Varieties of Emergence in Artificial and Natural Systems*

Achim Stephan

Institut für Philosophie der Universität Karlsruhe (TH), Postfach 6980, D-76128 Karlsruhe
at present: Hanse-Wissenschaftskolleg, Fischstr. 31, D-27749 Delmenhorst

Z. Naturforsch. **53c,** 639–656 (1998); received April 20, 1998

Emergence, Reduction, Connectionism, Qualia, Synergetics

In different disciplines such as philosophy of mind, dynamical systems theory, and connectionism the term 'emergence' has different jobs to perform. Therefore, various concepts of emergence are developed and examined. While weaker versions are compatible with property reductionism, stronger versions are not. Within philosophy of mind, particularly within the qualia debate there is a need for a strong notion of emergence, while in discussions of emergent properties of connectionist nets or of dynamical systems one can do with weaker notions of emergence.

Introduction

In addition to having various technical uses, the term 'emergence' also has a use in ordinary language. Thus, sometimes people use the expression 'the emergence of x' just to mean that x has appeared or that x has come up. The term 'emergence' is used in this way in book titles such as "The Emergence of Symbols" (Bates, 1979) and "The Emergence of Probability" (Hacking, 1975). Of course, one could speak about 'the emergence of animals with brains' or about 'the emergence of robots' in this sense of 'emergence'. However, I will not focus on this ordinary use of 'emergence' in what follows. Instead, I will focus on the technical uses of 'emergence'.

In most technical uses, 'emergent' denotes a second order property of certain first order properties (or structures), namely, the first order properties that are emergent. However, it is controversial what the criteria are by which emergent properties are to be distinguished from non-emergent properties. Some criteria are very strong, so that few, if any, properties count as emergent. Other criteria are inflationary in that they count many, if not all, system properties as emergent. One of the consequences of this controversy is a great confusion about what is really meant by an 'emergent property', when this term is used in such different disciplines as theories of self-organization, philosophy of mind, dynamical systems theory, or connectionism.

Therefore, the second section of this article is intended to discuss, in a systematic way, several theories and concepts of emergence of different strengths. It will be shown that the weaker versions are compatible with property reductionism. In contrast, stronger versions are incompatible with property reductionism. Also, the important distinction between synchronic and diachronic theories of emergence is developed within this section.

In the third section, I examine the different concepts of emergence, distinguished in section 2, as they apply to several natural and artificial systems. It will become evident that the concepts perform very different jobs, and so, one needs to be clear about which concept of emergence one wants to employ. For example, within philosophy of mind, particularly within the qualia debate there is a need for a strong notion of emergence, while in discussions of emergent properties of connectionist nets or of dynamical systems one can be content with weaker notions of emergence.

'Emergent' is not only attributed to properties in a strict sense, but also to dispositions, behavior, and structures. To simplify my presentation, I will use the concept of a *property* in a wide sense to apply to dispositions (e.g., being breakable) and behavior: Of a system which behaves such and

---

* This communication is a contribution to the workshop on "Natural Organisms, Artificial Organisms, and Their Brains" at the Zentrum für interdisziplinäre Forschung (ZiF) in Bielefeld (Germany) on March 8–12, 1998.

Reprint requests to Dr. A. Stephan, Nauheimer Str. 39, D-65428 Rüsselsheim.
E-mail: astephan@uni-bremen.de

0939–5075/98/0700–0639 $ 06.00

such, one can always say it has the property (or disposition) to behave such and such. I will, however, discuss emergent *structures* separately. As we will see, they are particularly important for the interpretation of dynamical systems and connectionist nets.

Weak Emergence, Synchronic Emergence, and Diachronic Emergence

There are three theories among the different varieties of emergentism deserving particular interest: *synchronic* emergentism, *diachronic* emergentism, and a *weak* version of emergentism. For synchronic emergentism the timeless relationship between a system's property and its 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, if it is *irreducible*, i.e., if it is not reducible to the arrangement and the properties of the system's parts. In contrast, diachronic emergentism is mainly interested in *predictability* of novel properties. For such a theory, those properties are emergent that could not have been predicted in principle before their first instantiation. By the way, 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 a common 'weak' theory, which at the present pervades emergentist theorizing mainly in connectionism 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 any problems. The stronger versions of emergentism can be developed from *weak emergentism* by adding further theses.

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 (or structures) consist of material parts only. According to the thesis, all possible candidates for emergent properties, such as, e.g., being alive or being in a mental state, are instantiated only by material systems with a sufficiently complex physical microstructure. It excludes all *vitalistic* positions which hold that properties like being alive can be instantiated only by a compound consisting of an organism and some *supernatural* entity, e.g. an *entelechy* or an *élan vital*.[1] Thus, all substance-dualistic positions are rejected; for they base having cognitive states on supernatural bearers such as a *res cogitans*.[2] Hence, the thesis of physical monism denies that there are any supernatural components responsible for a system's having emergent properties. Particularly, this means that living or cognitive systems – whether artificial or natural – consist of the same parts as lifeless objects of nature. There is no reason to suppose that there are some *specific components* that belong just to those systems which are alive or able to cognize, but are missing in systems which are lifeless or unable to cognize. Instead, it is nothing but *specific* constellations of physico-chemical processes that show vital behavior or have mental qualitites.

(i) *Physical monism*. Entities existing or coming into being in the universe consist solely of material parts. Likewise, properties, dispositions, behaviors, or structures classified as emergent are instantiated by systems consisting exclusively of physical parts.

Embracing a *naturalistic* position, emergentists subscribe to a scientific empiricist position, but in so doing, they do not subscribe to reductionism.

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

[1] 'Supernatural' properties are meant to be hyperphysical, i.e., as independent from (physical) nature and their laws.

[2] In the history of emergentism, however, there were theories of emergence that did not claim the thesis of *physical monism*; instead, they took *mental* or *neutral* building blocks as fundamental (cf. Broad, 1925, p. 610–653). Anyway, the thesis of physical monism is not questioned by main stream debate today.

possible candidates for emergents. It is the thesis of systemic properties.

This thesis is based on the assumption that *general* properties of complex systems fall into two different classes:[3] (i) properties which some of the system's parts also have, and (ii) properties that none of the system's parts have. Examples of the first class are properties such as being extended and having a velocity; sometimes such properties are called hereditary properties (however, 'hereditary' is not used in a biological sense). Examples of properties in the second class are walking, reproducing, breathing and having a sensation 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.[4]

Sometimes systemic properties are characterized as 'novel' properties, only by virtue of being systemic.[5] However, this does not attribute any temporal dimension; instead it characterizes a 'timeless' *systematic* relationship: in comparison to the properties of the system's parts, the system's properties are 'new'. Thus, one could, if one liked, distinguish between *diachronic* and *synchronic* novelties. However, I prefer to characterize *systematically* novel properties as does the thesis of *systemic properties*, only *diachronic* novelties in time should be characterized by a thesis of *novelty* (see below).

