the second criterion of reducibility is violated, *i.e.*, when the behavior of

4. However, whether reference to linguistic usage might answer questions concerning reducibility in a definite way is controversial. Particularly, P. Churchland has opposed arguments of the Broadian style (see 1988, 29 ff.).

5. Properties that are called "unanalyzable" for simplicity here might be analyzable in other ways than by behavioral features. A certain smell, for example, might be analyzed as a mixture of the smells of musk and fish-meal. This, however, would not be an analysis based on concepts of motion and behavior.

the system's parts does not follow from their behavior in other (simpler) constellations. Broad thinks that such examples of irreducible behavior might occur in chemical compounds and also in organisms.⁶ His central idea is that the parts of a genuinely novel structure, such as, *e.g.*, an organism in comparison to any inorganic compound, might behave in a way that is not deducible from the part's behavior in other structures. Implicitly, that means that the actual behavior of parts that interact in wholes does not result from their behavior in pairs.⁷ If the behavior of some system's parts is irreducible in this respect, then all properties that depend nomologically on the behavior of the system's parts (for example, reproduction) are irreducible too.

Thus, we can specify more precisely the second variant of a systemic property's irreducibility. It is based on the non-deducibility of the behavior of the system's parts:

- (vi.b) *Irreducibility of the components' behavior.* The specific behavior of a system's components within the system is irreducible if it does not follow from the components' behavior taken in isolation or in other (simpler) constellations.

A violation of the second criterion of reducibility, which is manifested in the irreducibility of the component's behavior, does not imply, however, a violation of the first criterion of reducibility. Systemic properties that cannot be reduced because the behavior of the system's parts is irreducible might nevertheless be behaviorally analyzable. Hence, the two necessary conditions of reducibility as well as those irreducibilities that are based on the violation of these conditions are independent of each other. Summarizing, we obtain from (vi.a) and (vi.b) the following modified version of systemic property irreducibility:

- (vi) *Irreducibility.* A systemic property is irreducible if (a) it is neither micro- nor macroscopically behaviorally analyzable,

6. Broad has also examined "in abstracto" under what conditions the behavior of a system's components can be irreducible (cf. 1919, 113 f.). Recently, W. Bechtel has held a similar position: "although studying the properties of amino acids in isolation may reveal their primary bonding properties, it may not reveal to us those binding properties that give rise to secondary and tertiary structure when the amino acids are incorporated into protein molecules" (1988, 95).

7. Scenarios of this kind are discussed already by J. S. Mill and G. T. Fechner (cf. Stephan 1999, sections 6.1. and 7.3.).

or if (b) the specific behavior of the system's components, on which the systemic property supervenes, does not follow from the component's behavior taken in isolation or in other (simpler) constellations.

Thus, we have to distinguish two completely different types of irreducibility of systemic properties; and the consequences that go along with them seem to be equally different. If a systemic property is irreducible because of the irreducibility of the parts' behavior on which the property supervenes, we seem to have a case of "downward causation". For, if the components' behavior is not reducible to their arrangement and the behavior they show in simpler systems or in isolation there seems to exist some "downward" causal influence exerted by the system itself (or by its structure) on the behavior of the system's parts. To be sure, if there existed such instances of "downward causation", this would not amount to a violation of some widely held assumptions such as, for example, the principle of the causal closure of the physical domain.⁸ Within the physical domain, we would just have to accept additional types of causal influences besides the already known basal types of mutual interactions.

In contrast, the occurrence of unanalyzable properties does not imply any kind of downward causation. Systems that have unanalyzable properties that depend nomologically on their bearer's micro-structures need not be constituted in a way that amounts to the irreducibility of their components' behavior. Nor is it implied that the system's structure has a downward causal influence on the system's parts. All the more, there is no reason to assume that unanalyzable properties themselves exert a causal influence on the system's parts. Rather, it is dubitable how unanalyzable properties might play any causal role at all. Since they are not behaviorally analyzable—that is to say, they neither seem to correspond to any "mechanism" nor do they seem to result from any "mechanism"—, it is hard to see how they themselves could be causally efficacious. And, if one cannot conceive of *how* unanalyzable properties might play a causal role, it is hard to conceive of them other than being epiphenomena.⁹

