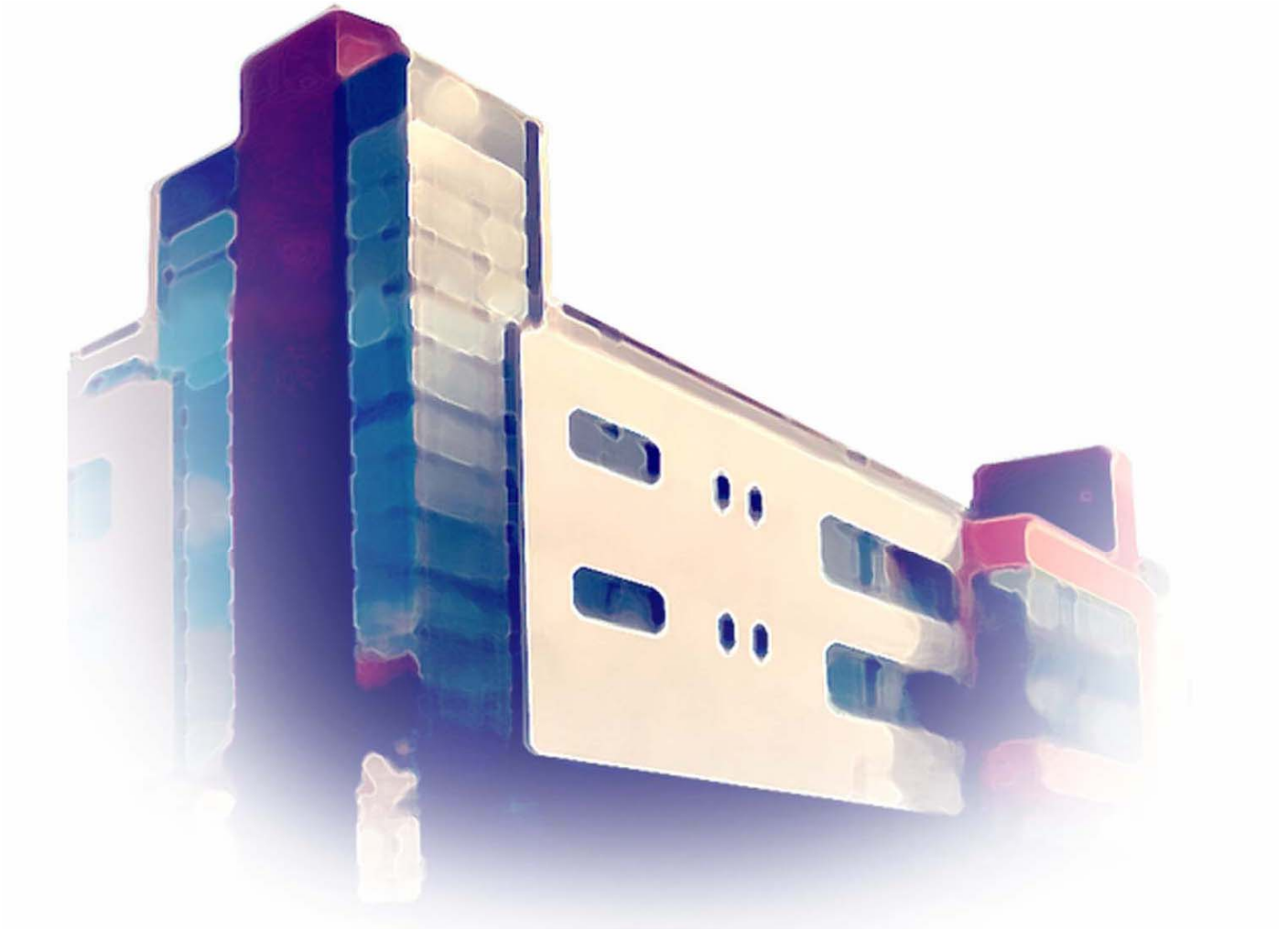
Jan Scholz

# *Emergence in Cognitive Science: Clark's Four Proposals to the Emergentists*



PICS

*Publications of the Institute of Cognitive Science*

*Volume 10-2004*

ISSN: 1610-5389

Series title: PICS
Publications of the Institute of Cognitive Science

Volume: 10-2004

Place of publication: Osnabrück, Germany

Date: November 2004

Editors: Kai-Uwe Kühnberger
Peter König
Petra Ludewig

Cover design: Thorsten Hinrichs

Bachelor Thesis

# Emergence in Cognitive Science:

## *Clark's Four Proposals to the Emergentists*

Jan Scholz / jscholz@uos.de

University of Osnabrück

2004

# Contents

# Acknowledgements

I would like to thank all those phenomena which are irreducible or unexplainable for being a constant source of wonder and surprise. I would like to acknowledge in particular Achim Stephan's rewarding introduction into the topic and his support. I am more than thankful for Pim Haselager's support. I gratefully acknowledge interesting discussions about emergence and anything else with John Bickle and Charbel Niño El-Hani. Last but not least I would like to thank Jacqueline Griego for brushing up my English.

# Introduction

Recently the notion of *emergence* became of interest again. In the last couple of years several conferences[1] were held, numerous papers[2] published, and books[3] appeared on the subject, approaching it from various points of view. Apparently not only philosophers but also scientists working in the most diverse disciplines are attracted by the concept of *emergence*. However, let us start at the beginning.

Since ancient times man has wondered how the extreme diversity of things his experience confronts him with came into existence. Baffled by this Arcanum he attributed it to hard-to-prove divine existences for a long time[4]. In the beginning of the 20th century the issue was brought up by the conflicting positions of *vitalism* and *mechanism*, each of which claimed to have an universal explanation for the plurality of phenomena and their properties which existed in the world. In particular, the emergence of life was at the centre of debate.

The vitalists tried to explain the difference between living and non-living entities by the postulated existence of a supernatural and non-physical entity or force like an *entelechy* or an *élan vital*. On the contrary the mechanists declared that qualitative differences between entities are solely due to the differences in number and arrangement of the same basal set of physical elements or components.

While the mechanists advocate a reductionistic position, vitalism can be called a substance-dualistic or -pluralistic position. *Emergentism* is located on the axis between those extremes. On the one hand emergentism comprises *naturalism* in the form of *physical monism* and on the other hand it allows for the existence of genuine novel properties. However, while a property-dualistic position is acceptable for the emergentist she will oppose any substance-pluralistic positions.

In what may be called the heyday of British emergentism in the 1920s gifted philosophers and academic scientists like Alexander, Lloyd Morgan, Sellars, and Broad started serious attempts to give a philosophically sound definition of emergence. We will later see that they had different motivations and consequently put emphasis on different aspect of their theories.[5]

Before we start to examine philosophical notions of emergence, let us take a look at the intuitive concept of emergence. What could be meant by calling something "emergent"? The literature on the subject offers a wide range of opinions about the possible meaning of the predicate "emergent",

---

[1] See, e.g., V Brazilian-International Cognitive Science Congress: *Life, Robots and Emergence* (2003), São Paulo; Institut Jean Nicod, *Reduction and Emergence*, (2003), Paris.
[2] See, e.g., references.
[3] See, e.g., Johnson (2001), Holland (1998).
[4] First ideas of emergence might have surfaced already in AD 129. Kim (1999, fn. 3) points out that "Galen had a clear statement of the distinction between emergent and nonemergent properties of wholes". (On the Elements according to Hippocrates, 1.3, 70.15-74.23)
[5] For excellent historical accounts of the concept of emergence see, e.g., Stephan (1999, part II) or McLaughlin (1992).

from its ordinary sense "coming into being" to precise definitions spanning several pages. Frequently definitions such as the following can be found.

*Novel structures, patterns or properties resulting from the interaction of a complex system's multiple components are called* emergent.[6]

The adjective "novel" is often substituted or supplemented by "unexpected", "nontrivial", or "unpredictable". In addition to the above-mentioned entities, "events", "laws", and "behaviors" are nominated as possible candidates for emergence.

Major differences in definition and considerable confusion seem to result from two important points.

- What can be emergent?
- What are the conditions for something to be called emergent?

The first point should not be too difficult to clarify. A given entity, for example a structure, is emergent, if it instantiates emergent properties or has emergent dispositions. The same holds for more abstract entities like events or behaviors. Therefore it is in general more appropriate to speak of emergent *properties* or *dispositions*. However this point is often ignored, so it should be remembered that an expression like "emergent structure" is a short form for the fact that the structure has emergent properties or dispositions.

The second question is harder to answer and lies at the heart of the emergentist's struggle. The predicate "novel" is at best ambiguous. It has at least two fundamentally different meanings: *diachronic* and *synchronic*. If something has not existed before and suddenly comes into existence at a certain point in time it is diachronically new. On the other hand synchronic novelty is time independent. This concept focuses on the relationship between a system's properties and the arrangement and the properties of the system's parts. We will later see that this distinction led to the differentiation between synchronic and diachronic theories of emergentism.

It is the right choice of conditions that have to be fulfilled for something to be emergent that defines the efficacy of a particular kind of emergence. Depending on the "strength" of the preferred concept, different kinds of phenomena and different numbers of phenomena fall under a particular notion of emergence, or as Clark (2001, p. 113) put it, each particular brand of emergence cuts the "emergent/ nonemergent cake" somewhat differently.

Traditionally philosophers advocate a stronger notion of emergence. In particular the *Philosophy of Mind* regards the concept of emergence as a handy tool for examining the mind-body problem. *Intentionality* and especially *phenomenal qualities* are the last remaining subjects of the investigation of (strong) emergence. The *natural sciences* typically work with weaker notions of emergence. Consequently the possible range of application is far more extended. The notion of emergence is

[6] See, e.g. `http://de.wikipedia.org/wiki/Emergenz`.

used in the description or explanation of phenomena such as quantum entangled states in quantum physics[7], molecular structures in chemistry[8], morphogenesis and evolution in biology[9], development of communities in the social sciences and price development on free markets in economics[10].

In the following I will focus on *cognitive science*, which features a wide range of phenomena like, for example, connectionist networks, artificial intelligence, artificial life, robots, and autonomous agents. In his book *Mindware* (2001) Clark proposed four classes of emergent phenomena. On the basis of his suggestions I will investigate the use of the notion of emergence in cognitive science. It will quickly become apparent that those classes are not exclusive and in fact are part of a bigger class of phenomena, namely complex dynamical systems.

This work comprises two major parts. In the first part I introduce the philosophical concepts of emergence according to Stephan (1999) thereby laying the foundations for the second part. In the beginning of the second part basic knowledge of general system theory and the theory of self-organization are presented. This will later be helpful to analyze the phenomena that Clark uses to illustrate his notions of emergence. Furthermore Clark's notions of emergence will be discussed in detail as well as the general theme behind them.

The purpose of this work is twofold. First, I want to make the philosopher and the (cognitive) scientist acquainted with the other's ideas of emergence. The concepts put forward by both sides shall be presented and analyzed for differences and similarities. Second, it shall be discussed if those differences can be overcome. If this is not possible, I will point out the reasons for the irreconcilability of both positions. In the end it will be shown whether there is legitimate need for several concepts of emergence.

[7] See, e.g., Hüttemann (2000, section 3).
[8] See, e.g., Müller and Kögerler (1999, p. 103).
[9] For treatments that focus especially on the issue of emergence in connection with biology, see works of Niño El-Hani, e.g. El-Hani (2000).
[10] See, e.g., Bonabeau (2002).

# Emergentism in Philosophy of Mind

## The Layered Model and Varieties of Emergentism[11]

The concept of emergentism is closely related to a multilayered model[12] of the world that views the world as stratified into different levels organized in a hierarchical structure: elementary particles at its lowest levels and molecules, cells, and multicellular organisms at progressively higher levels. According to the multilayer model all entities belonging to a given level, except the most elementary, can be exhaustively decomposed into entities belonging to lower levels. Thus each entity can be arranged in the hierarchy according to its mereological[13] relation to other entities.

However this model raises more questions than it can possibly answer. Is there a bottom or top level? How are entities at different levels related to each other? How are characteristic properties of a given level, i.e. properties that make their first appearance at a specific level, related to properties at the adjacent levels?

Emergentists are particularly interested in the latter question. While they agree with reductionists that higher-level properties somehow depend on or are determined by lower-level properties[14], they oppose the view that higher-level properties must necessarily be deducible from lower-level properties.

It should be noted that especially in discussions of the mind-body problem, regarding the relation of mental properties to neural or physical properties, the question of whether properties of a whole can be taken to be ontologically independent of the properties of the parts plays a prominent role. As a side note let me point out that the relation between higher- and lower-level properties is the subject of various philosophical concepts that are not necessarily all in competition with emergence such as *supervenience*[15], *realization*, and *reduction*.