It should be uncontroversial that both artificial and natural systems with systemic properties exist. Those, who would deny their existence would have to claim that *all* of a system's properties are 'hereditary' properties, that is to say, that they 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, 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 properties and dispositions to behave depend nomologically on its micro-structure, that is to say, on its parts' properties and their arrangement. There can be no difference in the systemic properties without there being some differences in the properties of the system's parts or their arrangement.[6]

Vollmer illustrates this case with an example from chemistry: The system 'graphite' consists of carbon atoms arranged in honeycomb layers. However, a tetrahedral arrangement of the same atomic parts constitutes a different system, e.g., 'diamond', with different systemic properties. By choosing different atomic building blocks, but the same structure type, again we obtain another system, e.g. 'silicon'. In each of these examples the systemic properties are nomologically dependent on properties of the system's parts and their arrangement (cf. to Vollmer, 1988, p. 93). Anyone who denies the thesis of the system's properties synchronic determination either has to admit 'free floating' properties, i.e., properties that are not bound to the properties and arrangement of its bearer's parts, or she has to suppose that some other factors, in this case non-natural factors, are responsible for the different dispositions of systems that are identical in their microstructure. In the case mentioned above, she would have to ad-

[3] General properties are properties of a general type, such as having a weight, or being liquid; they are not specific properties, such as having a weight of 154.5 pounds or being liquid by a temperature of 1200 °C.

[4] The distinction between systemic and non-systemic properties goes back to Broad (1933, p. 268). A more recent version is from Bunge (1977, p. 501 f.).

[5] Cf., e.g., Vollmer (1992, p. 188) and Lorenz (1988, p. 48).

[6] In recent debate, the thesis of *synchronic determination* is sometimes stated in a less stronger version as the thesis of *mereological supervenience*, which claims that a systems's properties (or dispositions) supervene on its parts' properties and their arrangement. Then, too, there is no difference in the systemic properties without differences in the part's properties or their arrangement (see Stephan, 1994, p. 109). The thesis of mereological supervenience, however, is weaker than the thesis of synchronic determination, since it does not claim the *dependence* of the system's properties from its micro-structure, it only claims their *covariance.*

mit, 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, that may have hardness 2 instead of hardness 10 on the Mohs-scale. This seems to be totally implausible. Equally beyond thought is that there may exist two micro-identical organisms, one 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 can distinguish colors in the ordinary way, does not hold a naturalistic-physicalistic position.[7]

Weak emergentism as sketched so far comprises the minimal conditions for emergent properties. It is the common base for 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., Bunge, and Vollmer), but also by cognitive scientists (e.g., Hopfield, Rosch, Varela, and Rumelhart) in exactly its weak form. The three features of weak emergentism – (i) the thesis of *physical monism*, (ii) the thesis of *systemic properties*, and (iii) the thesis of *synchronic determination* –, however, are compatible with contemporary reductionist approaches without further ado. Particularly, this is true because recent reductionist approaches – contrary to older variants – take into consideration the system's structures too. Thus, a system's systemic property just being non-additive does not by itself make it irreducible. Some champions of *weak* emergentism credit the compatibility of 'emergence' and 'reducibility' as one of its merits compared to stronger versions of emergentism (e.g., Bunge, 1977).

Synchronic emergentism

We come to the essential features of more ambitious theories of emergence, the theses of *irreducibility* (or *non-deducibility*) and of *unpredictability* of certain systemic properties. These theses are closely connected: Irreducible systemic properties are eo ipso unpredictable, in principle, before their first appearance. But besides irreducible properties, there also seem to be properties that can't be predicted before their first appearance on other grounds. Therefore, the thesis of unpredictability is more complex than the thesis of irreducibility of systemic properties. Thus, it is reasonable to start with a discussion of the thesis of irreducibility, which is easier to analyze, and a discussion of synchronic emergentism.

Broad's attempt to explicate a (strong) theory of emergence 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 occuring 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 'isolated' or in other (more simple) systems.[8]

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 having quite dif-

[7] However, 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.

[8] Properties that might be ascribed to a system's 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, p. 112f.).

ferent consequences are concealed.[9] 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 show when they are part of the system follows from the behavior they show in isolation or in simpler systems 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 from 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 simpler constellations 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 the system's parts. In other words: From the behavior of the system's parts it should 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 violation of the second criterion, what means that from the behavior of the system's parts in other (simpler) systems it would not follow how they will behave in the actual system. In contrast, he claims that the irreducibility of secondary qualities (e.g., the color or taste of certain objects) and phenomenal qualities (e.g., the 'how it is for us' when experiencing the colors or tastes of those objects) results already from a violation of the first condition, since they were neither adequately characterizable by the macroscopic nor by microscopic behavior of the system's parts, 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.[10] Broad has illustrated the fundamental distinction between (behavioral) *analyzable* and *unanalyzable* properties by pointing to characteristic properties of organisms and secondary qualities, respectively.

If secondary and phenomenal qualities are not analyzable,[11] 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:

(a°) *Unanalyzability.* Systemic properties which are not behaviorally analyzable – be it micro- or macro-scopically – 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) sys-

[9] As we will see, one type of irreducibility seems to imply 'downward causation', while the other seems to imply epiphenomenalism.

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

[11] 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 recent debate about qualia, particularly about the limits of functionalism can be seen as a repetition of Broad's argumentation (see below).

tems. 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 side, however, even analyzable systemic properties can be irreducible and therefore emergent. This is the case when the second criterion of reducibility will be violated, i.e., when the behavior of 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.[12] 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.[13] 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:

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

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

(iv) *Irreducibility.* A systemic property is irreducible if (a) it is neither micro- nor macro-scopically behaviorally analyzable, or if (b) the specific behavior of the system's components, over which the systemic property supervenes, does not follow from the component's behavior in isolation or in other (simpler) constellations.

Thus, we have to distinguish two totally different types of irreducibility of systemic properties. Equally different seem to be the consequences that result from them. If a systemic property is irreducible because the behavior of the system's parts, over which the property supervenes, is itself irreducible, this seems to imply that we have a case of 'downward causation'. For, if the components' behavior is not reducible to their arrangement and the behavior they show in other (simpler) systems or in isolation, then there seems to exist some 'downward' causal influence from the system itself or from its structure on the behavior of the system's parts. To be sure, if there would exist 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. 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 implied that the system's structure has a downward causal influence on the system's parts. All the more, there is no reason

[12] Bechtel holds a quite 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, p. 95). In "Mechanical Explanation and its Alternatives" Broad has examined 'in abstracto', under what conditions the behavior of system's components can be irreducible (cf. 1919, p. 113f.).

