Biology and Philosophy (2006) 21:501–521
DOI 10.1007/s10539-005-9003-9

On emergence, agency, and organization

STUART KAUFFMAN[1,*] and PHILIP CLAYTON[2]
[1]*The Institute for Biocomplexity and Informatics, The University of Calgary, Calgary, AB T2N 1N4, Canada;* [2]*Philosophy, Claremont Graduate University and CST, 1325 N. College Avenue, Claremont, CA, 91711, USA; *Author for correspondence (e-mails: skauffman@ucalgary.ca, skauff@telus.net; phone: +403-220-8349; fax: +403-242-8771)*

Received 16 July 2004; accepted in revised form 26 July 2005

Key words: Autocatalysis, Autonomous agents, Emergence, Preadaptation, Reductionism, Theory of organization, Semiotics, Teleology, Underdetermination of biology by physics, Work cycle

Abstract. Ultimately we will only understand biological agency when we have developed a theory of the organization of biological processes, and science is still a long way from attaining that goal. It may be possible nonetheless to develop a list of necessary conditions for the emergence of minimal biological agency. The authors offer a model of molecular autonomous agents which meets the five minimal physical conditions that are necessary (and, we believe, conjointly sufficient) for applying agential language in biology: autocatalytic reproduction; work cycles; boundaries for reproducing individuals; self-propagating work and constraint construction; and choice and action that have evolved to respond to food or poison. When combined with the arguments from preadaptation and multiple realizability, the existence of these agents is sufficient to establish ontological emergence as against what one might call Weinbergian reductionism. Minimal biological agents are emphatically not conscious agents, and accepting their existence does not commit one to any robust theory of human agency. Nor is there anything mystical, dualistic, or non-empirical about the emergence of agency in the biosphere. Hence the emergence of molecular autonomous agents, and indeed ontological emergence in general, is not a negation of or limitation on careful biological study but simply one of its implications.

> An organized being is then not a mere machine, for that has merely *moving* power, but it possesses in itself *formative* power of a self-propagating kind which it communicates to its materials though they have it not of themselves; it organizes them, in fact, and this cannot be explained by the mere mechanical faculty of motion. (Immanuel Kant, *Critique of Judgment* [1987]: 221)

Introduction

The past decade has seen a re-emergence of interest in the concept of emergence, and not only in the pages of this Journal. Of course, the debate between reductionist theories and those that endorse emergent properties and entities has a long history: it bloomed briefly in the early years of the 20th century with the British Emergentists, was dormant for some decades, and has again blossomed (cf. McLaughlin 1992; Clayton 2004: chapter 1). Clearly, though, the

recent discussion has produced more nuanced arguments, and a more subtle interplay between philosophy and biology, than one finds in the earlier literature.

The aim of this article is to defend one very precise and limited form of emergence: the emergence of biological agency. Five minimal physical conditions, we will argue, are necessary for applying teleological or agential language in biology; and taken together, we suggest, they are sufficient: autocatalytic reproduction, work cycles, boundaries for reproducing individuals, self-propagating work and constraint construction, and choice and action that have evolved to respond to (e.g.) food or poison.

Although our argument is clearly relevant to the broader philosophical question of the emergence and nature of conscious agency, we here remain agnostic on that topic. From the standpoint of the philosophy of biology, as we hope to show, it is much more urgent to develop an account of the organization of biological processes and to specify the necessary and sufficient conditions for biological agency. Science will only understand the emergence of autonomous agents in the biosphere when it develops an adequate theory of the organization of processes, and in particular when it learns how biochemical processes self-organize. The account of autonomous agents proposed here is at least a first step toward comprehending biological self-organization.

The philosophical context

We assume the so-called reduction vs. emergence debate is familiar to readers and thus provide only the briefest summary of the conceptual background presupposed by our argument. Reduction involves the assumption that one can construct a hierarchy of scientific disciplines, and that the causal and explanatory connections move in particular directions across the hierarchy. The causal arrow points upward from the fundamental microphysical causes. Hence what we call chemical or biological or psychological causes are merely more complex manifestations of the fundamental causes and are ultimately reducible to them. Explanations map causal relations: the complex phenomena of biology or psychology are to be explained by reducing them to their fundamental causes, i.e., microphysical particles and forces.