The notion of emergence as it is put forward by Broad or Kim is an *ontological* concept and directly addresses the question whether certain properties of a system can be deduced *in principle* from the properties of the system's components and their organization. The absolute and metaphysical character of this concept should be stressed once again.

> "Assuming emergence to be an ontological notion implies *not* to accept as examples of emergence cases that depend on the *epistemological inaccessibility* of explanations to human beings." (Hüttemann 2000, section 2; my italics)

---

[11] "Varieties of Emergentism" is a title "borrowed" from Stephan (1999a, p. 49).
[12] See, e.g., Kim (1998, p. 15ff), Kim (1999).
[13] part-whole
[14] "So it's unsurprising that supervenience qua mere necessary covariation fails to be sufficient for emergence." (Marras 2003, p. 1)
[15] Supervenience: *B*-properties supervene on *A*-properties if no two possible situations are identical with respect to their *A*-properties while differing in their *B*-properties.

The notion of emergence is a hierarchical concept. According to Stephan (Stephan 1999), there are three theories among the different varieties of emergentism deserving particular interest: *weak* emergentism, *synchronic* and *diachronic* emergentism. Weak emergentism is what could be called the smallest common denominator of all plausible theories of emergence. It defines the base for stronger theories of emergence. It appeals through the fact, that it is still compatible with property reductionism. Stronger versions of emergentism add further conditions to weak emergentism and might no longer provide this merit.

At the heart of *synchronic* emergentism lies the non-temporal relationship between a system's properties and the arrangement and the properties of the system's parts. According to this theory a system's property is emergent if it is *irreducible*. This is the case if the system's property is not reducible to, nor deducible from, the properties and the relations of the system's parts.

In contrast, the theory of *diachronic* emergentism is based on the *unpredictability* of novel properties. Consequently a system's property is emergent if it is impossible to predict its instantiation before its first appearance. It should be noted that synchronic and diachronic theories of emergentism are not independent. Because irreducible properties are necessarily unpredictable, synchronic emergence necessarily implies diachronic emergence. In the following, weak emergentism, synchronic emergentism and diachronic emergentism will be discussed in detail.

## Weak Emergentism

Emergentism is based on *naturalism* – the thesis that explanations making use of spiritual or supernatural entities are not valid. Thus the first feature of weak, and therefore any stronger version of emergentism, is the thesis of *physical monism*. This thesis delimits the nature of systems that can instantiate emergent properties by allowing only systems that consist solely of material parts. Therefore any substance-dualistic positions are rejected.

In particular, the thesis of physical monism excludes any *vitalistic* positions or compound theories, which claim that the emergence of a property is due to the presence of some supernatural entity. It therefore denies the possibility that non-physical entities or forces could be responsible for the fact that a given system has certain properties.

That means for example, that living as well as non-living systems consist of the same physical components. It would not be admissible to state that there are *specific components* which alone determine the fact that a certain system is alive. Both systems comprise only material parts that are arranged in specific constellations.

> *Physical monism.* Entities existing or coming into being in the universe consist solely of material parts. Likewise, properties, dispositions, behaviours, or structures classified as emergent are instantiated by systems consisting exclusively of physical parts. (Stephan 1999a, p. 50)

This weak position of naturalism should not be confused with *physical reductionism*. In contrast to strong naturalism – the thesis that all higher-order properties can be completely reduced to physical properties – weak naturalism requires only the supervenience of higher-order properties over physical properties.

The second feature of weak emergentism is the thesis of *systemic properties*, which delimits the type of properties that are possibly emergent. The notion of collective or systemic properties is based on the differentiation between the system and its elements. According to this distinction properties can be assigned to systems and/or elements. Properties that can be instantiated by both the system and its elements are called *hereditary* or *resultant* properties. Examples of hereditary properties are basic physical properties such as spatial extension, mass and velocity. On the other hand there are properties that are instantiated by the system, but by none of its components. Those properties are called *systemic* or *collective* properties. Examples of systemic properties include properties such as living, breathing, bipedal locomotion and being angry.

> *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. (Stephan 1999a, p. 50)

The notion of systemic properties is thought to be uncontroversial. Countless examples dispel any doubt.

The third and last feature of weak emergentism is the thesis of *synchronic determination*, which specifies the type of relationship that holds between the properties of the system's parts and their arrangement on one side and the emergent properties on the other. It states the *nomological* dependence of a system's properties and dispositions on its *microstructure*[16].

> *Synchronic determination.* A system's properties and dispositions to behave depend nomologically on its microstructure, 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. (Stephan 1999a, p. 50)

Anyone who denies the thesis that a system's properties are synchronically determined has basically two choices. Either one admits properties that are not bound to the properties and arrangements of the system's parts or one allows for the existence of systems that are identical in their microstructure but have different properties or dispositions, a peculiarity, which can then only be explained by non-natural factors. Both alternatives result in a non-naturalistic position.

In summary, weak emergentism comprises three theses. The first thesis restricts the nature of a systems that could possibly instantiate emergent properties. The second thesis limits the type of

[16] A system's microstructure is defined by the properties of the system's elements and their arrangement: The microstructure $\langle c_1, c_2, \ldots, c_i \mid o \rangle$ of system $S$ is the set of the system's components $c_1, c_2, \ldots, c_i$ and their arrangement $o$.

properties that might be emergent, and the third thesis specifies the type of relationship that holds between the system's microstructure and its emergent properties.

These are the base conditions for all following theories of emergentism. However it is also a "theory of its own right" (Stephan 1999a, p. 51), that is held by some philosophers as well as cognitive scientist[17], who do not want to give up on reductionism because of emergentism.

## Synchronic Emergentism

To obtain stronger versions of emergentism, we have to add more conditions. There are two major possibilities: *unpredictability* and *irreducibility*. Diachronic emergentism can be obtained by adding the thesis of unpredictability to weak emergentism, and synchronic emergentism by adding the thesis of irreducibility (or non-deducibility) to weak emergentism.

Broad's classical explication of a strong version of emergentism – synchronic emergentism – is based on the latter strategy.

> 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*). (Broad 1925, p. 61)

According to Broad a systemic property that depends nomologically on the system's microstructure, 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 (simpler) systems. (Stephan 1999a, p. 51) A closer look reveals that Broad's statement includes two possibilities for a property to be irreducible, each depending on the violation of a different criterion of reducibility.

A property is called *reducible* if it fulfils two conditions: (i) from the behavior of the system's parts alone (in the actual system) it must follow that the system has some designated property, and, (ii) 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. (Stephan 1999a, p. 51)

Both conditions can be independently violated and consequently yield two different kinds of irreducibility: (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*, or (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*. (Stephan 1999a, p. 52)

[17] Stephan (1990, p. 51) mentions philosophers Bunge and Vollmer and cognitive scientist Hopfield, Rosch, Varela, and Rumelhart.

What are the implications of the latter version of irreducibility? It says that it would be impossible to determine the behavior of the system in question from the behavior of the system's parts in isolation or in other (simpler) systems. As long as that is the case, the system's behavior has to be taken as a brute fact. And that is exactly the situation scientists employing the mechanistic explanation strategy and dealing with complex systems are confronted with every day. They try to explain a complex system's property on basis of the properties of its isolated components and general laws of combination and interaction. Hüttemann advocates the view that since the mechanistic explanation strategy is deeply entrenched in scientific practice, accepting a system's behavior as emergent in the above sense is "tantamount to the admission that the mechanistic explanation strategy has its limits". He therefore considers such emergent systems as *anomalies* and suggests instead of giving up the methodology it is always an option to "modify either the descriptions of the components or the laws of composition or the laws of interaction" (Hüttemann 2000, section 6).

However, if that enterprise had to fail *in principle* in certain cases, the system in question would be irreducible according to case (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 component's behavior in isolation or in other (simpler) constellations. (Stephan 1999a, p. 52)

The first condition of reducibility is already violated if it is impossible in principle to characterize a systemic property by neither the microscopic nor the macroscopic behavior of the system's parts. Candidates for irreducibility due to *unanalyzability* are properties that are not definable by their causal role like phenomenal qualities[18].

> *Unanalyzability.* Systemic properties which are not behaviorally analyzable – be it micro- or macroscopically – (are necessarily) irreducible. (Stephan 1999a, p. 52)

The two different types of irreducibility of systemic properties entail likewise two different consequences. Irreducibility due to irreducibility of the components' behavior implies *downward causation*[19]. "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." (Stephan 1999a, p. 53) It should be noted that this would not necessarily violate the principle of the causal closure of the physical domain. We just had to accept additional causal influences.

If however a system's property were irreducible because of its unanalyzability, we had to ask if it had any causal role at all. A property is behaviorally unanalyzable if it cannot be characterized by

[18] See section "Are Phenomenal Qualities Synchronically Emergent" on p. 11.
[19] See, e.g., Kim (1992).

its causal role. If we do not want to deprive unanalyzable properties of any existence at all[20], it seems that the status of those properties can only be the one of epiphenomena[21]. Both consequences represent the two horns of what Stephan (1999, ch. 16) coined the *Pepper-Kim-dilemma* – the problem of whether theories of emergentism allow mental properties to have causal efficacy and at the same time leave the causal closure of the physical domain intact.

## Diachronic Emergentism

As noted before, Broad favored the synchronic version of emergentism. He was interested in the characteristic differences of systemic properties independent from their temporal occurrences. The notion of irreducibility seems to be an adequate tool to capture this idea, because it can account for such difference even in non-evolutive universes.

On the other hand, Alexander and Llyod Morgan found temporality to be an important feature of their concept of emergence. Within the scope of evolutionary theories of emergence, they were less interested in reducibility and asked instead if genuine novel properties could be *predicted* in principle before their first occurrence.

The notions of *novelty* and *unpredictability* will prove to be decisive features of theories of diachronic emergentism. In the following we will encounter two varieties of diachronic emergentism: *weak diachronic emergentism* and *diachronic structure emergentism.*

To obtain weak diachronic emergentism we have to extend weak emergentism by the thesis of *novelty.*

> *Novelty.* In the course of evolution exemplification 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. (Stephan 1999a, p. 53)

However, this theory is still too weak, i.e. too similar to "pure" weak emergentism, to be of any theoretical interest. The further addition of the thesis of *unpredictability* yields a stronger notion of emergence.

Systemic properties can be unpredictable in principle for two entirely different reasons: (i) if the micro-structure of the system that exemplifies the property for the first time in evolution, is unpredictable or (ii) if the systemic property is itself irreducible. (Stephan 1999a, p. 53)

If a systemic property is irreducible it is ipso facto unpredictable before its first instantiation, i.e. unpredictability is necessarily implied by irreducibility. Therefore the notion of unpredictability in the sense of the second criteria offers no real advantages beyond the plain notion of irreducibility.

---

[20] According to *Alexander's Dictum* "to be real is to have causal powers" only those things exist that have causal efficacy (c.f. Kim 1992, p. 134ff).

[21] Epiphenomena are phenomena that are secondary effects which cannot themselves cause anything.

The first case – *unpredictability of structure* – is much more promising, especially concerning dynamical systems and chaotic processes. What could be the reasons for a structure to be unpredictable? First, if the universe were *indeterministic*, it follows that there could be unpredictable structures. To preclude such trivial causes of unpredictability, most emergentists claim that the development of new structures is governed by deterministic laws.

Second, as Stephan (1999, p. 56) suggests, the emerging of novel structures could be unpredictable in principle if their formation is governed by laws of deterministic chaos. As we will later see[22], functions that exhibit chaotic behavior are essentially unperiodic in their chaotic regimes. In addition the outcome of those functions depends critically on their initial values, i.e. small differences of initial values can later on lead to radically different developments.

The ability to predict or simulate the behavior of chaotic functions depends on three prerequisites. First, the exact *initial conditions* have to be known. This is however impossible in empirical systems where one would have to measure for an infinite time span to detect low-contrast information in grainy data[23]. Second and third the calculations required for the prediction of future states have certain *space* and *time complexities*, i.e. the computer has to be fast enough and supplied with enough memory.

The fourth problem is first of all to determine by which laws the phenomenon in question is governed and how it can be formalized. In the following we have to be careful to distinguish between what is impossible in principle and what is impossible due to the circumstances (e.g. epistemic or computational limitations). However these two situations are not as dissimilar as often supposed. In fact they coincide ultimately at the point where the necessary calculations have a spatial complexity that exceeds the number of particles in the universe.[24]

Early emergentists devised thought experiments wherein they endowed hypothetical creatures with special talents. According to Alexander[25] the so called Laplace's calculator is a creature that knows several earlier states of the world and all the natural laws that govern changes in the world. Stephan takes this knowledge as the maximum of what could be known in principle and concludes that as chaotic processes are aperiodic, it would be impossible to determine from a finite number of states the exact formulas, which govern the development of these processes[26]. "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 they are governed." (Stephan 1999a, p. 54)

---

[22] See section on "Complex Dynamical Systems".