[13] Scenarios of this kind are discussed already by Mill and Fechner (cf. Stephan, 1998, sections 6.1. and 7.3.); see also McLaughlin who seems to think that emergentism has to allow configurational forces (1992, 52ff.).

to assume that unanalyzable properties themselves exert a causal influence on the system's parts. Rather it is to ask, how unanalyzable properties might have any causal role to play 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 could be causally effective themselves. If, however, one can not see *how* unanalyzable properties might play a causal role, then, it seems, such properties are epiphenomena.

Diachronic emergentism

In a systematic examination of systemic properties that are exemplified in our world – for example, chemical, vital, or mental properties –, the question that arises is whether the properties are reducible. In contrast, their predictability is, so to speak, of no account. Within the scope of 'emergent evolution', however, importance of the two questions seems the reverse: While the question concerning the reducibility of emerging properties seems less relevant, it is of particular interest what systems and properties might have been predictable, at least in principle, before they were actually exemplified. Likewise, predictability of properties that are be expected plays an important role in the development of novel artefacts. Here, however, it is not predictability in principle that matters. What matters is practical predictability – it's better to know before you make the 'Elch test' whether your newly constructed car will pass it or not. Unforeseen events of this type have little to do, however, with emergence in the theoretically interesting sense.

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.

(v) *Novelty.* In the course of evolution exemplifications of 'genuine novelties' occur again and again. Already existing building blocks will develop new constellations; new structures will be formed that 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 fundamentally different reasons: (i) they can be unpredictable because the micro-structure of the system, which exemplifies 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 depend nomologically on it. (ii) However, a property can be unpredictable even though the 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. This does not preclude that further occurrences of this property might be predicted adequately. Since in the second case criteria for being unpredictable are identical with those for being irreducible, this notion of unpredictability will offer no theoretical gains beyond those afforded by the notion of irreducibility.[14]

Let us focus, therefore, on the first case: *unpredictability of structure.* This version of unpredictability passed almost unnoticed in 'classical' literature on emergentism during the 1920s, but because of strong interest in dynamical systems and chaotic processes this notion gains considerable significance.

The structure of new formed systems can itself be unpredictable for several reasons. Thus, belief in an indeterministic universe implies that there

[14] A difference in extension between both notions could result only in respect to those properties, which, although reducible, are not predictable before their first appearance. A reducible property is unpredictable before its first appearance if the behavior of the system's components upon which it supervenes does not follow from their behavior in those systems that exist at the time of prediction. The notion of unpredictability widened in this way would depend, however, in a very contingent way on the chronological order of systems coming into being in evolution. Therefore, such a notion should not be important in qualifying the notion of emergence.

will be novel, unpredictable structures. However, from an emergentist perspective it would be of no interest, if a new structure's appearance would be unpredictable only because its coming into being is not determined, not to mention that most emergentists claim, anyway, that the development of new structures is governed by deterministic laws. But still deterministic formings of new structures can be *unpredictable in principle*, if they are governed by laws which are attributed to deterministic chaos.

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 exercises an influence on 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, thus, the origin is the attractor. For $1 < \mu < 3$ 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 slowly – an indication for fundamental change in its behavior. For larger values dynamic 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 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 as if just 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 – we today do not count properties and dispositions of chemical compounds any more among *synchronic* 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 creature of phantasy to have. For example, in Alexander's considerations (cf. 1920, ii, pp. 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.[15] I will leave it

[15] Laplace himself assumed that his calculator knows all laws governing nature, but he took those laws to be laws of Newtonian physics. During his time, nobody knew anything about aperiodic chaotic processes.

open whether or not it is plausible to ascribe such a 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.

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

Summing up, it may be said a *systemic property* is *unpredictable* in principle before its first appearance, if (i) it is irreducible, or if (ii) 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.

Synopsis

The following figure depicts the logical relationship that holds between different versions of emergentism.

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 sciences, particularly for characterization of systemic properties of connectionist nets, 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 for debating nonreductive physicalism and qualia. It is not compatible with reductive physicalism any more. *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 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 far weaker than synchronic emergentism. Theories of deterministic chaos (in dynamical systems) can be acknowledged as a type of structure emer-

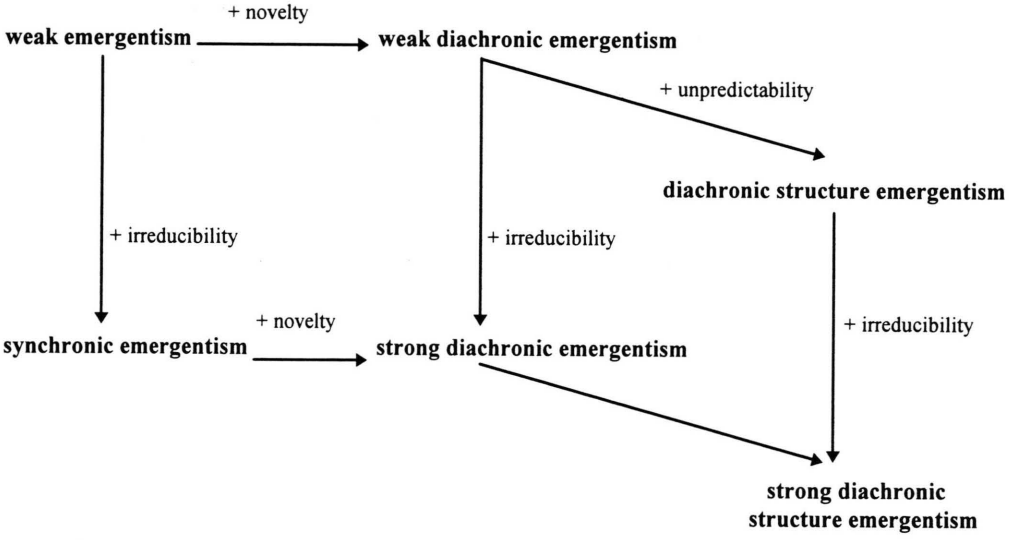

Fig. 1 depicts the logical relationships that hold between different versions of emergentism.

gentism. Likewise, its perspective is important for evolutionary research. In comparison to the above mentioned versions of emergentism the synthetic position of *strong structure emergentism* has no equivalent in recent discussion. Most important from a theoretical point of view are *weak emergentism*, *synchronic emergentism*, and *diachronic structure emergentism*.