Among scientists, the physicist Steven Weinberg (1992) offer a classic formulation of the strongly reductionist stance. Although he is hesitant to claim that scientists will someday be able to *deduce* upwards from a hoped-for final physical theory, Weinberg does argue strongly for explanatory reduction, for example from social behaviors to individual organisms to cells to molecules to chemistry and ultimately to physics. Many physicists still pre-suppose some form of classic reductive physicalism, but not all; Philip W. Anderson's influential 'More is Different' (Anderson 1972) represents a well-known counterexample, and the Nobel laureate physicist Robert Laughlin has just published a frontal attack on reductionism in physics (Laughlin 2005). There

are, it seems, as many varieties of reduction as there are of emergence, and the number is not small. Nonetheless, we do not agree that the nuancing of 'reduction' and 'emergence' in the recent literature erases all conceptual difference between them or renders the distinctions otiose. To the contrary.

Phenomena that cannot be deduced from underlying laws (e.g., quantum mechanics and deterministic chaos) are generally called epistemically emergent. Emergence in this sense presupposes the existence of levels of organization in the natural world. Wimsatt offers the classic definition: 'By level of organization, I will mean here compositional levels – hierarchical divisions of stuff ... organized by part–whole relations, in which wholes at one level function as parts at the next (and at all higher) levels ...' (Wimsatt 1994, 222). We take it to be uncontentious that at least some epistemically emergent phenomena exist.

Ontological emergence, by contrast – the view that new 'higher' levels of entities arise and have causal powers not possessed by the parts – is sharply contested. Candidates for such higher-level causes include mental causes and the emergence of causal agents in the biosphere. Rightly or wrongly, talk of ontological emergence raises the specter of an anti-scientific dualism, especially for philosophers of science (Clayton 2004, chapter 4). For those who emphatically reject dualism, as we do, it will be crucial to identify potential cases of ontological emergence prior to consciousness if this view is to be taken seriously at all. Does life represent an emergent organization of matter, energy, and process? And, in particular, do molecular autonomous agents – distinctively biological causal entities – exist?

Minimal holism

There is a sense of holism – call it 'robust' or 'maximal holism' – that stands in tension with causal explanations in the natural sciences. A system S is *maximally holist* if it does not admit of analyses in terms of the laws, regularities, particles, or causal powers that underlie S, or if such analyses do not help to explain the phenomena associated with S. By contrast, a system S′ is *minimally holist* if (1) S′ is not maximally holist and (2) analyses of S′ in terms of lower-level laws, causes, etc. do not fully or adequately explain the phenomena associated with S′. Minimal holism is not opposed to (suitably defined) reductionist explanations; as Wimsatt notes, 'it is possible to be a reductionist and a holist too' (Wimsatt 1994, 225). We suggest that functional explanations in biology are minimally holist.

The function of the heart is, roughly, to pump blood, not to make heart sounds. Now suppose we grant that physics might one day be able to give a complete account of all the causal properties of the heart. That account would nevertheless not be able to pick out the particular subset of causal consequences that constitutes the function of the heart. To find that function, it

would seem, requires an analysis of the organism – in its environment, through its life cycle, and in light of its particular selective situation.

In short: biological functions, for example the functions of organs, are subsets of their causal consequences and must be analyzed at the level of whole organisms, their environment, and their evolutionary history. The physicist cannot identify the particular subset of consequences that are the function of an organ without using explanations involving selection, thereby relying on biological theory. Hence, in at least this limited sense, biological explanations necessarily move beyond explanations at the physical level alone.

We find the minimal holism that is inevitably entailed by functional explanations to be both non-mysterious and completely consistent with biological science. In biology, to say that a given causal consequence of a part is the function of that part is to say that this particular causal consequence was selected for. Since discriminating biological function invariably includes a reference to evolutionary history, a physicist's account of that function would have to incorporate an evolutionary component, which would necessarily include reference to Darwin's mechanisms of evolution. In short, physicists will still have to play at the level of evolutionary explanations if they are to discriminate the functions of parts of organisms (as they must).

Human agency and teleology

Volumes have been written on human agency and teleological explanations; thankfully we do not here need to defend a particular position on the nature of human agency. But the argument that follows will have to use this well-established way of speaking of agency as a touchstone – and, in part, as a point of contrast – for developing a theory of biological agency. We thus note merely that ascriptions of agency and purpose represent a standard language game (in a loosely Wittgensteinian sense) in accounts of human behavior. Humans regularly offer teleological or 'means-ends' explanations in which reasons appear as causes of behavior (von Wright 1971; Chisholm 1976; Bishop 1983; O'Connor 1995). If human action, reasons, motives, intents, and purposes do not offer an adequate vocabulary and framework for the attribution of teleological explanations, nothing does. And if the attribution of agency is (as it seems) a viable language game, human action offers a vast array of situations upon which to practice that language game.