[23] Note also the constraints imposed by Heisenberg's "uncertainty principle": If *a* and *b* are canonically conjugate variables then the more precisely the value of *a* is determined, the less precisely value of *b* can be determined, and vice versa.

[24] The same holds, of course, for temporal complexity. Calculations become impossible in principle if they take more time than is left in the universe until it collapses or becomes absorbed in total thermal equilibrium.

[25] See, Alexander (1920, ii, pp72f, p328) cited in Stephan (1999a, p. 54).

[26] In general, finite measurements can never demonstrate characteristics which mathematically demand infinite precision.

> *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. (Stephan 1999a, p. 54)

In summary, a systemic property is unpredictable in principle before its first instantiation, if it is irreducible, or if the structure by which it is instantiated is unpredictable in principle before its first formation. It should be noted that the second case allows for unpredictable properties that are not necessarily irreducible. It follows that diachronic structure emergentism is compatible with reductive physicalism.

## Are Phenomenal Qualities Synchronically Emergent?

The typical argument put forward against the existence of strongly emergent properties is based on the trust in the future advancements of science. It says that even if a given property is currently *not* reducible it is still possible that it might be reducible in future, i.e. the property's emergence is relative to the current body of knowledge. This line of reasoning fails to appreciate what kind of concept the emergentists are after. What they want to determine is a property's *in-principal* irreducibility.

Whether or not a behavioral analysis can be found for a previously behaviorally unanalyzable property, or if such an analysis is impossible in principle, is a question that proved to be notoriously difficult to settle.

In the following I will give a quick introduction into the problem of phenomenal qualities. It will point us into the direction where strongly emergent phenomena might be located before we turn to the cognitive science, a field of knowledge where the weak kind of emergentism is favored.

Qualia, the terminus technicus for phenomenal qualities, have until now escaped any precise physical or causal definition. In a standard broad sense of the term qualia refer to those subjective qualities of experience that accompany so many perceptual acts, e.g. the taste of salt, the scent of a rose or the sensation of pain. There is something it is *like* to have a specific experience – a qualitative *feel* – that distinguishes it from any other experience. In short, qualia refer to the introspectively accessible, phenomenal aspects of our mental lives.

Qualia therefore reveal an *epistemic asymmetry* in our knowledge of the mental. While some mental properties such as memory and learning can mainly be characterized objectively by their causal roles, our knowledge that phenomenal qualities exist derives primarily from our own case. The former are called *psychological* properties, properties that can be defined by their causal roles, i.e. functionally analyzed. The latter are called *phenomenal* properties, properties that evade functional analysis.

How can a property evade functional analysis? To reductively explain a phenomenon one has to specify its causal role, and then show how certain lower-level (e.g. physical) mechanisms fulfill this

role[27]. However, no matter what functional account of perception is given, there will always be a further question: Why is this kind of perceptual state accompanied by this qualitative feel? It follows that phenomenal states are not sufficiently characterizable by their causal role. As Levine (1983) put it, there exists an *explanatory gap* between the mental and the physical.

Systemic properties, which are not behaviorally analyzable like phenomenal qualities, are according to Broad irreducible. In this case no progress in science will ever close the gap between the physical and the mental. In other words: Phenomenal qualities are synchronically emergent. This implication is however disputed.

> "[The] (true) claim that the concept of pain is not a functional concept, or that there is no functional analysis of pain is [..] irrelevant to the question of its reducibility." "[..] if we insist that irreducibility is part of the very definition of an emergent property [..] then we should say that there are no emergent properties." (Marras 2003, p. 10ff)

In summary, it can be said that the status of synchronically emergent phenomena is a hotly debated issue. As mentioned in the previous chapter irreducibility of components' behavior as a criterion for emergence runs contrary to common scientific practice and is therefore fiercely rejected (by those following this practice). In addition the relation between unanalyzability and reducibility is controversial. On the other hand the "gap" still persists. It seems that we have to wait for either side – be it a philosopher or a scientist – to come up with an argument that has the potential to settle the question of the existence of synchronically emergent phenomena.

[27] This is of course a simplified account. According to Kim's model of functional reduction *properties* are reduced (Kim 1998, p. 97ff). According to its akin predecessor – the Nagelian model of reduction (Nagel 1961, ch. 11) – *theories* are reduced.

# Emergentism in Cognitive Science

## Emergence Outside Philosophy

In the first part of this thesis I explicated the basic features of theories of emergentism as they are put forward by philosophy. While synchronic emergentism in particular seems to be a useful tool with which to tackle problems of phenomenal qualities, "non-philosophical" disciplines such as artificial intelligence, artificial life, and theories of complex dynamical systems and self-organization are mainly associated with versions of weak emergentism. However, the two extant types of emergence – weak and stronger concepts of emergentism – have both been used in cognitive science and consequently been mixed up. As Walmsley (2003, p. 10ff) points out "[..] we have a situation where two sides are simply talking past each other. Those interested in the [..] scientific import of emergence (à la Clark) and those interested in the metaphysical aspect of emergence are simply talking about *different things*." I will later come back to this point.

What role could emergentism play in a field of knowledge where by common consent everything is reducible to, and explainable in terms of lower level properties, i.e. physics? For example neurobiologists are already investigating the role that single molecules and ions play for behavior and cognition[28].

What then is the job of emergentism outside philosophy? Should emergentism point out the limits of science, beyond which the reductionists' project will fail in principle or should it guide our scientific endeavors? Are emergent properties a brute fact to be taken with natural piety or an incentive to develop new explanatory frameworks? Should emergentism be of metaphysical or of methodological importance?

In order not to get lost in the wide range of slightly different concepts of emergence and the multitude of possibly emergent phenomena that are put forward in cognitive science, I will focus on the four notions of emergence and the phenomena that fall under them as they were suggested by Andy Clark (2001, ch. 6): emergence as *collective self-organization*, emergence as *unprogrammed functionality*, emergence as *interactive complexity*, and emergence as *uncompressible unfolding*.

It will later become evident that those classes are not wholly disjunctive. In fact they are all related to a larger class, namely *complex dynamical systems*. Two suggestions – collective self-organization and interactive complexity – are closely related to *theories of self-organization*, which fall under *dynamical system theory*.

It will therefore be necessary to give a short introduction into General System Theory, with emphasize on complex dynamical systems, supplemented by an overview of theories of self-organization and synergetics.

[28] See, John Bickle (2003) for a "ruthlessly reductive account" of neuroscience.

## Complex Dynamical Systems in General System Theory

Almost all phenomena that are of interest in emergentism are concerned with complex systems and their properties. While the human body is one of the most complex systems we know, we will later see that even much simpler systems can behave in unexpected and hard-to-explain ways. Among those systems are robots, artificial agents, and ant colonies.

I will yield to the "natural affinity between dynamics and emergence" and treat those systems no matter how diverse they may appear at first glance as complex dynamical systems. "After all, the language of dynamical systems theory seems to be ideal for describing and explaining emergent properties." (Walmsley 2003, p. 1) General System Theory will not only provide the basic concepts to describe complex dynamical systems formally, but it is itself a bridge between philosophy and the natural sciences[29].

General System Theory[30] is based on the notion of a system. Intuitively a system is a set of interrelated elements comprising a unified whole. A system typically consists of components or elements, which are connected together in order to facilitate the flow of information, matter or energy. If a system is a component of another system, it is called a subsystem.

According to the analytic notion of a system, the system's boundaries are no intrinsic property of the system. Pragmatics and context govern the boundaries chosen in practice by the external observer. Typically those boundaries are drawn so that the interactions between components of the system are maximized and the interactions of the system with its environment are minimized. However by doing so one is liable to neglect the influence of the system's context[31]. Systems that are explicitly considered as isolated from their environment are called *closed* systems. Because it is in general impossible to completely isolate a system from its environment most systems are in fact *open* systems – systems that exchange matter, energy or information with their environment.

There are two classes of systems: systems that are *far* from thermal equilibrium and systems that are *near* or in thermal equilibrium. As we will later see, only systems which are far from thermal equilibrium, namely *dissipative* systems, exhibit interesting behavior.

There are several reasons to call systems complex: number of elements, variety of elements, complexity of elements, and complexity of interaction (between elements). The complex self-organizing systems we will later encounter usually consist of a large number of homogeneous and simple[32] elements that interact in a complicated fashion. Usually those interactions are nonlinear (see below), circular, and of short range. The emergence of systemic properties is often attributed to the complex nature of the system's elements interactions.

---

[29] See von Weizäcker (1987).
[30] For a classic introduction see, e.g. von Bertalanffy (1969).
[31] C.f. "contextual simplification" in Wimsatt (1997).
[32] in relation to the whole system

Systems can be defined more formally as sets of interdependent variables[33]. A *variable* is an entity that can change, i.e., be in different states at different times. If the change of a variable depends on others, the variables are said to be *interdependent*. The *state* of the system is simply the state or value of all its variables at a time integrated into a *state vector*.

$$\boldsymbol{x} = (x_1, \ldots, x_n)$$

Systems that change over time are called *dynamical* systems. They can be modeled by a set of (a) *differential* or a set of (b) *difference* equations. The former is used if time is considered to be continuous the later if time is modeled in discrete steps, as in numerical computer simulations.

$$d\boldsymbol{x}/dt = f(\boldsymbol{x}) \quad \text{(a)}$$

$$\boldsymbol{x}_{t+1} = f(\boldsymbol{x}_t) \quad \text{(b)}$$

*Linear* differential[34] equations can be solved easily. To solve *nonlinear* differential equations is however very difficult if not impossible and can only be accomplished in special cases or approximately. The fact that even numerical procedures are usually confronted with huge problems depends on special features of dynamical systems, which will now be discussed in detail.

Dynamical systems reside in so called *phase space*, which includes any possible state of the dynamical system and has therefore the same dimensionality as the system's state vector. The set of all states a system travels through during a well defined period of time is called a *trajectory* or the system's *behavior*. A trajectory represents part of the system's history and can be visualized in a *phase portrait*.

A characteristic feature of dynamical systems is the so called *attractor*. Intuitively attractors are trajectories in phase space that attract neighboring trajectories. Trajectories that enter the *basin of attraction* of a given attractor have no chance but to converge towards that attractor. There are basically three classes of attractors: *fixed point attractors, periodic attractors*, and *chaotic attractors*.

Systems that evolve towards a single stable state, like a damped pendulum, exhibit fixed point dynamics. Their final states are usually states of minimal energy. If the system is slightly disturbed it will return to its stable state after some time.

Systems that cycle through a loop of states are in a *limit cycle* or periodic attractor. If for example a wall clock is slightly disturbed it will after a period of time return to its stable phase of oscillation. Such systems can somehow paradoxically be describes as being in a stationary state of ongoing activity.

---

[33] "We must therefore distinguish objects (parts of the world such as the sun and planets, Macintoshes, and cognitive agents) from the systems they instantiate. Any given object will usually instantiate a great many systems of different kinds." (van Gelder 1998, section 3.1)

[34] In the following I will confine myself to continuous dynamical systems. However most of what is said should be applicable to discrete dynamical systems, too.

Chaotic attractors consist, in contrast, to the first two classes of *infinitely* many attractors that are arranged in a fractal pattern. Systems that can be described by chaotic attractors display *chaos* – which is defined to be an "*aperiodic bounded* dynamics in a *deterministic* system with *sensitive dependence to initial conditions*" (Kaplan 1995, p. 27). *Aperiodicity* means that the system never visits the same state twice. Computers with finite precision must eventually return to some former state, which makes it difficult to model chaotic systems numerically. *Bounded* means that the system, however erratic its dynamics may be, will not leave a certain region of phase space. *Deterministic* means that if the exact initial conditions are known any future state can theoretically be calculated. *Sensitivity on initial conditions* is an essential aspect of chaos. It means that two states that are initially close will drift apart as time proceeds.[35]

Aperiodicity and sensitivity to initial conditions are responsible for the fact that chaotic systems are *effectively* (long-term) unpredictable, although they are deterministic. Limited empirical methods and bounded computational resources make it impossible to model a given system's dynamics with infinite precision. Especially in the case of chaotic dynamics, this eventually leads to large deviations between the simulation of a system and its original development. In fact any degree of error (in measurement or computation) no matter how small will grow exponentially[36]. Greater precision will improve prediction of future states, but in chaotic systems, precise and accurate predictions about the state of the system in the arbitrarily distant future can in general not be made.

As noted above dynamical systems can exhibit a wide variety of different dynamics. A single dynamical system is not confined to one type of behavior, but can display various types of qualitative behavior for different values of its *control parameters*[37]. The point where the system changes from one form of qualitative behavior to another is called a *bifurcation*. A *bifurcation diagram* shows the possible values a dynamical system's function can have as a function of the system's control parameters. The well-known *pitchfork* or *period-doubling* bifurcations[38] are classic examples where a previous stable state becomes unstable, and a transition to bistability occurs.