Emergence in the Qualia Debate, in Connectionism, and in Synergetics

Let us turn to some specific cases to consider how the different concepts of emergence are applicable. We will see that in one of the central debates within philosophy of mind, namely the debate about the nature of qualia, there is need for a strong notion of emergence such as that of synchronic emergentism. In contrast, in connectionism and synergetics weaker notions of emergence are employed.

Emergence in the debate over whether qualia are physical

In recent debate about qualia,[16] Nagel, Block, Jackson, Levine, and McGinn, among others, have argued in one way or another that qualitative mental phenomena are not reducible to physical or functional states, respectively. If their arguments succeed, they imply emergentist or substance-dualistic positions. Most interesting and powerful seems to be Levine's so-called 'explanatory gap'-argument, which I will consider closely in the following.

Levine starts with comparing two statements, namely (i) 'pain is the firing of C-fibers', and (ii) 'heat is the motion of molecules'. The decisive difference between the two identity statements, according to Levine, is that the second is *fully explanatory*, while the first is not:[17] "there is a 'gap' in the explanatory import of these statements" (1983, p. 357). The second identity statement is assumed to be 'fully explanatory', because knowledge of natural laws helps us understand why the motion of molecules has exactly the causal role usually ascribed to heat. In doing so, it is presupposed – and this is of great importance – that the macro-physical concept of heat can be fully explicated by heat's causal role.

"[Statement (ii)] is explanatory in the sense that our knowledge of chemistry and physics makes intelligible how it is that something like the motion of molecules could play the causal role we associate with heat. Furthermore, antecedent to our discovery of the essential nature of heat, its causal role [...] exhausts our notion of it. Once we understand how this causal role is carried out there is nothing more we need to understand" (1983, p. 357).

In other words: statement (ii) is fully explanatory because some system's property of heat is reducible in respect to the motions of molecules, and so far it is not emergent. On the other hand, the reason for statement (i) not being fully explanatory is, according to Levine, that the notion of pain (as are notions of other phenomenal states) is not exhausted by the causal role of pain. The decisive point is that there is no reason to believe that firing of C-fibers fits better to typical pain experiences than to any other phenomenal experience.

In a series of articles, Levine and Hardin argued about the 'explanatory gap'-argument's core thesis by concentrating on experiences of colors (see Levine, 1983; 1991; and Hardin, 1987; 1988; 1991). Levine's last 'move' is the following:[18]

"If inverted qualia are possible, then the question why red things look reddish and not greenish has no adequate answer in physical/functional terms. Whereas, if inverted qualia are not possible, then the question why red things look reddish, and not some other (perhaps unimaginable) way or no way at all, still has no adequate answer in physical/functional terms. An explanatory gap persists" (1991, p. 39).

Levine here seems to refer to objections that were originally directed at functionalism, namely the 'inverted qualia'-argument, and the 'absent

[16] The term 'qualia' denotes the experiential properties of conscious states, i.e. of sensations, feelings, perceptions, and of propositional attitudes.

[17] Talk about 'firing of C-fibers' has its firm place in literature about the qualia-problem. It is, however, only a fill-in for an adequate neurophysiological analysis of a pain's material base.

[18] Articles Levine has published subsequently have not changed the debate considerably.

qualia'-argument.[19] In his article "On Leaving Out What It's Like" he confirms this: "The basis of my argument for the existence of this explanatory gap was the conceivability of a creature's instantiating the physico-functional property in question while not undergoing an experience with the qualitative character in question, or any qualitative character at all" (1993, p. 130). The 'explanatory gap'-argument, however, does not depend on the empirical possibility of absent or inverted qualia. Even if they were not actually possible, there is no way to see how qualia could be reduced to neurophysiological or functional states. That is because it does not even seem to be possible in principle to deduce from generally obtaining natural laws that the micro-structures of those systems that have phenomenal states have all features characteristic of phenomenal states.

In response to Levine, Hardin has tried to make clear which explanatory gap he thinks is closable at least in principle, and which one must be left open. If we assume that color experiences are phenomenally simple and unstructered, then it looks difficult indeed to tie them explanatorily to physiological processes. In fact, however, we should acknowledge that at least some colors show structures that fit well to neuronal processes.

"We would for example, expect the neural process associated with orange to be relatively more complex than the process associated with red, and in fact, expect that the 'red-making' process would be in some fashion incorporated in the 'orange-making' process on the grounds that perceived redness is an ingredient in perceived orangeness" (1991, p. 44).

Hardin concedes, however, that it is far easier to explain the difference between experiencing red and experiencing green than to explain experiencing red or experiencing green themselves. We could develop an understanding of the differences between both experiences without being able to explain each experience in itself. Thus, Hardin seems to admit that prospects are slim to answer the really central question whether perfect knowledge of neuronal processes could allow to deduce statements about the qualitative states that are correlated with (or supervene upon) such processes.[20]

Let's see why qualia are so resistent to reduction. According to Levine, reduction that is explanatory requires two stages: "Stage 1 involves the (relatively? quasi?) *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. Stage 2 involves the empirical work of discovering just what those underlying mechanisms are" (1993, p. 132).

If one claims that a reduction that is explanatory is impossible *in principle*, as is claimed for qualia, that does not imply a failure of the second task. What is implied is a failure, in principle, of the first task. Apparently, phenomenal properties cannot be individuated by their causal roles: "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. Reduction is explanatory when by reducing an object or property we reveal the mechanisms by which the causal role constitutive of that object or property is realized. Moreover, this seems to be the only way that a reduction could be explanatory. 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). Thus, Levines *synchronic qualia-emergentism* is based on two theses:

(1) The reduction of a systemic property P is explanatory if and only if the realization base exhausts exactly the causal role which is constitutive of P.

(2) Phenomenal properties (or states) are not fully graspable by the features of their causal role.[21]

[19] Inverted qualia would exist, e.g., if things we call 'red' look to somebody else the way things look we call 'green', and vice versa. Qualia are called absent, if a system has no phenomenal experiences at all, although it shares with us functional descriptions of phenomenal states. See, for example, Block (1978; 1980).

[20] In the meantime, others have suggested further ways to close the 'explanatory gap'. See, for example, Kirk (1996) and Kurthen (1996).

[21] Notice the analogies between Levine's theses and the criteria for irreducibility in the section above.

Several responses are possible to this analysis: (i) one tries to avoid emergentism by showing either how properties that are not fully graspable via their causal roles, might yet be reducible explanatorily, or by showing that phenomenal qualities are graspable adequately via causal roles. For both variants, however, there exists nothing more than first attempts or statements of intentions.[22] (ii) One accepts both theses and, thereby, an emergentist position. But then, additional questions arise: To what type of psycho-physical theories does synchronic qualia-emergentism belong? Can it still be seen as a physicalistic position, or has it to be treated as a kind of property dualism? And furthermore: What solution can emergentism offer to the problem of mental causation?