Whatever theoretical stance one may ultimately take regarding full-blown human action, it is a stunning fact that the universe has given rise to entities that do, daily, modify the universe to their own ends. We shall call this capacity *agency*.

We dwell no further on the human case, not because there are no important issues here – indeed, the issues are of paramount importance – but because there is a prior question that philosophers of biology must address before the behaviors of *homo sapiens* could ever be adequately explained: what is the

minimal natural system to which one might attribute teleological or purposive explanations? We shall supply a candidate answer by attempting to define *molecular autonomous agents*; we then explore some of the ramifications of this definition. We note in advance that definitions are neither true nor false, but fruitful or barren.

Molecular autonomous agents

As noted, the goal is to identify the minimum natural system to which it makes sense to attribute teleological explanations. Consider, then, a bacterium swimming up a glucose gradient. This is a case in which biologists normally say that the bacterium is 'going to get food.' That is, in a sense to be explicated below, the bacterium is acting on its own behalf in an environment. We shall call a system able to act on its own behalf *an autonomous agent*. By 'autonomous' we do not mean that the system is isolated, but only that it can act on its own behalf. Now if the bacterium is an autonomous agent, able to act on its own behalf, then *a fortiori* all free-living organisms are autonomous agents.

There is nothing mysterious or anti-scientific about this ascription of agency. From another perspective, the bacterium is 'just' a physical system. But then one wants to know: what must a physical system be such that it can act on its own behalf?

We propose a tentative five-part definition of a minimal molecular autonomous agent: such a system should be able to reproduce with heritable variation, should perform at least one work cycle, should have boundaries such that it can be individuated naturally, should engage in self-propagating work and constraint construction, and should be able to choose between at least two alternatives.

'Choose' is, of course, teleological language; applied to bacteria it will have precious few of the connotations that it has in the language game of human agency. The term must therefore be pared down to its absolute minimum, since we are seeking the minimal physical system to which one might apply teleological language.

These considerations lead to the idea of a hypothetical chemical autonomous agent (see Kauffman (2000) for more detail). The first step for conceiving this form of agency is the existence of an open, far from equilibrium, thermodynamic system. The argument then turns on a fuller understanding of what is involved in the notion of chemical work. Consider in particular the Carnot cycle. The Carnot cycle is an idealization that expresses the most efficient heat engine cycle allowed for by natural law. It consists of two isothermal processes and two adiabatic processes. To picture the cycle, consider a hot and a cool heat reservoir, a cylinder with a piston inside, and working gas between the piston and the head of the cylinder. During the work cycle there is a power stroke in which the piston starts high in the cylinder, the working gas at the high temperature of the hot reservoir. During the power stroke, the working

gas does work to push the piston down the cylinder in the isothermal part of the power stroke – that is, the temperature of the gas does not fall significantly due to contact with the hot reservoir. Next the cylinder is taken out of contact with both reservoirs during the adiabatic part of the power stroke, when the gas cools due to expansion. The power stroke is a spontaneous process. After the power stroke the now cool gas needs to be recompressed to return the piston to its initial condition high in the cylinder. The cylinder is placed in contact with the cool reservoir and work is done on the piston, in a non-spontaneous process, to recompress the gas. The way a heat engine works depends upon the fact that it takes less work to recompress the gas if it is cool. Thus the contact with the cool reservoir sustains the low gas temperature during this first, isothermal, part of the compression stroke. Then the cylinder is brought out of contact with both reservoirs and more non-spontaneous work is done on the piston to recompress the gas and heat it to the initial state of the engine in the adiabatic part of the compression stroke, completing the work cycle.

A number of features about the work cycle are important. First, the cycle links a physically spontaneous process and a non-spontaneous process. Second, at equilibrium, no work cycle can occur. Since we have defined a molecular autonomous agent as a physical system that does a work cycle, it follows that agency is a non-equilibrium concept. Third, the work cycle brings the operating organization of the engine back to its initial arrangement of parts ready for another cycle. Thus, the Carnot engine is an example of a cyclic organization of process. It appears that we have no formal language to talk about such organization. We return to this below.

In a chemical reaction system, the concept of a spontaneous process is an exergonic chemical reaction – the approach of the system to equilibrium. A non-spontaneous process is an endergonic process, where work is done on the chemical system to 'push' it beyond equilibrium, towards the excess synthesis of some chemical component compared to its equilibrium concentration. This requires linking the endergonic reaction to some form of spontaneous process, often by linking the reaction to a second exergonic reaction. With this in mind, we can describe our hypothetical molecular autonomous agent.