8. C. Gillett, however, takes another line of reasoning while discussing configurational forces; cf. his (2003, 95–121).

9. In Stephan (1997) I tried to find a way how we nevertheless can attribute causal powers to unanalyzable properties.

It is instructive to see how Kim's and Levine's notions of *reductive explanation* (which may also serve to define the notion of synchronic emergence) relate to the foregoing analysis. Both authors note that in a first step we have to work the concept of the property to be reduced "into shape" for reduction. Kim also calls this the "priming procedure" in which we must construe, or reconstrue, the property to be reduced relationally or extrinsically. He considers a domain **B** of properties (or phenomena, facts, etc.) serving as the reduction base, and a property *E* that is going to be reduced (cf. 1999, 10–12):¹⁰

Step1 *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**.

Now, this condition which we may call the *functionalization condition* is nothing but a new guise of Broad's analyzability condition. If the concept of a property cannot be worked "into shape" for reduction, that is, if the property cannot be "functionalized" then reduction necessarily fails. The property will turn out to be irreducible, hence synchronically emergent. But notice, as we just have seen, if irreducible properties exert "downward causation" they do not do so *qua* being unanalyzable (or "unfunctionalizable").

As a second step both Kim and Levine mention the empirical task of finding the realizers of a property that could be worked into shape for reduction: "Stage 2 involves the empirical work of discovering just what those underlying mechanisms are" (Levine 1993, 132). This task is not explicitly mentioned by Broad, since he discusses situations where we already have nomological correlations between a system's microstructure and some systemic property, but are not able to deduce the property from our reduction base, *i.e.*, from the behavior of the system's parts in simpler systems or in isolation.

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*

10. Levine has put it as follows: "Stage 1 involves the ... *a priori* process of working the concept of the property to be reduced 'into shape' for reduction by identifying the causal role for which we are seeking the underlying mechanisms" (1993, 132).

is instantiated on this occasion in this system (similarly, for classes of systems belonging to the same species or structure types).

Kim, in opposition to Levine, postulates a third step in which he refers to the development of theories that can explain how the realizers of *E*, the property to be reduced, perform the particular functional role that is characteristic for *E*.

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.

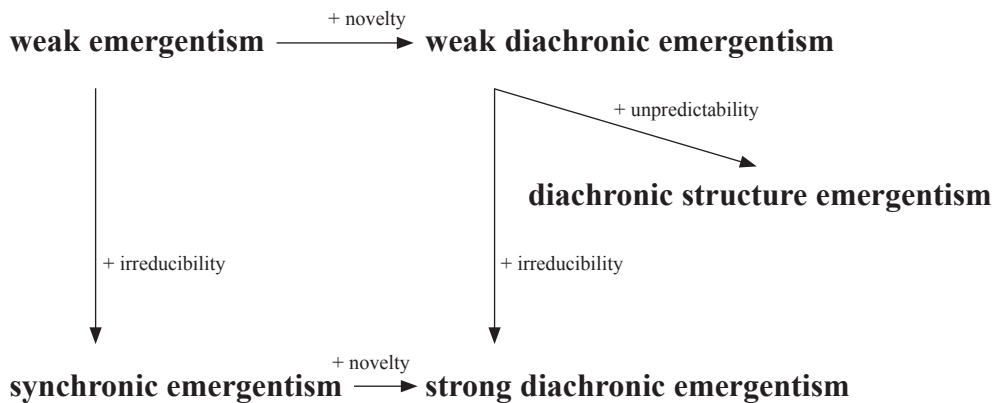
Although *prima facie* Kim's third step seems to correspond to Broad's second criterion, a closer look reveals important differences. Kim only considers theories that connect the reduction base of a systemic property (usually the system's microstructure) with the property to be reduced. Thus, nothing can be said about any "downward" causal influence by the system or its structure. The detailed analysis of Broad's approach, however, has allowed us to identify cases that imply downward causation. These are cases of irreducibility, where no theory is capable to explain the behavior of a system's parts within this very system by reference to their behavior in simpler systems.