The natural affinity between dynamical system theory and emergentism is probably due to the concept of bifurcation that allows dealing with systems not only quantitatively, but also qualitatively.

One might now ask whether the qualitatively new dynamics beyond a bifurcation are emergent. We can easily see that the behavior of real-life systems like cardiac tissue that can display periodic, quasi-periodic, and chaotic dynamics are at least weakly emergent. If the behavior is governed by laws of deterministic chaos, it is unpredictable in principle due to the inaccessibility of exact initial

[35] See, e.g., Kantz (1999, p. 76).
[36] Two neighbouring points deviate from each other exponentially: $\Delta x(t+1) = k\, \Delta x(t)\, e^{\lambda t}$, with Lyapunov exponent $\lambda > 0$.
[37] See section on "Self-Organization and Synergetics".
[38] See appendix, figure 1.

conditions and the behavioral laws as shown above. In that case we can speak of diachronic structure emergentism.[39] However, what about the points of transition, can they be predicted? It is possible to build a model that behaves according to the characteristics of cardiac tissue (Glass 1984). Although this model is an extreme simplification, it predicts the stimulus frequency that is necessary to elicit chaotic behavior in the tissue.

The precision of the model's predictions depends of course on how well the set of equations mirrors the empirical system's dynamics and on how precise the models parameters can be determined. Given the simplicity of the model, its predictions are fairly accurate, a fact that is often emphasized in dynamical system theory. We will later see that complex systems with a large numbers of elements can be described by only a few simple parameters that seem to be relatively independent of the system's physical substrate.

But let us return to the question whether bifurcations can be predicted. The bifurcation diagram of the logistic map beautifully shows that the parameter intervals between successive period-doubling bifurcations rapidly decrease.[40] Algebraically it is only possible to calculate the first few period-doubling bifurcations because it involves the solution of high-degree polynomial equations[41]. Thereafter only numerical approximations can be used. Therefore the *exact* determination of the onset of chaos or any higher-order bifurcation is impossible in principle.

It follows that it is impossible in principle to predict higher-order bifurcations or the onset of chaos in systems that have similar dynamics as the logistic map, i.e. period-doubling cascades with infinite precision. However often fairly accurate approximations are possible.

In summary, dynamical systems that evolve according to laws of deterministic chaos provide good examples for diachronic structure emergentism. The exact determination of initial conditions and the laws that govern the formation of a system's structure provide in-principle obstacles for predictions of empirical systems. Computational limitations can even make predictions of abstract systems impossible in principle. However, chaos theory has made it possible to at least short-term predict what before seemed only random and allows to determine the error bonds of those predictions.

---

[39] More accurately: The formation of the system's structure is governed by laws of deterministic chaos. Therefore the system's structure is unpredictable in principle before its first formation, which implies the unpredictability of properties (resp. the behaviour) instantiated by the structure.

[40] See appendix, figure 2.

[41] Abel's Impossibility Theorem: In general, polynomial equations higher than fourth degree are incapable of algebraic solution in terms of a finite number of additions, subtractions, multiplications, divisions, and root extractions.

## Self-Organization and Synergetics[42]

Theories of complex dynamical systems are closely related to the theory of self-organization – in fact all self-organizing systems are dynamical systems. Intuitively self-organization can be defined as a process in which the internal organization of a system increases automatically without specific interference from the system's environment. A system that is self-organizing establishes or maintains its order by virtue of the interactions of its own elements.

Phenomena of self-organization can be found in various disciplines, such as physics, chemistry, biology, social science, and economics. For example magnetization, crystallization, morphogenesis, political views and stock markets can all be described by laws of self-organization[43]. Attempts to simulate these phenomena revealed that their dynamics share many common characteristics.

The introduction of the concept of self-organization in this work has two reasons. First, the theory of self-organization will help us to understand and examine the emergent phenomena that fall under Clark's concepts of emergence, and second, the concept itself is often associated with emergence.

Bénard rolls are a much cited example for self-organizing systems. A liquid is evenly heated from below, while cooling down evenly at the surface. For small temperature differences, heat is exchanged by mere diffusion. If the temperature difference is increased heated liquid starts to move towards the surface, since it has a lower density. At the same time cooler liquid at the surface begins to sink to the bottom. Since these two movements cannot take place at the same time and in the same volume the flows in the liquid coordinate and form rolls or hexagonal patterns.

There are several fundamental traits that distinguish self-organizing systems like the one mentioned above from more traditional mechanical systems[44].

Self-organizing systems display global patterns of organization. Those ordered spatio-temporal patterns arise spontaneously as the result of large numbers of interacting elements[45]. The nature of the interaction is *nonlinear* and *local*.

Nonlinearity is best explained by the *feedback* relations that hold between elements. Each element affects other components, which again affect the first element. The nature of this circular cause-and-effect relation can be *positive* or *negative*. If the first element's change is amplified when it returns via its effects on other components, feedback is said to be positive. Positive feedback makes deviations grow in a runaway, explosive manner. It leads to accelerated development. Feedback is

---

[42] *Synergetics* is an interdisciplinary field dealing with principles of self-organized pattern formation in non-equilibrium systems. Haken (1988, p. 23) considers synergetics "as a theory of the emergence of new qualities at a macroscopic level". For our purposes synergetics can be considered to be a theory of self-organization.

[43] See, e.g. Haken (1977).

[44] See, e.g. Kelso (1995, p. 16ff), Heylighen (1999, p. 5ff).

[45] There are several theories about the emergence of order in self-organizing systems, such as "order from noise" (von Foerster) and "order through fluctuations" (Prigogine).

said to be negative, if the other components reactions are opposite to the initial change. Negative feedback stabilizes the system by suppressing or counteracting change. Feedback relations allow a system to adapt to a changing environment.[46]

*Local interactions* can lead to *global order*. In a disordered system any influence that propagates through the system is quickly dispersed by random fluctuations of its elements, i.e. remote parts of the system that are basically independent and uncorrelated. In an ordered state, however, the system's parts are coupled and therefore strongly correlated. If the system is in a stable state random changes of single elements are quickly counteracted by the "system", i.e. the single element's neighbors that are again controlled by their neighbors and so on.

Self-organizing systems are *robust* and *resilient*. By virtue of their redundant and distributed organization, self-organizing systems have the capacity to restore themselves after perturbations or errors. Unaffected parts of the system can usually make up for the affected ones. Another reason for this inherent robustness is that fluctuations are an intrinsic property of self-organizing systems. They allow the system to discover new stable states.

A central property of self-organizing systems is the *absence of centralized control*, i.e. the system's organization is not established by some external or internal agent. All components contribute more or less equally to the establishment and maintenance of the organization.

Self-organizing systems are *far from thermal equilibrium*. Thermal equilibrium is a state of minimal energy in which no energy or matter is dissipated. The second law of thermodynamics states that in closed systems entropy can only increase. To maintain its order a self-organizing system has therefore to dissipate entropy by exchanging energy or matter with its environment. It increases its own order at the expense of the order in the environment. Therefore self-organizing systems have to be open and dissipative systems.

Self-organizing systems can be described formally by *order parameters* and *control parameters*. *Order parameters* or collective variables are relevant degrees of freedom, characterizing the state of a self-organizing system. It is an interesting feature of self-organizing systems that their states can usually be described by few order parameters. Convection rolls can be described by their direction of rotation, lasers by their light's phase, etc. An order parameter is created by the coordination of the system's components and describes the system's macro pattern. It compresses an enormous amount of information (e.g. the individual molecules' positions and velocities in the Bénard experiment, the individual atoms' positions and phases in the laser).

[46] Heylighen (1999, p. 15) goes as far as to say that in some sense "self-organization implies adaptation". "For example, the pattern and speed of flow in the Bénard rolls will be adapted to the specific temperature difference between bottom and surface, whereas the orientation of the spins will tend to be parallel to any outside magnetic field."

I will not cite the much-quoted "enslaving principle" of synergetics, because it is easy to misinterpret. Let me however point out that to attribute causal powers over the system's elements to the "whole" of the system or to an "order parameter" would mean to commit a category mistake. Higher-level entities have no *modus operandi* by which they can act on lower-level entities.[47] For example, to state that a Bénard roll forces the liquid molecules to move in a particular direction, should be understood as a short form for saying, that the liquid molecules comprising the Bénard roll have such and such effect on particular (other) liquid molecules.

It should be noted that in general collective variables do not critically depend on a system's (physical) realization. Many systems that can be described in terms of collective variables show remarkable substance independence. Groups of artificial agents, birds or humans[48] can all display the same behaviors. It follows that the collective phenomena are the result of the organization and interactions, rather than due to the specific physical substrate.

*Control parameters* are parameters that lead the system through different states or patterns (e.g. temperature gradient in the Bénard experiment, or the energy feed to a laser). If the value of the control parameter is increased the system changes its dynamics. With increased energy influx dynamical systems often leave their stable states and start to oscillate (limit cycle dynamics). If the energy influx is increased even further the system eventually shows chaotic behavior. To put it differently, the more energy is pumped into the system, the more amplification of small differences is possible, and therefore, more variance in the types of behavior.

Unlike order parameters, control parameters are quite unspecific in nature, i.e. "in no sense do they act as a code or a prescription for the emerging patterns" (Kelso 1995, p. 16). The resulting dynamics are not encoded in the value of the control parameter and therefore cannot be predicted solely on the basis of the control parameter.

The increase of an order parameter can lead the system to a point of bifurcation beyond which several stable states exist. It is not possible to determine a priori which state will be realized in an empirical system, because this depends on small fluctuations that are amplified by positive feedback. "In practice, given the observable state of the system at the beginning of the process, the outcome is therefore *unpredictable*." (Heylighen 1999, p. 12)

In disordered systems, the individual components' states have the same probability. Because we talk here about complex systems with a large number of elements, it can be assumed that the different microscopic states cancel each other out, i.e. cannot result into a coherent macroscopic pattern. At a global level the system is homogeneous or symmetric. After self-organization, however, symmetry is lost, i.e. one configuration dominates all other. This process is called

[47] See Stephan (1999, p. 237).
[48] See, e.g. "aggressor-defender game" in Bonabeau, E., Funes, P., & Orme, B. (2003).

*symmetry breaking*. From an empirical point of view it is virtually indeterminable why the system prefers one configuration instead of its alternatives.

Let me point out the fact that self-organizing systems are frequently associated with a quite different class of emergentism – *dynamic* emergentism – which is consequently mixed up with the philosophical concepts we encountered in the first part of this thesis.

The *dynamic*[49] or *phenomenological*[50] notion of emergence is applied to previously uninstantiated systemic properties that come into existence through system-internal processes, for example by means of self-organization or a kind of evolutionary process. The "emerging" of solutions in connectionist networks or general problem solvers, or adaptation of organisms in an evolutionary setting are some of the most frequently mentioned phenomena associated with this notion of emergence. It tries to capture the intuition that the "solution" is somehow implicitly contained in the system and produced or generated when needed. This process is inherently temporal and often stepwise improvements or sequences of *subgoals* on the way towards a possible solution can be distinguished.

The distributed and parallel character of this process and the fact that it can take different routes and can result in different "final" states is taken as evidence that it is unpredictable or irreducible. In addition the process is often, as in the case of biological evolution, open-ended. Typically there are no fully adapted organisms or best solutions, but only approximations. In the context of evolutionary settings there usually arise two types of questions: questions concerning the status of the "final solution" and questions concerning the "route" to it[51]. The latter question is asked by proponents of the dynamic notion of emergence.

> "[..] I propose to look at this question [what characterizes emergent properties] not from the traditional static viewpoint but from a dynamic, evolutionary viewpoint, replacing the question 'How can a property *be* emergent?' by 'How can a property *become* emergent? (i.e. how can it *emerge*?)'." (Heylighen 1989a, p. 1)

However, Heylighen does not define a dynamic notion of emergence, he rather focuses on a certain class of systems evolving according to "generalized variation-and-selection dynamics". It turns out that his notion of emergence is really "static" and depends on the informational restrictions that he imposes on the predictor.

> "The property of 'being a solution' is *emergent*, i.e. it cannot be explained or predicted at the level where the search is carried out. It only appears when the state is actually reached [..]." (Heylighen 1989a, p. 3)

---

[49] See Heylighen (1989a).
[50] See Stephan (1999, p. 224ff).
[51] Given the initial state and learning rules, both, the evolution and final state of a connectionist network can be predicted. Stephan (1999, p. 230-231) concludes that connectionist networks do not provide an example for *structure emergence* or any stronger notion of emergence.