The first part of the system consists in a *molecular autocatalytic system* in which two DNA trimers join end to end in a proper 3′–5′ phosphodiester bond to form a hexamer. This reaction is catalyzed by an initial copy of the hexamer single-stranded DNA sequence. Hence the molecular system is autocatalytic and self-reproducing. The second part of the system is the *work cycle*, which is used to drive *excess* synthesis of the hexamer from the trimers. More technically, exergonic chemical energy from the work cycle is used to drive the endergonic, hence non-spontaneous, synthesis of excess hexamer compared to its equilibrium concentration. The work cycle motor consists in coupled reactions. Its driving force consists in pyrophosphate (PP), which is present in excess of its equilibrium concentration. PP breaks down exergonically, that is spontaneously with a loss of free energy, to two monophosphates, P + P. The

breakdown of PP to P + P is coupled to the synthesis of hexamer from the two DNA trimers and provides the chemical energy to drive that reaction endergonically such that excess hexamer is synthesized compared to its equilibrium concentration. Once this has occurred, PP must be restored to its initial concentration such that another cycle of work can be done. Restoring PP to its initial concentration is accomplished by another coupled pair of reactions where an exergonic reaction drives the non-spontaneous, hence endergonic, resynthesis of PP. An electron in its ground state absorbs a photon and is lifted to an excited state. Note that work has been done on the electron. As the electron falls back spontaneously to its ground state exergonically, that energy is coupled to the endergonic resynthesis of PP from P + P, until it reaches the initial concentration of PP. The motor in this system is the net rotation of monophosphates around the reaction cycle. Thus the system both reproduces and does a work cycle.

In a heat engine, the organization of processes is accomplished by cams, gears, and so forth. In order to coordinate the reactions of the molecular agent such that the forward synthesis of hexamers occurs with the depletion of PP, the PP subsequently being restored to its initial level (that is, that the two reactions occur successively like the power and compression strokes of the Carnot cycle), we suppose that monophosphate, P, feeds back to activate the hexamer catalyst, while PP feeds forward to inhibit the resynthesis of PP from monophosphate, a reaction catalyzed by one of the DNA trimers. These regulatory couplings assure that the two reactions occur reciprocally, as demonstrated by writing down the differential equations for the system and finding a limit cycle oscillation for its dynamics (Daley et al. 2002).

We want the molecular system to have boundaries in order to be individuated. But an even deeper reason to require boundaries is that the molecules would otherwise diffuse out of effective contact with one another. Thus we suppose that the remaining DNA trimer is able to act as a catalyst to ligate two molecules, X and Y, which jointly create a lipid-like molecule Z. In an aqueous medium lipids will fall to a low energy state in which they form vesicles that are lipid bilayers, called liposomes, which are deeply similar to cell membranes. Thus we suppose that Z synthesis leads to the formation of a bounding membrane containing the reaction members discussed. For the purpose of the following discussion, let us suppose that the formation of Z requires that chemical work be done. For example, a high energy molecule such as PP might break down exergonically and be coupled to the ligation of X and Y to drive the endergonic synthesis of Z.

This system is a perfectly legitimate, if hypothetical, thermodynamically open chemical reaction network. It does not cheat the second law of thermodynamics, for its food sources are X and Y, the two DNA trimers, and photons. Indeed, we must assume that these food molecules can diffuse across the bilipid membrane to reach the interior of the 'cell.'

Finally, we wish our system to be able to 'choose,' in the sense that it can exhibit different and appropriate behaviors in the presence of different choice

situations. We will therefore assume that its aqueous interior can undergo a sol– gel transition that allows the entity to move, which may require another – here unspecified – work cycle. We further suppose that the system can synthesize receptor molecules, which bind to food and a specific toxin, and that it can move toward the food and away from the toxin. (Obviously, this would require substantially more molecular machinery than we have described here.) In an appropriately minimal sense of 'choice,' a system of this sort would be able to exercise choice. Real cells accomplish just such choice behavior.

A variety of points should be made about this minimal system. First, for a work cycle to be performed, the system must be displaced from equilibrium; hence agency is a non-equilibrium concept. Second, the system links spontaneous and non-spontaneous processes in its work cycles. This linking has led to a biosphere in which sunlight is captured by plants and used by herbivores and carnivores to create the complex web of linked exergonic and endergic reactions within the global chemical reaction network that undergirds the biosphere. Third, the system stores energy, in particular in the excess of hexamer over its equilibrium concentration. As pointed out by P.W. Anderson (personal communication), this might later allow error correction. Fourth, the system is a perfectly legitimate coupled non-equilibrium reaction network as noted above. Fifth, such systems ought soon to be constructable and testable. In fact, they may augur a technological revolution, since they can perform work and hence build things. In sum: systems of this type are not merely philosophical examples; they are objects of current scientific research. At the same time, they exhibit the five principles which, we claim, are essential to a minimal autonomous agent, that is, the sort of agent to which one can validly apply teleological language.