Perspectives on whether an emergentist theory of qualia can be seen as a physicalist position have changed considerably during the past twenty years. The older arguments presented by Nagel, Kripke, and Jackson had, among other things, the aim to show that physicalism is an inadequate position. Levine changed the situation by pointing to the fact that various 'conceivability'-arguments do not necessarily have ontological consequences:

"If these thought experiments show that physicalism leaves something out, it can't be in the sense that there are facts that physicalistic descriptions fail to pick out, [...] perhaps, Cartesian conceivability arguments can't demonstrate that qualia aren't physical states or processes, but they at least throw the burden of argument back onto the physicalist to show why we should think they are physical states or processes" (1993, 125 f.).

Levine's interpretation would allow, therefore, to treat synchronic qualia-emergentism as a physicalistic position. As things stand, however, physicalists have turned the tables. They demand that genuine physicalist positions may not leave out *explanatorily* anything (cf. Beckermann, 1996b and Horgan, 1993). According to this criterion, qualia-emergentism can not be seen as a physicalistic position any more. One reason for the physicalists' becoming more fastidious might be that emergentists seem to face a hopeless dilemma when confronted with the problem of mental causation.[23] In the presence of this, the challenge for the physicalist to manage those 'portions' of qualia that escape individuation by causal roles seem to be negligible compared to the emergentist's problems with mental causation. For, if qualitative states can not be individuated via causal roles, then it is questionable whether they can play any causal role at all.

Synergetics and emergentism

Synergetics – an interdisciplinary theory of cooperation – addresses an original subject of emergentism: the *coming into being of novel systemic properties.* Particularly, it is concerned with qualitative changes in a system's behavior that have become instable by change of specific controlling parameters (see Haken, 1996, pp. 587 and 593). Synergetics' starting point is the observation that both in animate and in inanimate nature, there exist numerous systems that spontaneously form by themselves – that is to say, in a self-organizing way – *new* structures. Classical examples are the entirely novel light of lasers in comparison with that of ordinary lamps,[24] or the structure formation of slime mold.[25] Both examples and many similar ones have in common that small disturbances of control parameters give rise to instable states out of which new structures emerge by processes of self-organization. The new structures often instantiate qualitatively new properties compared to those the system instantiated before – a subject that has been in the center of emergentists' considerations.

Synergetics looks for the regularities that lie behind various self-organizing systems which usually

[22] Cf. to Beckermann, Kirk, Levine and Rey (all 1996).

[23] The seeming dilemma, however, is not hopeless for the emergentist, as I argued in "Armchair Arguments Against Emergentism" (1997).

[24] When only a small amount of energy is pumped into the device, the laser operates as a lamp. Then, its main elements, namely specific atoms, emit lightwaves of about 3 m coherence length independently of each other. If the pump power is increased to a certain threshold, the atoms oscillate in phase and emit a giant wavetrack of about 300,000 km length.

[25] Slime mold (*Dictyostelium disciodeum*) normally exists in the form of single amoebic cells. If nutriment runs short, the cells assemble at a certain place, pile up, and then differentiate into spores and stalk cells. Slime mold can move then as a whole. It wriggles on the ground like a snake (cf. Haken, 1988, p. 99–102).

are studied by such distinct disciplines as physics, chemistry, or biology. Such as theories of emergence synergetics has as a characteristic feature a naturalistic attitude: It takes it for granted that processes which lead to new structures are triggered and maintained by the system's parts' cooperation, and not by external 'organizers' or 'supernatural' entities. In contrast to stronger versions of emergentism, synergetics claims that formation of new structures as well as properties instantiated by them can be explained and forecasted by reference to the parts' cooperation. Hence, from a synergetic point of view neither should new arising structures be seen as *structure emergent*, nor should properties instantiated by these structures be seen as *synchronic emergent*.

Of central significance for synergetics are such notions as 'order parameter' or 'enslaving' introduced by Haken himself. Quantities that are seen as order parameters are thought to do a double job: On the one hand, they are used to *describe* very complex processes of self-organization in a simple but yet adequate way; on the other hand, they are used to *explain* causally the ongoing processes (cf. 1996, p. 588, and 1987, p. 139f.). Haken's 'thesis of description' (TD) is supported by numerous examples, and by a mathematical formalism developed in (1983):[26]

(TD) By specifying the behavior of order parameters of a system, the system's behavior is adequately characterized. Instead of describing a system's behavior by specifying the behavior of all its single parts, it is sufficient just to specify the behavior of few relevant order parameters.

Thus, the state of a laser can be described, in principle, in two ways: first, on a microscopic level by specifying the individual states of all electrons, or second, by exploiting that laser's light has come up to a macroscopic state. Here, a far smaller amount of information, e.g., wavelength and amplitude of emitted light, is sufficient to describe characteristic features of lasers. Thus, transition from microscopic to macroscopic descriptions significantly reduces the information which is necessary to characterize systems adequately (cf. 1987, p. 140). However, from obvious *compressibility of information* Haken seems to infer *compressibility of causal factors*; for he is interpreting order parameters that are employed to describe a system in a simple way as causal agents:

(TC) The behavior of the system's parts is *fixed* by just a few quantities, namely the order parameters.

Sometimes Haken puts the 'thesis of causality' (TC) in a specific jargon of synergetics, and calls it the 'enslaving principle' (EP), which strictly speaking means that "slowly changing quantities (e.g., gradually increasing order parameters) enslave rapidly relaxing motions" (Haken and Haken-Krell, 1989, p. 27, my translation; see also Haken, 1988, p. 20 and 1996, p. 588).

(EP) Order parameters *enslave* a system's individual parts, that is to say, they determine the parts' behavior.

The principle of 'enslaving' a system's parts by order parameters is a specific case of downward causation: Order parameters are macroscopic quantities. They refer to states or properties of the whole system. As systemic quantities, they influence the system's part's behavior. Since Haken supposes that order parameters emerge only from cooperation of the system's parts, he prefers to use the notion 'circular causality' to characterize the presumed mutual interaction between system's parts and order parameters.