The "dynamic" notion of emergence put forward by Heylighen thus refers to a special class of complex dynamical systems that evolve according to some kind of evolutionary process and display properties, which somehow "emerge" over time. The dynamics of those systems are often hard to explain and predict. However, Heylighen did not succeed in working the intuitive idea of dynamic emergence into a formal and metaphysically valuable notion of emergence. Even Stephan (1999, p. 224) does not consider what he calls phenomenological emergence to be connected with any specific theory of emergence. So far more traditional "static" notions of emergence seem to be more adequate for dealing with metaphysical or ontological questions of "evolutionary" systems.[52]

## What Do Dynamical Explanations Explain?

Classical componential explanations that explain a system's functioning by detailing the individual roles and the organization of its parts often fare badly in explaining weakly emergent phenomena. As Wimsatt (2002) points out, componential explanations are best suited for "aggregate systems", i.e. systems in which components display explanatory relevant behavior even in isolation from one another.

However, "as the complexities of interaction between parts increases, the explanatory burden increasingly falls not on the parts but on their organization" (Clark 1997, p. 114). Dynamical system theory facilitates understanding of complex systems by providing a common language for system, environment and their interactions by treating system and environment as coupled systems whose mutual evolution is described by a set of interlocking equations. It thus provides a new kind of explanatory framework in which a system's behavior is explained by identifying a set of variables and their evolution over time. The distinct patterns that emerge as the system unfolds can be modeled by a set of equations and described by the mathematically precise terminology of phase space, attractors, bifurcations, etc.

It is often argued that dynamical system explanations are merely descriptions[53]. However, they owe their status as *explanations* to their ability to work as counterfactuals, i.e. they do not only inform us about the actual observed behavior of a system but also about how it will behave in various other circumstances. But can they work as *covering-law* explanations?

> According to the covering-law model of explanation, a phenomenon is explained when a statement describing it can be derived from statements specifying a law or set of laws and relevant initial conditions. Thus, an explanandum is, according to the model, a *deductive consequence* of an explanance. (Walmsley 2003, p. 3)

According to the strictest deductive-nomological model of explanation the laws cited cannot be any old laws, they must be basic laws of physics. However, dynamical explanations explain a system's behavior with reference to points and trajectories in an abstract state space; they make no reference

[52] I consider the attempt to formally define "dynamic emergence" as highly problematic and as a topic of a separate treatment.
[53] See van Gelder (1998, section 6.6.).

to the actual structure of the system whose behavior they are explaining. Because the system parameters tracked in dynamical system explanations can be arbitrarily far removed from the system's (physical) structure, dynamical system explanations only tell us about the topographical structure of the system's dynamics, but "not what it is about the way the system itself is constituted that cause those parameters to evolve in the specified fashion" (van Gelder 1991 in Clark 1997, p. 118).

It is generally supposed but often not explicitly shown that the equations cited in a dynamical explanation can be derived from the laws of physics. Walmsley argues that, nevertheless, the absence of such derivations does not rob dynamical explanations of any of their explanatory force, "since the derivation of a dynamical law from a law of physics is not *itself* part of the dynamical explanation". (Walmsley 2003, p. 5) He then concludes that dynamical explanations are a distinct species of covering law explanations where the general law under which the specific case falls is not a law of physics, but a higher-level dynamical law.

In summary dynamical explanations offer many advantages over traditional componential explanations. They provide a common language for all levels of analysis, have an intrinsic temporal focus and have counterfactual powers – not to mention their explanatory successes[54].

However, dynamical explanations are often regarded as provisional[55] or incomplete because they do not constrain the implementation of the dynamics described. According to Clark and Kelso scientific practice therefore has to resort to a mix of various explanatory tools (cf. "the tripartite scheme" in Kelso 1995).

Let me remark that the often criticized disconnectedness of the dynamical domain from the physical domain has the potential to be fittingly be described in emergent terms. It could be possible that the dynamical covering law that is featured in a dynamical explanation cannot be derived in principle from a law of physics.

> "In this case, it would turn out that any behavior of the system could be predicted on the basis of some higher-level law, but that this higher-level law could not be predicted from statements about (say) neurophysiology or physics." (Walmsley 2003, p. 12)

Emergence conceived in this sense would be *nomological*, because it is the law itself which is emergent. Unfortunately Walmsley does not elaborate on this concept and he does not give any more specific conditions for dynamic laws not to be derivable from physical laws.

---

[54] See the "empirical success argument" of van Gelder (1998, section 5). For successful dynamical models, see, e.g., Smith and Thelen (1993).

[55] Jaeger (1995, section 3.4) considers those "non-reductionistic" dynamical explanations that do not refer to the physical substrate of the empirical system as "provisional". Though they are already "real explanations" it should be possible in principal to extend those explanations with a reduction to the physical substrate according to him.

## Two Different Demands for Emergence in Cognitive Science

In the introduction I already pointed out that a difference of opinion about the "right" concept of emergence is prevalent in cognitive science. For one group the notion of emergence depends critically on either unpredictability or irreducibility for the other group the notion of emergence is primarily connected with systemic properties and (intuitively) "unexpected" or "unplanned" behavior. The tension between weak and stronger concepts arises in the literature because, "as is frequently found, dynamical explanation and emergence are referred to in the same breath, and the former is alleged to illuminate the latter." It is especially a problem for "those who [..] want to make room for both emergence and dynamical systems in cognitive science." (Walmsley 2003, p. 8) There are two potential paths of escape from the apparent tension. One can either deny that dynamical explanations are covering-law explanations or one denies that emergence is precluded by covering law explanations.

Regarding the first possibility it should have become clear from the previous section that dynamical explanations – even if they do not comply with the strictest deductive nomological model – are still of a kind that is incompatible with notions of emergence appealing to unpredictability or non-deducibility, because the final state of a system is still deducible from the set of equations together with the initial conditions. The first path is therefore blocked off, because dynamical explanations are nonetheless a form of covering-law explanation strong enough to rule out emergence *qua* non-deducibility or unpredictability.

The second path can be taken by adopting a weak concept of emergence, which is compatible with covering-law explanations. Next to the weak notion of emergence presented in the first part of this thesis, several other notions have been suggested, which are closely related to weak emergentism. Four of those suggestions, which were made by Clark, will be discussed in detail as follows.

In summary there are two different demands for emergence in cognitive science that should be distinguished. On one hand there is the intuition that there exist phenomena that are irreducible or unpredictable in principle. On the other hand there are (not necessarily irreducible or unpredictable) collective phenomena that somehow "emerge" from complex dynamical systems. The first concept is covered by stronger notions of emergentism such as synchronic emergentism and diachronic structure emergentism. The ground for the second conception is prepared by the notion of weak emergentism.

## Clark's Four Proposals

Clark addresses the notion of emergence in chapter six of his book *Mindware*[56]. The chapter's title is tellingly named "Robots and Artificial Life". Indeed, theories of artificial life are today what theories of evolution were in the last century – a playground for emergentists.

Andy Clark wishes for a notion of emergence that is neither too liberal nor too strict. A notion that is too liberal would allow too many phenomena to count as instances of emergence, i.e. it would be of little explanatory or descriptive power. On the other hand, a notion that is too strict would be in danger to "effectively rule out any phenomenon that can be given a scientific explanation" (Clark 2001, p. 113). Because Clark does not want to insist that only currently unexplainable or unpredictable phenomena should count as emergent, it should be clear that he favors a concept of emergence wholly compatible with covering-law explanations[57], a fact that makes his concepts irreconcilable with stronger notions of emergence.

Clark's concept of emergence is *classificatory*. According to Clark the concept of emergence should pick out the "*distinctive way* in which basic factors and forces may conspire to yield some property, event, or pattern" (Clark 2001, p. 113; my italics). To define the "distinctive way" he refers to the structure of the system (e.g. the nature of the system's elements, their interconnecting topology, and the types of couplings), the system's interactions with the environment, and the concepts that figure in a good explanation of the system's behavior (e.g. collective variables, uncontrolled and invisible variables). Thus Clark's concept of emergence incorporates *structural*, *interactive*, and *explanatory* aspects. It should be noted that both sides cannot be strictly separated. Explanations make use of structural and interactive properties of a system and at the same time the concepts that are used in describing a system's structure and interactions are influenced by the explanatory approach.

In summary, Clark's notion of emergence focuses on a special class of systems that stand out due to their structure, their interactions with the environment, and the type of explanatory concepts that are used to deal with them. It should be noted that the explanatory aspect implies a "weakly observer-dependent notion"[58]. (Clark 1997, p. 113)

In the following I will present Clark's four suggestions, each concerned with a slightly different class of systems. I shall point out the characteristics of each class and the properties of the phenomena therein. It will further become apparent that these classes are not all mutually exclusive. The attempt to put in more abstract terms what Clark illustrated by interesting examples, will

[56] See Clark (2001, ch. 6) and Clark (1997, ch. 6).
[57] "Emergence, thus defined, is linked to the notion of what variables figure in a good explanation of the behaviour of the system." (Clark 1997, p. 113)
[58] "[..] it [the notion of emergence] turns on the idea of a good theoretical account and hence builds in some relation to the minds of the human scientists." (Clark 1997, p. 113)

allow for a check as to whether a stronger notion than merely weak emergence is applicable. In particular I will try to define more formally which systems are emergent according to Clark.

## Emergence as Collective Self-Organization[59]

Clark suggests that emergent phenomena might be found in self-organizing systems that consist of a large number of elements that interact according to simple rules. To get a better idea of what kind of system he has in mind, we will have a look at two examples: *flocking behavior* and *termite nest building*.

The computer artist Craig Reynolds (1987) showed that quite realistically looking flocking behavior can be replicated on a computer using a group of simulated agents. His goal was to model the characteristic contrasts that a flock exhibits.

> It [a flock] is made up of discrete birds yet overall motion seems fluid; it is simple in concept yet is so visually complex, it seems randomly arrayed and yet is magnificently synchronized. Perhaps most puzzling is the strong impression of intentional, centralized control. (Reynolds 1987).

Reynolds (1987) supposed that "flock motion must be merely the aggregate result of the actions of individual animals, each acting solely on the basis of its own local perception of the world". He based his model on the intuition that flocking birds try to balance two opposing forces: the desire to stay close to the flock and the desire to avoid collisions within the flock. To achieve this balance a bird might be aware of three perceptual categories: itself, its two or three nearest neighbors, and the rest of the flock. The opposing forces of *collision avoidance* and the *urge to join the flock* were modeled by only three simple rules.

- *Collision Avoidance*: avoid collision with nearby flockmates
- *Velocity Matching*: attempt to match velocity with nearby flockmates
- *Flock Centering*: attempt to stay close to nearby flockmates

The patterns of on-screen activity of a group of boids, following these rules, share many characteristics with the flocking behavior of real birds, schooling fish or herds of terrestrial animals. Each boid makes subtle speed and heading adjustments in order to avoid colliding with its neighbors, ranks are closed, and if an obstacle is encountered the flock parts and washes around it only to reform elegantly on the other side.

The example showed that "interesting collective effects can emerge as a result of the interactions between multiple simple agents following a few simple rules" (Clark 2001, p. 108) What Clark calls here interesting, is the *complexity* and *adaptability* of the behavior of a flock of boids. The abilities of the flock seem to exceed by far the individual boids' abilities.

There is no explicit rule for the flock to part before and reform after an obstacle. This ability seems to be implicitly contained in the structure of the "flocking system". The behavior of the flock is

[59] Clark (2001, p. 113ff)

obviously a systemic property and clearly weakly emergent, since it can be predicted by the initial conditions and the behavioral rules, which are both available.

The agent-agent-interactions between the elements of a "flocking system" seem to be mainly responsible for the resultant behavior. There are several characteristics of these interactions, many of which transfer to the second example.

The interactions are *nonlinear*, for example the metrics of attraction and repulsion are proportional to the inverse square of the distance, and support *feedback* cycles. The changes a single boid causes in its surroundings are fed back via its neighbors. Boids have no global knowledge (e.g. the flock's center is inferred from the position of the neighbors' highest density) and can likewise directly affect only their local flockmates.

The interactions are governed by three *simple* and *complementary* rules. "Static *collision avoidance* and dynamic *velocity matching* are complementary. [..] Static collision avoidance is based on the relative position of the flockmates and ignores their velocity. Conversely, velocity matching is based only on velocity and ignores position. [..] [I]f the boid does a good job of matching velocity with its neighbors, it is unlikely that it will collide with any of them any time soon. With velocity matching, separations between boids remain approximately invariant with respect to ongoing geometric flight. Static collision avoidance serves to establish the minimum required separation distance; velocity matching tends to maintain it." (Reynolds 1987)

The boids' interactions only yield flocking behavior if they take place in *large number*. A small number of boids will for example pass an obstacle on one side instead of parting. A greater number of boids helps to "smooth out" the individual boids' behaviors so that systemic behavior becomes apparent. The more interactions take place the less dependent is systemic behavior on individual behavior. In other words, multi-agent-behavior requires a critical number of agents to appear.