Work cycles and self-organization

We turn now to some puzzles concerning work cycles, puzzles that seem to point us towards the kind of self-propagating organization that Kant was referring to in the initial quotation. To a physicist, work is force acting through a distance – a scalar quantity. Yet in any concrete case of work one finds an organization of processes that is not captured in that scalar quantity. That there is something odd about the concept of work can be surmised from the fact that an isolated thermodynamic system, say a gas in a bounding membrane, can do no work. By contrast, if the gas volume is partitioned by an elastic membrane into two parts, A and B, and if the pressure in B is higher than in A, the difference in pressure will cause the membrane to bulge into A, with the result that B does work on A. Oddly, this seems to require that the universe be divided into at least two parts for work to be done.

We find most congenial Atkins' definition of work as 'the constrained release of energy' (Atkins 1984). Picture a cylinder with a piston mounted inside and a compressed working gas at the head of the cylinder. The compressed gas has

molecules moving in all directions with a distribution of velocities, and it exerts pressure on the piston. The result is a spontaneous process that pushes the piston down the cylinder. The gas does work on the piston, and energy is released into a few degrees of freedom – the translational motion of the piston.

But this raises a new question: what are these constraints, and how do they originate? In the case just mentioned the cylinder, the piston, and the location of the piston inside the cylinder are the constraints. Where did they come from? The answer is that it took work to make these constraints. So we come to a crucial cycle: it appears to take constraints to make work, and work to make constraints. (While this may not be necessarily true, it is typically true.) It is critical to the entire discussion of emergence that biology presently lacks a theory of the organization of processes in general, and of biochemical processes in particular, and we suspect that this cycle of work– constraint– work is a part of the required theory of organization. We will return to this topic in a moment.

We are now far enough that we can begin to make sense of Kant's idea of a formative self-propagating organization communicated by the whole to the parts, though they have it not of themselves. The first concept for applying Kant's conjecture to actual biological systems is that of *propagating work*. Unfortunately, there is as of yet no formal definition of the concept. But picture a whimsical Rube Goldberg contraption: a cannon fires a cannon ball that hits a paddle wheel, setting the wheel spinning. A rope attached to the wheel is wound up around its axle, thereby lifting a pail of water from a well and tipping the pail's water into a funnel that leads to a tube. The water flows down the tube, opens a flap valve, and waters our bean field. This Rube Goldberg contraption accomplishes what we want to call propagating work. In the example a series of macroscopic changes occurs in the world, occasioned by the explosion in the cannon that sent the cannon ball flying, and the resulting changes constitute a chain of propagating work.

In this example, as in the earlier piston example, we have ignored the work involved in constructing the constraints that allow the whole assembly to function. But in attempting to understand minimal autonomous agents in biology, one cannot ignore the interrelations between work and the constraints that make it possible. In hypothetical and real cells, there is a crucial (and complex) cycle between work, constraint construction, and more work. It takes the constrained release of energy to do the work to construct more constraints to do more work – all of this depending on a controlled release of energy in a web that closes on itself until the real cell builds a copy of itself.

Consider our minimal molecular autonomous agent with some chemicals, A and B, in its aqueous phase. Let us suppose, as noted above, that building the molecule Z (that is, the lipid) required chemical work: work was done to construct the lipids, which then fell to a low energy state, the liposome. Let A and B be capable of undergoing three different reactions, to form C and D, or E, or F and G. Each reaction can be drawn in a Cartesian coordinate system with reaction coordinates on the X axis and free energy on the Y axis. Typically the substrates and products sit in their own potential wells, separated by an

energy barrier. Each of the three reactions has its own reaction coordinate × energy profile. Now let A and B diffuse into the bilipid membrane bounding our agent. In the membrane the rotational, vibrational, and translational degrees of freedom are altered compared to the aqueous phase. These alterations alter the reaction coordinate × energy profiles of the three reactions. *But this is precisely the manipulation of constraints*, for the changing heights of the energy barriers just are the constraints on the reactions. Thus our hypothetical autonomous agent has done chemical work to synthesize lipids, which fall to a low-energy bilipid membrane state that serves to modify constraints on chemical reactions. Suppose the barrier between A and B reactants and C and D products is lowered significantly when A and B enter the membrane, such that this one reaction, and not the other two (making E, or making F and G), takes place at a high rate. As long as these differentiated relations hold, chemical energy is released in constrained ways. It is obvious, then, that our agent does work to construct constraints on the release of energy. That released energy may then propagate to modify other constraints, thereby allowing more work to be done. For example, D may diffuse to a transmembrane receptor and, in real cells, activate a work-requiring transport mechanism that brings a molecule across the membrane into the cell up its own gradient.