The term 'enslaving principle' is not to be understood as a metaphor, although Haken's diction is full of anthropomorphisms and metaphors when he undertakes to illustrate it. Originally, it is introduced and justified within a presentation of a mathematical formalism that is developed to describe adequately processes of self-organization (see 1983, p. 208ff.). In his own interpretation of the mathematical formalism, Haken, however, jumps from a purely structural reading to a causal reading of the mathematical relations – a clear

[26] To give some examples, order parameters are, according to Haken, prevailing light waves in a laser which 'force' all emitted electrons to oscillate in their way (cf. 1988, p. 68), spiral fields of concentration of cAMP-molecules (cyclic Adenosin Monophosphate) of the slime mold (1987, p. 142) as well as language, form of government, or working climate in social systems (1996, p. 590).

case of the 'post hoc, ergo propter hoc'-fallacy.[27] For, the functionally described relation of succession between order parameters and system's parts, does not warrant an extrapolation to a causal relation between order parameters and system's parts.

Haken's causal theses appear rather odd when applied to order parameters within the scope of psycho-physical theories or approaches in the social sciences.

"Let us choose here an extreme case, the brain. We treat neurons and their connections as sub-systems. Chemical and electrical activities of neurons can be described by numerous microscopical variables. However, strictly speaking it is the thoughts that are efficacious as order parameters. Both components are dependent on each other" (1983, p. 15f., my translation).[28]

Haken's claim implies that thoughts 'enslave' neurophysiological processes. Both in this case and in social systems, Haken ascribes to order parameters a causal role they just do not have. There is no *modus operandi* according to which the working climate 'enslaves' some clerk (see Haken, 1996, p. 590). The working climate does nothing at all. It is an expression of the atmosphere within a group, it tells something – by compressing information! – about the manner of how members of a group deal mostly with each other, but it does not deal itself with the members of the group. The same is true for other order parameters of social systems listed by Haken (ib.): Neither national character, public opinion, nor ethics cause anything by themselves. Nor is anybody 'enslaved' by them.

To sum up, we can say that synergetics does neither establish a strong nor a novel version of emergentism. It does not treat novel structures that are formed at instabilities as unpredictable and thus as structure emergent.[29] Nor does it treat systemic properties or macroscopic quantities such as order parameters as irreducible and thus synchronically emergent. Instead, synergetics attempts to explain the coming into being and persistence of novel structures and their properties by processes of self-organization. Moreover, synergetics does not establish a scientifically based kind of downward causation, since a reading of the mathematical formalism that jumps from a sequential relation to a downward causal relation is not warranted.

Emergence in connectionism

In the last decade, connectionism has received great attention in cognitive science. Its core idea is to assume a network of elementary units that have a certain level of activation. Units are connected with each other. Units whose activation exceeds a certain threshold, can activate or inhibit other units according to certain weights that specify prevailing connections.

To see to what extent emergentist considerations are relevant for connectionism, I shall first examine more closely the parameters that specify a connectionist net. Each net is determined essentially by three factors: (i) by the number of units and connecting links which hold between them; (ii) by the function that determines the level of activation for each unit; and (iii) by the rule that determines how connection weights will change.

In each case the number of units and the links between them are fixed; they make up, so to speak, the 'skeleton' of a network which is static under ordinary circumstances: neither the number of units nor the structure of their links will change. A system's actual dynamic results from the possibility of modifying the weights of its internal connecting links. From a macroscopic point of view, these continuous processes of accomodation can be seen as learning procedures. Thus, a connectionist net 'learns' by locally determined changes of its connections' weights, and not by adding some further propositions to its data base. It's a big challenge for connectionist researchers to construe nets that are able to optimize link weights by themselves in a 'training phase' such that they are able to produce given outputs. When a net is

[27] From a mathematical point of view Haken's approach just means that solutions of differential equations depend essentially upon slowly changing quantities, the order parameters, while fast-moving quantities can be neglected. A more detailed presentation of Haken's mathematical considerations is given in Stephan (1998, section 18.1.).

[28] See also Haken and Haken-Krell (1989, p. 134).

[29] Thus, Haken says, "it is exactly calculable which collective motion will win in the end" (1988, p. 49, my translation). It is only that breakings of symmetry depend on very small deviations, and so are often unpredictable for practical reasons.

fed with various inputs after its practice time, it will calculate its outputs or 'answers' with stable weights in general.

Nets are most efficient in pattern recognition and generalization (e.g., in acquiring rules and forming schemata). Rumelhart, Smolensky, McClelland, and Hinton describe a simple net that is an ideal case in point to illustrate how connectionist nets instantiate schemata:[30]

We all have ideas of what typical functional rooms look like. Living rooms have sofas and easy chairs, but they usually do not have bath tubs and refrigerators; in offices and studies we might find desks, bookshelfs, and computers, but no toilets. Rumelhart *et al.* represented this knowledge about specific rooms' typical furniture and equipment in a connectionist net. The net consisted of 40 units each of which represented a specific room feature. Since this example was to illustrate schemata formation they refrained from training the net; instead, they set the link weights in advance.

By way of calculation there are 2^{40} possible binary states in which the system could potentially settle. But actually, only five maxima which correspond to five specific room types crystallized. For example, by beginning a cycle with the 'oven'-unit, the system successively adds the units for 'ceiling', 'coffe-cup', 'sink' and 'refrigerator'. After 400 updates all features of a prototype kitchen were activated (see Rumelhart *et al.,* 1986, p. 25). Complete schematas for kitchen, bathroom, office, living- and bedroom could be activated by starting with one of these room's characteristic features, respectively. Besides these main schemata several 'subschemata' are hidden. They can be activated by activating, e.g., 'bed'- and 'desk'-unit together. Then we get a schema of, say, a student's room.

Behavior and properties of connectionist nets give rise to emergentist considerations in many ways. Three aspects should be discerned: first, connectionist nets obviously have systemic properties, that is they have properties their parts do not have. Thus, a net's properties are at least weakly emergent. However, it remains to be determined whether systemic properties of nets are emergent in a strong sense, namely synchronically emergent. Secondly, many systemic properties of nets seem to be emergent in a 'phenomenological' way. By this, I mean that the properties appear, or come into being by themselves – as is the case with the five rooms' schemata –, if nets get adequate stimuli. These facts of the case are referred to in English by the word 'emergent' in its ordinary use, however. No specific theory of emergence is implied by this usage. Since some connectionists make use of the word 'emergent' in its ordinary use intermingled with a more technical use, it is very important to tell the notions apart. Eventually, connectionist nets develop during their training phase – in a somewhat mini-evolutionary process – their 'soft' structures, by which I understand the specific distribution of link weights.[31] The net's systemic properties discussed above supervene upon these structures. Now, the question arises whether this formation of structure is an interesting case of structure emergence.