Following from the previous point the behavior is *robust* relative to the agents, because it does not depend on each and every boid behaving exactly according to the rules, for example the flock is not affected by the departure or arrival of a small number of boids. Though it has probably not been checked, it is imaginable that a lunatic boid erratically flies through the flock without really affecting the overall course of the flock.[60]

Connected with the robustness of the flock's behavior is the feature of *distributed control*. The flock does not follow any designated leader nor is the spatial structure determined by any single boid. Spatial as well as temporal control is distributed across the whole flock. Therefore single

[60] It would be interesting to examine how large a group of boids with deviating behavior has to be, and of what kind deviations from the behavioral standard repertoire must be, to have effects on the flock's behaviour. It should be checked if suitable modifications could yield oscillating or even chaotic behaviour.

failures or deviations can be compensated for. Note that behavior is even *distributed across behavioral modules*, i.e. over different competing rules.[61]

A single boid's behavior is *adaptive*. Because the rules are ordered according to their precedence individual boids will try to stay close to nearby flockmates as long as they are not in danger to hit an obstacle. If the acceleration demanded by the rule of *collision avoidance* reaches a certain threshold value, all less important rules are temporarily disregarded. A boid will not commit suicide just because it does not want to loose its flockmates.

The flock's behavior is *adaptive* to the environment. The spatial organization of the flock changes having the side-effects of maximized cohesion and minimized steering. The cohesion of the flock is only temporarily destroyed by obstacles, in a long run the flock stays together[62]. The flock's parting and re-joining allow the flock to continue on its course despite obstacles.

The above mentioned features illustrate that the properties of the "flocking system" are due to the organization and the interactions of its elements, which are quite independent from the system's (physical) substrate.[63] While the boids example illustrated how behavioral patterns emerge from *agent-agent* interaction the next example – termite nest building – will show the same for *agent-environment* interactions.

Termite nest building can be explained by the principle of *stigmergency*. In a *stigmergic routine*, repeated agent-environment interactions are used to control and guide a collective construction process. Termites modify their environment in response to the triggers provided by previous alterations to the environment.

Each termite acts according to two basic rules. First, they produce little mud balls that are simultaneously impregnated by a chemical trace and second, they deposited them wherever the chemical trace is strongest. At first mud balls are deposited at random, but wherever mud balls happen to lie close to each other, the marker concentration rises which leads to the deposition of even more mud balls until column-like buildings appear. If two columns are proximal to each other the drift of chemical attractants from the nearby column inclines the termites to preferentially add mud balls to the side of each column that faces the other. As this process continues the tips of the columns incline together until an arch is formed. Similar stigmergic routines yield complex structures like cells, chambers and tunnels.

---

[61] The theme of "opposing forces" will come up in the next section again. It seems that it allows an agent or a multi-agent system to settle in some kind of "equilibrium", i.e. a stable behavioral state.

[62] In this simplistic model this depends on the environment, e.g. the space and size of obstacles. Reynolds therefore suggests that his model could be improved and made more realistic by altering the spherical radius of perception to a more realistic long-range perception. (Reynolds 1987).

[63] "At the level of the individual behav*ors*, we have a clear difference in kind: Boids are not birds. [..] at the level of behav*iours*, flocking Boids and flocking birds are two instances of the same phenomenon: flocking." (Langton 1996 ,p. 68)

As in the boids example there is once again no centralized control or designated leader. The simple and complementary behavioral rules are robust to change and maintain their effectiveness in highly varied environments. Termites only interact with the features of their local environment. However, unlike the direct interactions between single boids, communication between termites is mediated by the environment. Environmental features thus act as a kind of memory and persist even if the originating individual is no longer present (Beckers et al. 1994, p. 188).

The difference between the two examples is captured by Clark's notion of *direct* and *indirect* emergence (Clark 1997, p. 73ff). While the notion of direct emergence refers to cases where multiple agent-agent interactions yield collective effects, indirect emergence refers to cases where interaction between agents is mediated by the environment. Or in explanatory terms: The difference concerns the extent to which we may understand the emergence of a phenomenon by focusing largely on the properties of the individual elements (direct emergence), versus the extent to which explaining the phenomenon requires attending to quite specific environmental details. (Clark 1997, p. 74)

Such collective effects can, according to the theory of self-organization, be described by "collective variables". A collective variable's value reflects the result of multiple agent-agent or agent-environment interactions. Similar to the Bénard experiment one could define position and acceleration vector of the flock as collective variables, which describe the flocking behavior of boids[64].

Clark takes this already as a defining feature of emergence. "[..] [A] phenomenon is emergent if it is best understood by attention to the changing values of a collective variable." (Clark 1997, p. 112) According to Clark one can define emergence as collective self-organization as follows.

> *Emergence as Collective Self-Organization.* Any adaptively valuable behavior arising as the direct result of multiple, self-organizing (via positive feedback and circular causation) interactions occurring in a system of simple and homogeneous elements is called emergent. Collective variables and control parameters play an important role in the description and explanation of the behavior.

This notion is clearly a weak notion of emergentism, because it neither presupposes unpredictability nor irreducibility, it even suggests a specific type of explanation. It focuses on a special class of systemic properties – properties of the dynamics (resp. behaviors) of self-organizing systems with multiple simple elements.

Clark's first concept of emergence manages to select a certain class of dynamic systems, and certain explanatory concepts best suited to deal with them. However, he does not indicate how crucial he considers control and order parameters. Especially for real-life systems Clark does not indicate how appropriate parameters can be found.

[64] It is less clear how suitable collective variable could be defined for the termite nest building. Clark himself does not indicate how.

## Emergence as Unprogrammed Functionality[65]

Clark's second suggestion concerns systems that display adaptive behaviors that are "not supported by explicit programming or any fully agent-side endowment" (Clark 2001, p. 114). The behaviors arise instead as a *side-effect* of repeated *agent-world interactions*. Goals and strategies are not explicitly represented or encoded in the system's structure.

A simple example for a system exhibiting such kind of behavior is a *wall-following robot*[66]. It is possible to construct a robot that is able to follow walls by equipping it with two basic behavioral rules. First, if it hits a wall, it should make a single course adjustment by a constant angle that leads away from the wall. Second, it should have bias to veer to one direction. If the robot starts near a wall, which is on the side of its bias, it will constantly drift towards it until it hits the wall and bounces off only to drift again towards it. By this bounce and veer technique the robot is able to follow walls on one hand side. To make the robot faster or to allow it to follow curved walls or a sequence of connected walls these two parameters have to be appropriately tuned.

Obviously the goal to follow a wall is not explicitly encoded in the robot. Only by interacting with an appropriate environment the robot's disposition to follow walls is realized. The wall is not explicitly represented in the robot's structure and no appropriate deflection angles are calculated. It just happens to be the case that its simple design amounts to wall following in the right context. The very implicit character of the robot's behavior entails that it can only be very *indirectly* manipulated. It is for example not possible without any great effort to instruct the robot to follow a sequence of randomly aligned walls with the greatest possible speed. Despite the lack of any goals or plans the robot's behavior appears nevertheless to be somehow intentional or purposeful to the external observer.

A defining feature of systems of this class of phenomena is that the systems' behaviors are not supported by "explicit programming or by any fully 'agent-side' endowment" (Clark 2001, p. 114). Instead they result from an iterated sequence of *agent-environment interactions*. Steels introduced the notion of an *uncontrolled variable*. "An uncontrolled variable changes due to actions of the system but the system cannot directly impact it, only through side effects of its actions." (Steels 1993, p. 37) Corresponding to the notion of an uncontrolled variable at the output side of the agent, there is the notion of an *invisible variable* at the input side of the agent. "An invisible variable is a characteristic of the environment, which we as observers can measure but the system has no way to sense it, nor does it play a role in the [..] behavior."[67]

Steels defines emergence according to this two notions. "[..] when a behavior is emergent, we should find that none of the components is directly sensitive to the regularities exhibited by the

[65] Clark (2001, p. 114)
[66] For a more complex example see, e.g., Webb's (1996) "cricket robot".
[67] "Distance to the wall" is for example an *uncontrolled* and *invisible* variable, since the robot has no means to directly measure or impact it. (Steels 1993, p. 37)

behavior and that no component is able to control its appearance directly." Clark's second notion of emergence can thus be formulated as follows.

> *Emergence as Unprogrammed Functionality.* Any adaptively valuable and only indirectly manipulable behavior that does not depend on central or explicit control structures and arises from iterated sequences of agent-environment interactions is emergent. Uncontrolled and invisible variables play an important role in the description and explanation of the behavior.

This notion is again a weak notion of emergentism, since it neither presupposes unpredictability[68] nor irreducibility. It focuses on a special class of systemic properties – properties of systems that have no internal state encoding of their goals and that engage in repeated interaction with their environment. Steels' notions of uncontrolled and invisible variables provide the explanatory aspects of this notion of emergence.

Clark distinguishes emergence as collective self-organization from emergence as unprogrammed functionality by the feature of direct manipulability. Only phenomena that fall under the former notion "allow for a form of direct control by the manipulation of a single parameter [..]" (Clark 2001, p. 114).

## Emergence as Interactive Complexity[69]

Both previous suggestions – emergence as *collective self-organization* and emergence as *unprogrammed functionality* – were based on the intuition that complex interactions yield possibly emergent phenomena. In the case of collective self-organization, qualitatively new behavior emerged as the result of multiple *agent-agent* interactions. In the case of unprogrammed functionality, repeated *agent-environment* interactions were decisive.

Accordingly, each system is described on a different level. The analysis of self-organizing systems focuses on the nature of the interactions and how they relate to the properties of the whole system. Because the agents are relatively simple and homogeneous their structure is of less interest. On the contrary, in the second category we have a coupled system consisting of environment and a single agent, whose internal structure is considered to be far more important. As a result, systems of the latter kind can no longer be adequately described in terms of self-organization.

Both cases can also be distinguished according to the environment the agents are operating in. An agent in a multi-agent system as it was depicted in the Boids example operates most of the time against the backdrop of other agents that are similar to itself. Whereas in the second case the agent's interaction with a non-agent environment plays a critical role.

Both phenomena rely on a certain number of interactions that have to take place before a behavioral pattern can be distinguished. In general, in a multi-agent environment considerable more interactions take place per time than in the agent-environment case.

---

[68] "The point is not that such behaviors are necessarily unexpected or undesigned [..]." (Clark 2001, p. 114)
[69] Clark (2001, p. 114ff)

Clark's third suggestion tries to incorporate the superordinate theme of complexity of interaction. He proposes that one could do justice to both the proceeding accounts by understanding "emergent phenomena as the effects, patterns, or capacities made available by a certain class of complex interactions between systemic components" (Clark 2001, p. 114). He wants to depict "emergence as the process by which complex, cyclic interactions give rise to stable and salient patterns of systemic behavior" (Clark 2001, p. 114).

This definition for the concept of emergence comes very close to a definition of self-organization. All in all it seems that he is aiming here at those dynamical systems, whose interactions with the environment or other systems are complex and show some form of pattern. Unfortunately, complexity of interactions is hard to define. Clark gives a set of properties that come in various degrees and could define the complexity of interactions: number, nonlinearity, temporal asynchronicity, and circularity (Clark 2001, p. 115). Intuitively, Clark wants this notion of emergence to depend on how many of those conditions are fulfilled and to what degree.

But this rather imprecise suggestion leaves many questions unanswered. How large does the number of interactions have to be? What kind of nonlinearity is required? How complex does a feedback loop have to be? What are the necessary and sufficient conditions for a set or sequence of interactions to be complex?

Despite these problems, Clark proposes a graded measure of emergence. While bounce-and-veer wall following is a case of weak emergence, convections rolls are "a classic case of strong emergence" (Clark 2001, p. 115ff). In order not to confuse Clark's weak and strong emergence with the notion of emergence as defined by philosophy, I propose to speak instead of *strong* and *weak interactive complexity*. As the new nomenclature already suggests it is probably more adequate to define a measure of interactive complexity (because complexity is a quantitative concept), than to dilute the concept of emergence by inflationary use.

To allow emergence to come in degrees conflicts with the metaphysical notion of emergence we encountered in the first part of this thesis, which only allow for a property to be emergent or not to be emergent. It would rob the notion of emergence of its intended explanatory and classificatory power if nearly everything would be emergent to some degree.

However, it is still possible to use the intuitive connection between interactive complexity and emergence to develop an alternative concept of emergence by specifying an *onset* of emergence[70] or a threshold that divides the set of all entities according to their complexity into emergent and non-emergent entities. To do so, we need an objective and quantifiable measure for interactive complexity, a fact that disqualifies Clark's intuitive account of complexity. It is possible to resort

[70] See, e.g., Steels (1993, p. 9). "We can also quantitatively identify the onset of emergence once a suitable mathematical framework exists for defining the notion of a minimal description."

to algorithmic complexity[71], which is, however, especially difficult to define for empirical systems. The second problem is to motivate a given value for the onset of emergence. At what complexity is it useful to call a given phenomena emergent?