All the pieces are now in place for completing our account of the organization of minimal autonomous agents. The example has shown that cells do self-propagating work. This includes the construction of constraints on the release of energy, work that then constructs still further constraints on the release of energy, which in turn do work as well as constructing further constraints ... and so on. The astonishing fact is that, as cells carry out this complex web of work, constraint construction, and other construction projects (such as DNA replication and enzyme synthesis), a closure is attained in which the cell finally builds a rough copy of itself. *But this whole process is precisely the self-propagating organization to which Kant pointed.* Note that self-propagating organization in this sense does not involve matter alone, energy alone, information alone, or entropy alone. It is a process that involves all these – and something more as well. It appears that this self-propagating organization, 'communicate[d] to its materials though they have it not of themselves,' is a new form of energy-matter organization in the world; it is living matter, and it is ontologically emergent. The structural and functional features we have described meet the requirement for ontological emergence: the whole has causal powers not possessed by the parts. For example, in the case of cells the whole is capable of building copies of itself, hence capable of evolution by natural selection. We need merely imagine that mutant variants of the minimal autonomous agent, or real cells such as bacteria, can have heritable variants and be selected.

Against our reductionist critics we want to stress the minimal holism of autonomous agents. There is a closure of work tasks that is completed when a cell constructs a rough copy of itself. Work tasks, like the functions of the parts discussed earlier, are a subset of their causal consequences. And, once again, physical descriptions are precluded in principle from identifying the

appropriate subset of causal consequences unless they can refer to the roles that these consequences play in the self-reproduction (and other functional aspects) of the autonomous agent. It's therefore impossible to specify the tasks and functions without discussing the entities in question *at the level of autonomous agents*, that is, biologically. This goal cannot be achieved at the level of pure physical description alone.

Notice that what is needed for comprehending minimal autonomous agents is a theory of the organism-level organization of biochemical and other processes. Unfortunately, no adequate theory of the organization of such processes currently exists in the scientific or philosophical literature, even in outline. And yet a reproducing cell does it. Had we an adequate theory of how organismic processes self-organize, we would be able to conclude something more interesting about the ontological emergence of minimal autonomous agents than the bald fact that it occurs. We return to this issue below.

Implications of the argument

One of the ways to understand and evaluate an argument is to consider what it appears to imply. We thus pause to consider three plausible extrapolations from the argument to this point. It goes without saying that extrapolations from an argument are always more speculative than the core argument itself.

The underdetermination of the biological by the physical

Darwinian evolution is neutral with respect to the – possibly indefinitely many – physical bases that might support reproduction with heritable variation and hence evolution by natural selection. It is possible that life and evolution would arise in a non-denumerably infinite family of universes similar to our own.[1] The possibility that evolution might run on multiple 'platforms' undercuts the claim that explanatory arrows from physics are necessary and sufficient for explaining biological phenomena. To say everything that can be physically said about a biological process still leaves it less than fully explained.

Darwinian adaptive evolution is agnostic with respect to the precise physical mechanisms of reproduction and heritable variation such that natural selection can act to yield evolution. We will argue that this fact means that Darwinian selection is not reducible to – and hence cannot be explained by – any specific set of lower-level explanations. Indeed, the set of possible mechanisms for reproduction and heritable variation in this specific universe is systematically

[1] The 'multiple platforms' argument is analogous to the argument for the 'multiple realizability' of mental states, although we do not here take a position on the latter. For an introduction to the 'multiple realizability' concept in the philosophy of mind, see Heil (2003) and Clayton (2004, chapter 4.)

vague, in the sense that one cannot (even in principle) pre-list all members of this set. Almost all organisms of all reproducing species are capable of reproduction and heritable variation, even though we cannot pre-state all possible modes of reproduction nor what those organisms and species will become as evolution proceeds. Given that one cannot infer the modes of reproduction and heritable variation from the Darwinian mechanisms of evolution, and given the actual experimental demonstration of more than one mode of molecular reproduction, it follows that some unknown – and in the case of new species, unknowable – number and kinds of entities are capable of reproduction and heritable variation.