Let's examine first the relationship between global net properties and their realization base, namely the net's structure and its part's properties. Considerations of connectionist net's architectures and their modes of operation reveal that only trained nets show typical macroscopic properties such as 'rule following', 'schemata formation', or 'pattern recognition'. Untrained nets do not have those (cognitive) properties, they have only the disposition to acquire them. Macroscopic properties of trained nets supervene upon both their given hard structure, and their acquired soft structure. They are fully reducible to the organization of the net in consideration, the properties of its units (namely their activation formula), and the properties of links consisting between its units (nameley distribution of weights, and formula for changing weights). If these quantities are known, the output-behavior of any net can be predicted exactly and explained. It is obvious that a net's parts, namely its units and the links between them, do not have any of those macroscopic (cognitive) properties. So far, these properties a net acquires by training are typical systemic properties. However, since they are not irreducible systemic prop-

[30] The term 'schemata' refers as related terms as 'scripts', or 'frames' to coherent structures of knowledge, which are assumed to be fundamental for reasonable interactions with the world.

[31] The 'hard' structure of a net (its 'skeleton') is fixed by the number of units and the links between them. Usually, it is invariant.

erties, but are completely deducible from a net's structure, and its parts' and links' properties, a net's systemic properties are merely weakly emergent. They are not synchronically emergent.

Rumelhart and McClelland discuss in great detail net properties which they call 'emergent', and stress that connectionist approaches do not imply reductionistic, but interactionistic positions.

"We are simply trying to *understand* the essence of cognition as a property emerging from the *interactions* of connected units in networks. We certainly believe in emergent phenomena in the sense of phenomena which could never be understood or predicted by a study of the lower level elements in isolation. [...] This is the case in many fields. For example, we could not know about diamonds through the study of isolated atoms [...] Features such as the hardness of the diamond is understandable through the interaction of the carbon atoms and the way they line up. [...] Knowing about the individuals tells us little about the structure of the organization, but we can't understand the structure of the higher level organizations without knowing a good deal about individuals and how they function. This is the sense of emergence we are comfortable with" (1986, p. 128).

However, Rumelhart and McClelland's claim that connectionism amounts to a non-reductive position results from an inappropriate strong notion of reduction. Both authors, one must know, take a system's properties as reduced only when either some of the system's parts have these properties already, or when the system's property can be reduced to some linear interactions of the parts. Both possibilities of reduction, however, are given neither in the case of carbonaceous compounds, nor in the case of connectionist nets. Yet, almost no systemic property would be reducible according to these criteria. At the same time, Rumelhart and McClelland concede that a net's behavior is completely intelligible, if one takes into account all interactions between its units, that is to say, if one considers a system's hard and soft structure. Hence, even from their point of view, connectionism is not an instance of synchronic emergentism.

In discussions of connectionist nets' behavior ordinary use and technical use of the notion of emergence often gets intermixed. Thus, for example, Rumelhart and McClelland reiterate that macroscopic properties emerge from micro-level interactions: "[M]any of the constructs of macro-level descriptions such as schemata, prototypes, rules, productions, etc. can be viewed as *emerging* out of interactions of the microstructure of distributed models" (1986, p. 125; italics are mine). By this characterization, they lay stress upon the fact that a net's systemic properties come into being from the complex behavior of the net's components: rules and schemata can be available without being explicitly fed into the system. Clark discusses a net's ability to develop schemata from this point of view, too. He refers to Rumelhart, Smolensky, McClelland, and Hinton: "[they] detail a PDP model in which the properties of explicit stored schemata *emerge* simply from the activity of a network of units that respond to the presence or absence of microfeatures of the schemata in question" (1989, p. 93; my italics).[32] And he adds: "activating the unit for one prominent kitchen feature (e.g., the cooker) would activate in a kind of chain all and only the units standing for items commonly found in kitchens. Here, then, we have an emergent schema in its grossest form" (1989, p. 94).

Within the quoted texts, it is not emergence in a technical sense that is being discussed, but rather a system's abilities to acquire systemic properties by self-organizational processes, which can be ascribed quite adequately by an ordinary use of the term 'emergent'. The temporary manifestation of schemata that were already latently 'in' the link's weights is interpreted, then, as an emergent property of the net. Connectionists, thus, mainly point to 'emergent rules' or 'emergent schemata' to demarcate their position from 'classical' representationalism, accordingly to which all rules and schemata have to be fed in explicitely (see Horgan and Tienson, 1996).

There is a further feature of connectionist nets that provokes emergentist considerations: during the phase of training or learning a net runs through a mini-evolutionary process – link weights are adapted such that the net is enabled to handle the tasks it is supposed to master. Within

[32] The term 'PDP model' refers to connectionist nets, too. It is an abbreviation for a kind of processing going on in them, namely 'parallel distributed processing'.

this time the so-called 'soft structure' of the net develops.[33] Only when the links' weights are adjusted adequately does a net has available desired macro-properties, that is, only then can it develop schemata, recognize patterns, or make use of rules. Those are not implemented explicitely as they are in symbol manipulating devices, but are extracted from given material. However, this is not a case of genuine structure emergence: not only does the distribution of weights result from deterministic principles, the changes of weights are even calculable exactly, if we know the learning rule, the activation formula, the unit's initial activation, the intial weights, and the inputs. Were this not the case, Rumelhart *et al.* could not have skipped the training phase to calculate weights 'on foot' and feed them into the system in advance.

Even though we therefore should not speak of structure emergence in connectionist networks, regarding their soft structure, nets show a tremendous plasticity, when compared with other objects, even when compared with other dynamical systems. Chemical compounds, to give an example, have no degrees of freedom to change their internal structure. In this respect connectionist nets differ clearly from seeming analogic cases such as diamonds, which were referred to by Rumelhart and McClelland to explain 'emergent' system's properties (cf. 1986, p. 128). Diamonds are not dynamical systems that realize only after a certain number of reiterated steps the property of being hard. The diamond's property of being hard is always manifest, it does not emerge.

To sum up, connectionist nets do not instantiate any stronger type of emergence. Neither are the net's properties synchronically emergent, nor is the formation of a net's soft structure a case of structure emergence. In the weak sense in which macroscopic properties of nets are emergent, all systemic properties of complex systems are emergent. A difference to many other systems exists at best in the plasticity of nets and in their capacity to develop in a training phase by themselves adequate 'attractors' to cope with given tasks. Corresponding macroscopic properties will become manifest only temporarily during treatments. This second order property might justifiably be characterized as 'phenomenological' emergence. Only some dynamical systems have this property.

[33] The developing 'soft structure' depends on two factors, the net's hard structure, and the given inputs, that is external influences on it.