I think the general idea to find similarities between the two above mentioned classes of (weakly) emergent phenomena is surely not wrong. Complexity of interaction seems to be the right property that links both concepts. However, given the problems to define interactive complexity for empirical systems and to motivate either a quantitative measure of emergence or a point of onset for emergence, Clark's third notion of emergence must still be thought of as vague and ill-defined and far from a formal and well-defined concept. At this point I do not see how those problems can be overcome.

## Emergence as Uncompressible Unfolding[72]

Clark's last suggestion, which is based on Bedau (1996), is concerned with a quite different sense of emergence.

> This is the idea of emergent phenomena as those phenomena for which *prediction* requires *simulation* – and especially those in which *predication* [prediction; author] of some macrostate $P$ requires simulation of the complex interactions of the realizing microstates $M_1$ – $M_n$. (Clark 2001, p. 116)

Bedau's definition of what he calls "weak emergence" is based on the notion of *underivability without simulation*. (The following definition should be considered as what Clark takes for emergence as uncompressible unfolding.)

> Macrostate $P$ of $S$ with microdynamic[73] $D$ is weakly emergent iff $P$ can be derived from $D$ and $S$'s external conditions but only by simulation.[74] (Bedau 1997)

According to this definition, a systemic state of a given system is emergent if and only if it can be derived given the system's initial conditions and the sequence of all other external conditions, because the system's microdynamics completely determine each successive microstate of the system.

> To simulate the system one iterates its microdynamic, given a contingent stream of external conditions as input. Because the macrostate $P$ is a structural property constituted out of the system's microstates, the external conditions and the microdynamic completely determine whether $P$ materializes at any stage in the simulation. [..] What distinguishes a weakly emergent macrostate is that this sort of simulation is *required* to derive the macrostate's behavior from the system's microdynamic. (Bedau 1997; my italics)

Unfortunately Bedau does not define formally what he understands by *simulation*. It should become clearer if we analyze one of his examples.

---

[71] Complexity is usually defined as the minimal length of a binary program from which the information content of the target phenomena can be effectively reconstructed.
[72] Clark (2001, p. 116ff)
[73] A microdynamic governs the time evolution of $S$'s microstates.
[74] Note that Bedau explicitly restricts this definition to a given macrostate of a given system with a given microdynamic.

Conway's *Game of Life*[75] provides many structural macrostates that are weakly emergent according to Bedau's definition, such as *indefinite growth* and *glider*[76] *spawning*. In general it is impossible to tell whether a given initial *Life* configuration will grow indefinitely or spawn gliders. "[..] [A]ll we can do is let Life 'play' itself out." (Bedau 1997)

I think that what Bedau understands by simulation can be best captured roughly by a function (or set of functions) that is (are) applied to itself (themselves), similar to difference and differential equations. In fact, the *Game of Life* can be thought of as a difference function that given the states of all cells at time $t$ yields the states of all cells at $t+1$. To determine if a given *Life* pattern expands indefinitely we have to apply the *Life* function $f$ multiple times to itself

$$f(f(\ldots f(f(x, \alpha))\ldots))$$

until the resulting *Life* pattern somehow stopped growing. This iterated sequence of a function applied to itself, can be thought of as a simulation, simulating the evolution of state vector $x$ and external influences $\alpha$ in discrete time steps according to dynamics of $f$.

There are now two possible reasons to call the simulation of a given set of microstates *required* in the above sense. A simulation can either be required if there is no other simpler function $g$ than $f$ or if there is a sequence with less iterations.

Simpler functions are all functions that do not depend on each and every microstate (i.e. every element of the state vector) and on each and every environmental input. Such functions are abstractions. In fact, most simulations are already abstractions, because they usually disregard certain properties of the modeled systems and their environmental inputs. This is especially true for simulations of empirical systems. For the simulation of a pendulum one usually disregards the electromagnetic or chemical properties of the elements constituting the pendulum. A simpler *Life* function would be for example a function that yields the next state of the grid given only the states of every second cell.

The notion of "less iterations" has to be set in relation to the timescale defined by the microdynamics. A simulation is required if any other simulation works with an equal or finer timescale. The microdynamics of the *Life* function prescribe the timescale for the evolution of each cell. A simulation in the above sense is no longer required if it is possible to disregard the timescale and to simulate for example only half or none of the intermediary timesteps. (E.g. under certain conditions Newton's laws allows one to determine the position of a mass on which a force acts without determining all the intermediary positions.)

For *metaphysical* relevance Bedau's notion of emergence requires that it is impossible *in principle* to circumvent the above sketched simulation. This fact can indeed be established for certain

[75] The *Game of Life* is a cellular automaton, which runs on an infinite two-dimensional grid and evolves in discrete time steps. Each cell's state at $t+1$ is determined by the states of its eight neighbouring cells at $t$.
[76] A *glider* is a period four pattern that moves diagonally one cell per period.

algorithmic problems, such as semi-decidability problems (e.g. the halting problem). There is in general no shorter way for determining if a given Turing machine halts than to run it. However, this is a special case, since the simulation of the system is identical with the evolution of the original system.

In practice simulations are based on abstractions or simplifications of a system and its dynamics. Usually not the system in its entirety, but only certain aspects of it are simulated, thus reducing the high-dimensionality of the physical microdynamics. This makes it difficult to prove that the simulation of microstates $M_1$ – $M_n$ with microdynamics $D$ cannot be bypassed by simulating simpler microstates $N_1$ – $N_n$ or by using a coarser time scale than prescribed by $D$ or by completely circumventing the whole procedure with a law $E$ that, given the initial conditions, yields the macrostates at any point in time without calculating intermediary steps. Thus Bedau's notion of emergence can only be established for well-defined abstract systems so far.

Clark takes Bedau's notion of emergence as overly restrictive, because even in cases involving multiple complex interactions, it will often be possible to model systemic unfolding by simulating only a *subset* of the actual interactions. This is, as I explained above, true. But Clark's criticism somehow misses the point, because Bedau's intentions are complementary. His notion of emergence depends exactly on the fact that it is *not* possible to model a subset of the systemic unfolding.

In contrast to Bedau, Clark takes the possibility to describe a complex system's behavior by collective variables as significant for emergence (c.f. preceding accounts). To restrict the notion of emergence to phenomena that resist all such attempts of low-dimensional modeling would clearly contradict this conception.

In summary, Bedau's notion of "weak emergence" focuses on computational systems or abstract descriptions of empirical systems, because it requires the exact specification of the initial microstates and microdynamics. In the computational domain it is a concept that can usefully describe phenomena as they appear for example in the *Game of Life* or other cellular automatons. However, it should be noted that such phenomena are predictable by and reducible to the initial conditions and microdynamics, and thus only weakly emergent.

# Conclusion

It is hard to come to a conclusion without rushing to a conclusion. On the one hand we have emergentism qua irreducibility or unpredictability and on the other hand there are several different notions of emergence that are closely related to weak emergentism. Weak emergentism was originally introduced only to be the conceptual base for all stronger notions of emergence. "Do we need that many concepts of emergence?" might seem like a legitimate question, and is there room for all of them?

There are basically two different ways to ask those questions: (i) Do Clark's notions of emergence in general enrich the conceptual framework constituted by the philosophical notions of emergence? (ii) Is there need for any single notion of emergence that Clark suggests? I will try to answer the first question by confronting the metaphysical notions of emergence with their scientific import according to Clark, before I explain the merits and shortcomings of each of the four individual notions of emergence that Clark suggested.

We remember that the metaphysical concept of emergence tries to define the relation of higher-level properties of a system to the lower-level properties of its components. It is an ontological concept classifying properties according to their reducibility and predictability. In contrast, the scientific import of emergence that Clark promotes is not so much interested in *in principal* notions. It focuses more on methodological or empirical aspects. Clark asks *how* a given system can be explained not *if*.

It is his aim to classify complex dynamical systems according to their explanatory requirements. He takes the concept of weak emergentism and adopts it to special cases (e.g. self-organizing systems or autonomous robots). Rather intuitively he picks out systems that show "interesting" behavior, which he later specifies as autonomous, self-organized, adaptive, robust, or non-centrally controlled. The fact that these are all systemic properties of the analyzed systems technically allows him to call them emergent, but only in a weak sense.

But is there a reason to call the properties of those systems *emergent* instead of calling them simply systemic and specifying them in addition as adaptive or autonomous? I think the original gist of emergence qua irreducibility or unpredictability is definitely lost. Clark will always have to put up with those accusing him of using the notion of emergence for a purpose not intended.

However, he does not duck the issue, because he explicitly states that he does not insist that only unexplainable phenomena should count as emergent (Clark 2001, p. 113) nor does he preclude predictable phenomena from being emergent (Clark 1997, p. 109). This is however due to a slight misconception. He criticizes the notion of unpredictability as overly observer-relative, because he thinks in terms of *psychological* predictability (Walmsley 2003, p. 9).

Therefore Clark rightly contends that *psychological* predictability should not be the hallmark of emergentism. But his criticism falls short of attacking strong notions of emergence, because a definition of emergence in terms of the absence of *logical* predictability or "deducibility" is indeed observer independent (Walmsley 2003, p. 10). Thus, from a philosophical perspective there is no compelling reason to necessarily call those phenomena emergent, if "emergent" denotes stronger versions of emergentism.

But one has to give Clark credit for developing a quite different approach on the notion emergence. Let us momentarily leave aside metaphysical issues of stronger versions of emergentism and look objectively at how Clark is interpreting the concept of emergence. His goal to classify systems according to their explanatory requirements has its origin in the failure of traditional componential explanations in contemporary scientific research. Clark (1997, p. 114) notes that aggregate systems, i.e. systems in which the parts display their explanatory relevant behavior even in isolation from one another, are the ones for which componential explanations are best suited, but as soon as the complexity of interaction between parts increases new kinds of explanatory frameworks are needed.

Clark introduces dynamical explanations and emergent explanations. Dynamical explanations have the advantage of providing the same approach to tackle many different kinds of phenomena. They span the system and its environment as well as a system's different levels.

Emergent explanations provide understanding in terms of collective variables – variables that fix on higher-level features, but do not track properties of simpler components. "[..] [B]y plotting relations between the values of the collective variables and control parameters [..] we may come to understand important facts about the circumstances in which higher-level patterns will emerge, when one higher-level pattern will give way to another, and so on." (Clark 1997, p. 108)

However, Clark does not take those models as superior to more traditional analytical approaches. They are best seen as complementary and cognitive science cannot afford to do without any of the various explanatory styles. We must therefore ensure that the various explanations somehow interlock and inform each other (Clark 1997, p. 119).

Clark takes the pros and cons of the different styles of explanation as a powerful argument for an *explanatory liberalism* (Clark 1997, p. 119). This explanatory liberalism then quite naturally motivates the development of a concept for distinguishing systems according to their explanatory demand.

In summary, the fact that he is concerned with systemic properties allows him to speak of *weak* emergence, although he does not explicitly mention it. Beyond that he developed an explanatory approach to complex dynamical systems, which have only recently come into reach of decent scientific investigation and theorizing. His approach should not only be regarded as being methodological relevant, but also of metaphysical importance. Because traditional ways of

explanations have become powerful cousins we have to think about how to integrate them into our theorizing. Furthermore, let me point out that the big philosophical question "What are explanations?" is still far from settled. Thus several models of explanation exist that all have their advantages and disadvantages and it cannot be said which of those are really necessary and which can be replaced or conjoined. Clark at least makes the effort to test some of them and shows how they can be put to use.

If we take all of Clark's notions of emergence without discussing their legitimacy compared to metaphysical notions of emergence, we can analyze their legitimacy among each other. Here the first two – emergence as self-organization and emergence as unprogrammed functionality – comply with Clark's concept of an explanatory notion of emergence best. Both specify a class of systems, and particular explanatory concepts related to them. Emergence as interactive complexity still needs to be worked out. It is less sharply outlined, because of the difficulties in defining interactive complexity. However, it has the potential of integrating the previous two notions into one.

And finally emergence as uncompressible unfolding stands a little apart from the other notions, because it is more criticized by Clark than developed into a formal concept. It does not fit into Clark's explanatory approach on emergence, i.e. Clark does not mention any specific explanatory methods tied to it. So far it has only relevance in a restricted and well defined domain, which is however not necessarily less interesting. Cellular automatons and their kin still hold secrets that demand special modes of explanation and might even provoke diehard philosophical discussions.

That leaves us with two notions of weak emergence that correspond with Clark's explanatory approach on emergence and that have the potential to be picked up by the scientific community and put to their proper use. Emergence as self-organization, as well as emergence as unprogrammed functionality, can play a valuable role in the analysis and explanation of particular classes of complex dynamical systems.