It turns out, however, that the limitations on reductive explanation are actually far broader. Contemporary organisms are based on DNA, RNA, proteins, small and large metabolites, and so on. But self-reproduction is not limited to the template-based replication of DNA by protein enzymes. Recently it has been shown that peptide reaction systems are capable of autocatalytic reproduction (Lee et al. 1996; Ashkenasy et al. 2004). More generally, peptide systems are capable of what we shall call 'collective autocatalysis.' Here peptide A catalyzes the synthesis of B while B catalyzes the synthesis of A from appropriate precursors. But if two peptides can be collectively autocatalytic, why not three or a thousand? Indeed cells are, themselves, collectively autocatalytic wholes. Meanwhile, chemists are discovering other modes of molecular reproduction, including self-reproducing lipid vesicles.

Indeed, the irreducibility may even be more dramatic. Imagine that the constants of nature are real numbers and that at least a small range of those values are consistent with universes in which chemistry, self-reproduction, and heritable variation – hence natural selection – are possible. Darwin's mechanism of reproduction and heritable variation could then be realized in a non-denumerable infinity of universes. Hence, again, we cannot finitely pre-state all possible modes of reproduction and evolution, even in principle. It follows that *downward explanatory arrows in* this *universe are not explanatory of evolution in neighboring universes*. And yet that fact does not preclude us from hypothesizing that evolution by natural selection could occur in those universes.

This leads us to ask what it would mean to 'reduce' an explanation of organisms and natural selection and evolution to a physical or microphysical level. Although explanatory reduction is variously defined, most accounts agree that successful reduction – say, from some higher level L_2 to a lower level L_1 – requires that at least three conditions be met:

- the laws governing the behavior of the L_2 entities can be expressed in the language of L_1, and these laws are sufficient to explain the L_2 phenomena;
- the terms in the L_2 language are replaceable by a finitely pre-specified list of terms in the language of L_1;
- by virtue of this equivalence between an L_2 phenomenon and a set of L_1 terms, the L_1 theory and entities specify the causal mechanisms that are sufficient for explaining the L_2 phenomenon.

But meeting these conditions is just what is precluded by our incapacity, even in principle, to finitely pre-state all modes of reproduction and heritable variation. So, it seems, biology is cut off at the explanatory level from any final reduction to physics. Hence, we have no choice but to explain biological evolution as Darwin taught us to do – supplemented, of course, by those emendations to Darwin that subsequent advances in evolutionary theory have established.

Darwinian pre-adaptations

The fact that we cannot predict (i.e., deduce ahead of time) the future evolution of the biosphere might be taken to support epistemic but not ontological emergence. But Darwinian pre-adaptations appear to support the stronger, ontological construal of emergence.

It's well known that Darwin distinguished between adaptations and what he called pre-adaptations. An example of an adaptation would be a modification in the beak of a bird such that it was better fitted to the size and composition of the seeds available to it in its local habitat. The notion of a pre-adaptation is logically dependent on the concept of the function of a part of an organism. The standard account of this notion divides the causal consequences of parts of organisms into two logical subsets: those that constitute a part's (biological) function and those that do not.[2] For example, although the primary function of the heart is to pump blood, the heart also has causal consequences that are not functions, such as making heart sounds. Natural selection is said to have acted to modify those causal consequences that are the part's function. Thus pumping blood is, roughly, the causal consequence in virtue of which the heart was selected. In order to identify the function of a part of an organism, *we typically have to analyze the organism as a whole in its environment over a life cycle and in light of its specific selection pressures*. To do so is necessarily to make an evolutionary claim about selection for that particular causal consequence.

Consider the implications of the Darwinian concept of pre-adaptations. A causal consequence of a part of an organism which, in the normal environment, is not of selective significance might, in a different environment, become significant for selection. In such cases it turns out that, by virtue of the pre-existence of a specific causal consequence, the organism was *pre*-adapted to the novel environment. Many major adaptations in evolution, and even more minor ones, are thought to have been Darwinian pre-adaptations: lungs were derived from the swim-bladders of certain fish, the

[2] We speak of 'primary function' in the singular for the sake of convenience. In fact a set of functions may serve as the primary function of an organ. Still, it is never the case that *all* the causal consequences of an organ are its function.

inner ear bones that mediate hearing were derived from the jaw of an early fish, and so forth.