Conclusion

In particular, I have distinguished three versions of emergentism: weak emergentism, (strong) synchronic emergentism, and diachronic structure emergentism. Synchronic emergentism results from weak emergentism by adding the thesis of irreducibility. It turned out that this version of emergentism is important for the philosophy of mind, particularly for debating qualia. However, this does not establish that qualia are emergent phenomena in the strong sense. But they are good candidates for being so. Synergetics, on the other hand, does not treat any phenomenon as synchronically emergent, nor does it treat novel structures as candidates for structure emergence. Phenomena discussed by synergetics are emergent only in the weak sense. The same is true for properties of connectionist nets. As we have just seen, connectionist nets do not instantiate any stronger type of emergence.

Acknowledgements

First of all I wish to thank Judy McLaughlin who helped a lot to transform my German English into American English. I am also very grateful to Wolfgang Buschlinger, Joachim Schummer, Manfred Stöckler, Gerhard Vollmer and Henrik Walter for their comments on an earlier version of this article.

Alexander S. (1920), Space, Time, and Deity. Two Volumes. Macmillan, London.
Bates E. (1979), The Emergence of Symbols. Academic Press, New York.
Bechtel W. (1988), Philosophy of Science. An Overview for Cognitive Science. Lawrence Erlbaum Associates, Hillsdale, NJ.
Beckermann A. (1996a), Visual information-processing and phenomenal consciousness. In: Conscious Experience (T. Metzinger, ed.). Schöningh, Paderborn, 409–424.
Beckermann A. (1996b), Eigenschafts-Physikalismus. Zeitschrift für Philosophische Forschung **50,** 3–25.
Block N. (1978), Troubles with Functionalism. In: Perception and Cognition (C. W. Savage, ed.). Minneapolis, 261–325.
Block N. (1980), Are Absent Qualia Impossible? The Philosophical Review **89,** 257–274.
Broad C. D. (1919), Mechanical Explanation and its Alternatives. Proceedings of the Aristotelian Society **19,** 86–124.
Broad C. D. (1925), The Mind and its Place in Nature. Kegan Paul, Trench, Trubner & Co., London.
Broad C. D. (1933), Examination of McTaggart's Philosophy, Vol. **1,** Reprint. New York (1976).
Bunge M. (1977), Emergence and the Mind. Neuroscience **2,** 501–509.
Churchland P. (1988), Matter and Consciousness. Revised Edition. MIT Press, Cambridge, Ma.
Clark A. (1989), Microcognition. Philosophy, Cognitive Science, and Parallel Distributed Processing. MIT Press, Cambridge Ma., London.
Hacking I. (1975), The Emergence of Probability. Cambridge University Press, Cambridge.
Haken H. (1983), Synergetik. Eine Einführung. Zweite Aufl. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo.
Haken H. (1987), Die Selbstorganisation der Information in biologischen Systemen aus der Sicht der Synergetik. In: Ordnung aus dem Chaos (B.-O. Küppers, Hg.). Piper, München, 127–156.
Haken H. (1988), Erfolgsgeheimnisse der Natur. Synergetik: Die Lehre vom Zusammenwirken. Ullstein Sachbuch, Frankfurt, Berlin.
Haken H. (1996), Synergetik und Sozialwissenschaften and Replik. Ethik und Sozialwissenschaften **7,** 587–594, 658–675.
Haken H. and Haken-Krell M. (1989), Entstehen von biologischer Information und Ordnung. Wissenschaftliche Buchgesellschaft, Darmstadt.
Hardin C. L. (1987), Qualia and Materialism: Closing the Explanatory Gap. Philosophy and Phenomenological Research **47,** 281–298.
Hardin C. L. (1988), Color for Philosophers: Unweaving the Rainbow. Hackett, Indianapolis.
Hardin C. L. (1991), Reply to Levine. Philosophical Psychology **4,** 41–50.
Horgan T. (1993), From Supervenience to Superdupervenience: Meeting the Demands of a Material World. Mind **102,** 555–586.
Horgan T. and Tienson J. (1996), Connectionism and the Philosophy of Psychology. MIT Press, Cambridge, Ma., London.
Kirk R. (1996), How is Consciousness Possible? In: Conscious Experience (T. Metzinger, ed.). Schöningh, Paderborn, 391–408.
Kurthen M. (1996), On the Prospects for a Naturalistic Theory of Phenomenal Consciousness. In: Conscious Experience (T. Metzinger, ed.). Schöningh, Paderborn, 107–122.
Levine J. (1983), Materialism and Qualia: The Explanatory Gap. Pacific Philosophical Quarterly **64,** 354–361.
Levine J. (1991), Cool Red. Philosophical Psychology **4,** 27–40.
Levine J. (1993), On Leaving Out What It's Like. In: Consciousness (M. Davies and G. W. Humphreys, eds.). Blackwell, Oxford, Cambridge, Ma, 121–136.
Levine J. (1996), Qualia: Intrinsic, Relational – or What? In: Conscious Experience (T. Metzinger, ed.). Schöningh, Paderborn, 277–292.
Lorenz K. (1988), Die Rückseite des Spiegels. Neuausgabe. Piper, München & Zürich.
McLaughlin B. (1992), The Rise and Fall of British Emergentism. In: Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism (A. Beckermann, H. Flohr, and J. Kim, eds.). de Gruyter, Berlin, New York, 49–93.
Rey G. (1996), Towards a Projectivist Account of Conscious Experience. In: Conscious Experience (T. Metzinger, ed.). Schöningh, Paderborn, 123–142.
Rumelhart D. E. and McClelland J. L. (1986), PDP Models and General Issues in Cognitive Science. In: Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. **1** (D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, eds.). MIT Press, Cambridge, Ma., London, 110–146.
Rumelhart D. E., Smolensky P., McClelland J. L. and Hinton G. E. (1986), Schemata and Sequential Thought Processes in PDP Models. In: Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. **2** (J. L. McClelland, D. E. Rumelhart, and the PDP Research Group, eds.). MIT Press, Cambridge, Ma., London, 7–57.
Stephan A. (1994), Theorien der Emergenz – Metaphysik oder? Grazer Philosophische Studien **48,** 105–115.
Stephan A. (1997), Armchair Arguments Against Emergentism. Erkenntnis **46,** 305–314.
Stephan A. (1998), Emergenz. Von der Unvorhersagbarkeit zur Selbstorganisation. Dresden University Press, Dresden.
Vollmer G. (1988), Evolutionäre Erkenntnistheorie und Leib-Seele-Problem. In: ders., Was können wir wissen? Band **2.** Die Erkenntnis der Natur. Hirzel, Stuttgart, 66–99.
Vollmer G. (1992), Das Ganze und seine Teile – Holismus, Emergenz, Erklärung und Reduktion. In: Wissenschaftstheorien in der Medizin (W. Deppert, H. Kliemt, B. Lohff and J. Schaefer, eds.). de Gruyter, Berlin, New York, 183–223.