★

The lesson to learn from this thesis is that philosophers and (cognitive) scientists have essentially different concepts in mind when they speak of emergence. Although both ideas of emergence are related by the notion of weak emergence, they differ in the further development of this concept. While philosophers want to express the intention that there are phenomena that are unpredictable or irreducible in principle, scientists work with the opposite assumption. They admit that there are phenomena that are unexplainable by current methods, but at the same time they develop new empirical methods and explanatory frameworks to approach them.

Because both sides have already established their particular idea of emergence in literature and discussion, it would be futile to change either one of them. Therefore it is necessary to heighten the awareness of both sides for each others' concepts. This was one of the goals of this thesis and also

that of other works confronting philosophical notions of emergence with their scientific import[77]. A first step towards each other would be to explicitly call "*weakly* emergent" what is only weakly emergent. This shows that somebody who uses a concept of "emergence" only to denote systemic properties is well aware of its philosophical status. The notion of weak emergence provides just the right starting point for the clear classification of all concepts of emergence.

[77] See, e.g., Walmsley (2003).

# Appendix

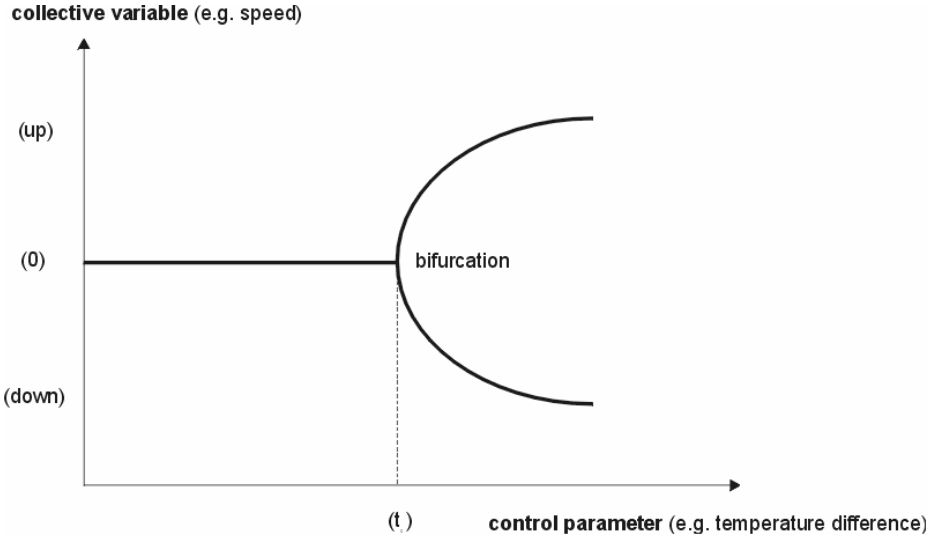



**Figure 1:** Schematic bifurcation diagram of a period-doubling bifurcation. The convection roll example is included in brackets.

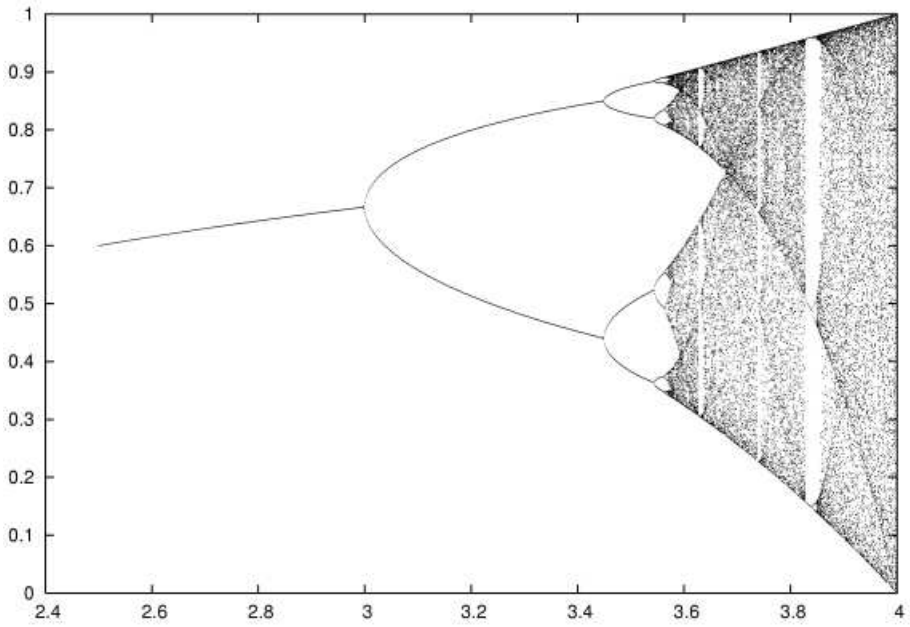



**Figure 2: Bifurcation diagram of the logistic map.**

# References[78]

Alexander, S. (1920). *Space, Time, and Deity. The Gifford Lectures at Glasgow 1916-1918.* Two Volumes. London: Macmillian.

Beckers, R., Holland, O., and Deneubourg, J. (1994). From local actions to global tasks: Stigmergy and collective robotics. In *Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, 181-189.

Bedau, M. A. (1996). The Nature of Life. In M. A. Boden (ed.), *The Philosophy of Artificial Life.* Oxford, England: Oxford University Press, 332-357.

Bedau, M. A. (1997). Weak Emergence. In J. Tomberlin, (ed.), *Philosophical Perspectives: Mind, Causation, and World. Vol. 11.* Malden, MA: Blackwell, 375-399.
`http://www.reed.edu/~mab/papers.html`

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bonabeau, E. (2002). Predicting the Unpredictable. *Harvard Business review*, March 2002.

Bonabeau, E., Funes, P., & Orme, B. (2003). Exploratory Design Of Swarms. In C. Anderson & T. Balch (eds.), *Proceedings of the Second International Workshop on the Mathematics and Algorithms of Social Insects*, Georgia Institute of Technology, 17-24.

Broad, C. D. (1925). *Mind and Its Place in Nature.* London: Routledge and Kegan Paul.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA: MIT Press.

Clark, A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science.* Oxford, New York: Oxford University Press, Inc.

El-Hani, C. N. and Emmeche, C. (2000). On Some Theoretical Grounds for an Organism-centered Biology: Property Emergence, Supervenience, and Downward Causation. *Theory Biosci.* 119: 234-275.

Glass, L., et. al. (1984). Global bifurcations of a periodically forced biological oscillator. *Phys. Rev.* A 29, 1348-57.

Haken, H. (1977). *Synergetics, an Introduction. Nonequilibrium Phase-Transistions and Self-Organization in Physics, Chemistry and Biology.* Berlin: Springer-Verlag.

Haken, H. (1988). *Information and Self-Organization. A Macroscopic Approach to Complex Systems.* Berlin, Heidelberg, New York: Springer-Verlag.

Heylighen, F. (1989a). Self-Organization, Emergence and the Architecture of Complexity. *Proc. of the first European Conference on System Science*, 23-32.
`http://pespmc1.vub.ac.be/Papers/PapersFH.html`

Heylighen F. (1989b): *Causality as Distinction Conservation: a theory of predictability, reversibility and time order*, (submitted to Synthese).
`http://pespmc1.vub.ac.be/Papers/PapersFH.html`

Heylighen, F. (1999). The Science of Self-Organization and Adaptivity. In *The Encyclopaedia of Life Support Systems*, EOLSS Publishers Co. Ltd.
`http://pespmc1.vub.ac.be/Papers/PapersFH.html`

[78] In case of doubt, page and section numbers refer to the online editions of the listed papers.

Holland, J. H. (1998). *Emergence: From Chaos to Order.* Oxford, New York: Oxford University Press, Inc.

Hüttemann, A. (2000). Emergence in Physics. *International Studies in the Philosophy of Science.* 14: 267-281. http://www.uni-bielefeld.de/philosophie/personen/huettemann/

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly,* 32, 127-136.

Jaeger, H. (1995). Dynamische Systeme in der Kognitionswissenschaft. Arbeitspapiere der GMD 925, GMD, St. Augustin.

Johnson, S. (2001). *Emergence: The connected lives of ants, brains, cities, and software.* New York: Scribner.

Kantz, H. (1999). Nichtlineare Zeitreihenanalyse in der Physik: Möglichkeiten und Grenzen. In K. Mainzer (ed.), *Komplexe Systeme in Natur und Gesellschaft. Komplexitätsforschung in Deutschland auf dem Weg ins nächste Jahrhundert*, Berlin: Springer, 74-88.

Kaplan, D., Glass, L. (1995). *Understanding Nonlinear Dynamics.* New York, Berlin, Heidelberg: Springer-Verlag New York, Inc.

Kelso, S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior.* Cambridge, MA: MIT Press.

Kim, J. (1992): "Downward Causation" in Emergentism and Non-reductive Physicalism, in: A. Beckermann, H. Flohr, and J. Kim (eds.), *Emergence or Reduction? - Essays on the Prospects of Nonreductive Physicalism*, Berlin, New York: de Gruyter, 119-138.

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation.* Cambridge, MA: MIT Press.

Kim, J. (1999). Making Sense of Emergence. *Philosophical Studies*, 95, 3-36.

Langton, C. G. (1996). Artificial Life. In M. A. Boden (ed.), *The Philosophy of Artificial Life.* Oxford, England: Oxford University Press, 39-94.

Levine, J. 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly,* 64, 354-361.

Lloyd Morgan, C. (1923). *Emergent Evolution.* London: Williams and Norgate.

Marras, A. (2003). *Being Realistic about Reduction: Reply to Kim.* (unpublished) Reply to Kim at the conference *Reduction and Emergence* of the Institute Jean Nicod 2003.

McLaughlin, B. (1992). *The Rise and Fall of British Emergentism.* In A. Beckermann, H. Flohr and J. Kim (eds.) *Emergence or Reduction? Essays on the Prospects of Non-reductive Physicalism.* Berlin, New York: Gruyter, 49-93.

Müller , A., Kögerler, P. (1999) Vom Einfachen zum Komplexen: Bildung von chemischen Strukturen. In K. Mainzer (ed.), *Komplexe Systeme in Natur und Gesellschaft. Komplexitätsforschung in Deutschland auf dem Weg ins nächste Jahrhundert,* Berlin: Springer, 103-116.

Nagel, E. (1961). *The Structure of Science.* Routledge and Kegan Paul, London.

Resnick, M. (1994). *Turtles, Termites, and Traffic Jams: Explorations in Massively Parallel Microworlds.* Cambridge, MA: MIT Press.

Reynolds, C. (1987). Flocks, Herds, and Schools: A Distributed Behavioral Model. *Computer Graphics.* 21(4), July, 25-34.

Sellars, R. W. (1922). *Evolutionary Naturalism.* Chicago: Open Court Publ. Co. Reissued 1969, New York: Russel & Russel.

Smith, L.B., Thelen, E. (Hrsg.) (1993): *A Dynamic Systems Approach to Development: Applications.* Cambridge, Mass: Bradford/MIT Press.

Steels, L. (1994). The Artificial Life Roots of Artificial Intelligence. *Artificial Life Journal*, vol. 1, nr. 1-2, pp. 89-125, Cambridge, MA: The MIT Press. http://arti.vub.ac.be/~steels/

Stephan, A. (1999). *Emergenz: Von der Unvorhersagbarkeit zur Selbstorganisation.* Dresden, München: Dresden University Press.

Stephan, A. (1999a). *Varieties of Emergence.* Evolution and Cognition Vol. 5, No. 1: 49-59.

von Bertalanffy, L. (1969). *General System Theory.* New York: George Braziller, Inc.

van Gelder, T. J. (1998) The dynamical hypothesis in cognitive science. Behavioral and Brain Sciences, 21, 1-14. http://www.arts.unimelb.edu.au/~tgelder/Publications.html

Weizäcker, E.U. (1987). Brückenkonzepte zwischen Natur und Geisteswissenschaften: Selbstorganisation, Offene Systeme und Evolution. Sozialökologische Arbeitspapiere 17, Forschungsgruppe Soziale Ökologie, Frankfurt/Main 1997, 1-27.

Walmsley, J. (2003). Dynamical Systems, Emergence, and Explanation. http://www.lehigh.edu/~interact/isi2003/isi2003.papersandnotes.html

Webb, B. (1996). A Cricket Robot. *Scientific American*, 275, 62-67.

Wimsatt, W. C. (1997). Functional Organization, Functional Inference, and Functional Analogy. *Evolution and Cognition*, 3 (#2): 102-132.

Wimsatt, W. C. (2002). *Emergence as Non-Aggregativity and the Bias for Reductionism.* (Forthcoming in Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality, Cambridge: Harvard U. P.) http://www.institutnicod.org/reduction.htm