A case in point concerning irreducible mental phenomena are qualia. Among others, Levine and Chalmers claim that we cannot reductively explain them, but not for empirical reasons. Apparently, phenomenal properties cannot be individuated by their causal roles, or as Levine says: "What seems to be responsible for the explanatory gap, then, is the fact that our concepts of qualitative character do not represent, at least in terms of their psychological contents, causal roles. [...] Thus, to the extent that there is an element in our concept of qualitative character that is not captured by features of its causal role, to that extent it will escape the explanatory net of a physicalistic reduction" (1993, p. 134). Kim seems to share this appraisal when he states: "To get to the point without fuss, it seems to me that the felt, phenomenal qualities of experiences, or qualia, are intrinsic properties if anything is" (1998, 102). Therefore, he thinks that the functionalization of qualia won't work. It is for the same reasons that Broad already took them to be

unanalyzable. Thus, qualia seem to be the best candidates for irreducible, *i.e.*, synchronically emergent properties—properties, however, that do not exert downward causal powers *qua* being unanalyzable.

3. Synopsis

Finally, I depict and summarize the logical relationships that hold between the different versions of emergentism with the aid of the following figure:



Weak diachronic emergentism results from *weak emergentism* by adding a temporal dimension in the form of the thesis of novelty. Both versions are compatible with reductive physicalism. Weak theories of emergence are used today mainly in cognitive science, particularly for a characterization of systemic properties of connectionist networks, and in theories of self-organization. *Synchronic emergentism* results from weak emergentism by adding the thesis of irreducibility. This version of emergentism is important for the philosophy of mind, particularly in the discussion of nonreductive physicalism and qualia. It is not compatible with reductive physicalism. *Strong diachronic emergentism* only differs from synchronic emergentism because of the temporal dimension in the thesis of novelty. In contrast, *structure emergentism* is entirely independent of synchronic emergentism. It results from weak diachronic emergentism by adding the thesis of structure-unpredictability. Although structure emergentism emphasizes the boundaries of prediction within

physicalistic approaches, it is compatible with reductive physicalism, and so it is weaker than synchronic emergentism. Theories of deterministic chaos in dynamical systems can be acknowledged as a type of structure emergentism. Likewise, its perspective is important for evolutionary research. Most important from a theoretical point of view are *weak emergentism*, *synchronic emergentism*, and *diachronic structure emergentism*.

REFERENCES

- Alexander, S. (1920) *Space, Time, and Deity*. Two Volumes. London: Macmillan.
- Bechtel, W. (1988) *Philosophy of Science. An Overview for Cognitive Science*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Broad, C. D. (1919) Mechanical Explanation and its Alternatives. *Proceedings of the Aristotelian Society* 19, 86–124.
- (1925) *The Mind and its Place in Nature*. London: Kegan Paul, Trench, Trubner & Co.
- Churchland, P. (1988) *Matter and Consciousness*. Revised Edition. Cambridge, Ma.: MIT Press.
- Gillett, Carl (2002) The Varieties of Emergence. Their Purposes, Obligations and Importance. *Grazer Philosophische Studien* 65, 95–121.
- Grimes, T. R. (1988) The Myth of Supervenience. *Pacific Philosophical Quarterly* 69, 152–160.
- Kim, J. (1990) Supervenience as a Philosophical Concept. *Metaphilosophy* 21, 1–27. Reprinted in J. Kim, *Supervenience and Mind*, Cambridge: Cambridge University Press, 1993, 131–160.
- (1996) *Philosophy of Mind*. Boulder: Westview Press.
- (1998) *Mind in a Physical World*. Cambridge: MIT Press.
- (1999) Making Sense of Emergence. *Philosophical Studies* 95, 3–36.
- Levine, J. (1993) On Leaving Out What It's Like. In M. Davies and G. W. Humphreys (eds.) *Consciousness*. Oxford: Blackwell, 121–136.
- O'Connor, T. (1994) Emergent Properties. *American Philosophical Quarterly* 31, 91–104.
- Searle, J. (1992) *The Rediscovery of the Mind*. Cambridge: MIT Press.
- Stephan, A. (1997) Armchair Arguments against Emergentism. *Erkenntnis* 46, 305–314.
- (1999) *Emergenz. Von der Unvorhersagbarkeit zur Selbstorganisation*. Dresden, München: Dresden University Press.