Incidentally, examples of pre-adaptation also play a crucial role in human technological evolution. The following story may be merely a story, but it will do. It is said that a group of engineers was trying to invent the tractor. They knew that they needed a massive motor, so they mounted a massive engine block on a chasse. The chasse promptly crumbled. They tried a bigger chasse, but again it broke. After multiple attempts one of the engineers said, 'You know, this engine block is so big, massive, and rigid that we could use *it* as the chasse and hang everything else off of it.' That, indeed, is how tractors are made. Now the rigidity of the massive engine block was a causal property of the engine block that, for normal purposes, would not be put to use. For this novel purpose, however, the rigidity of the engine block could be put to good practical use.

These examples raise a central question: is it possible to specify ahead of time – or more precisely, to finitely pre-state – all possible Darwinian pre-adaptations for the species alive today? We believe that the answer to this question is 'No.' In part, the difficulty lies in attempting to pre-state what all possible environments for organisms are. An environment for an organism, its niche, is, roughly, those aspects of its abiotic and biotic surroundings that bear on how it makes a living in the world. Organism and niche fit and mutually co-define one another. Yet note that the concept of a niche, or an environment, is systematically vague. There seems to be no way to enumerate all possible environments, and thus to specify in advance what all possible pre-adaptations might be. Hence it appears impossible in principle to finitely pre-state all possible Darwinian pre-adaptations.

Clearly, this limitation is epistemological: when such a pre-adaptation occurs, it could not have been foretold or deduced ahead of time. It thus represents *at least* a case of epistemic emergence.[3] But pre-adaptations support ontological emergence as well. It's not just that, after the fact, *we* cannot explain a given pre-adaptation in purely physical terms (though this is certainly true). Much more, before the particular causal feature began to confer selective advantage the world was such that the particular causal feature *was not yet distinguished* from among all the other causal properties of the organism and its parts.[4]

If valid, this conclusion has important consequences. For example, it implies that we cannot pre-state the configuration space of the biosphere. One cannot know ahead of time the kinds of entities, processes, and functionalities that will come to exist. This directionality suggests the existence of a distinctively

[3] This is not to deny, of course, that *after the fact* we can explain how it was that the pre-adaptation functioned as it did. Here we have a case of the well-known distinction between prediction and retrodiction, which reflects, once again, the existence of a biological arrow of time. In biology, as Broad (1925) argued, we have to see a case before we can explain it.

[4] Incidentally, note that 'all possible features of organisms' also cannot be finitely pre-specified, and for similar reasons.

biological arrow of time. Contrast this essential time-dependence with statistical mechanics. Physicists can pre-specify all possible combinations of positions and momenta in a liter of gas, and hence pre-specify the phase space of the gas system as a whole; indeed, the constructs of statistical mechanics are derived from that phase space. For a biosphere this is impossible.

Now a classical physicist might object that one could consider the solar system, say, as a classical 6N dimensional phase space, in which all possible biospheres can be represented as points or trajectories. Thus, she concludes, we *can* pre-specify the configuration space of the biosphere. For the sake of argument, let us grant her this possibility as far as it goes. Unfortunately, however, there seems to be no way of pre-stating the *collective* variables – the lungs, wings, mitochondria, and so forth – that will come to exist in the ongoing evolution of the biosphere. Note that it is precisely those collective variables that play a *causal role under selection* in the ongoing evolution of the biosphere, and hence play the key explanatory role for what comes to exist in the biosphere. This fact suggests, again, that Darwinian pre-adaptations are ontologically emergent. The collective variables should be taken as entities at their own level associated with their own causal powers, e.g., it was lungs that allowed amphibians and later taxa to invade the land.

For the reasons given, pre-adaptations resist analysis in terms of the familiar frequency theories of probability. When one asserts that a fair coin will tend to land on its head about 50% of the time, one knows ahead of time the full configuration space of, say, 10,000 coin flips. But we do not know the macroscopic collective entities that will arise in the evolution of the biosphere. We do not even know, ahead of time, what entities will be involved in causal interactions. So we cannot make the familiar frequency-based probability statements about 'the future of the biosphere in general.'

The role that Darwinian pre-adaptations play in evolution forces us, then, to accept the thesis of *ontological emergence*. For the new adaptation surely is a newly existing entity, a new functionality; it could not be pre-specified, even in principle; and it has causal powers of its own (wings allow the creature to fly). In science in general, what plays a causal role (say, under selection) also plays an explanatory role: it determines what comes to exist in the biosphere. Those things that are causes and are essential for scientific explanations are things that scientists are justified in treating as existing. Hence the emergence in question is ontological.

Downward causation

Is the claim that explanatory arrows always point upward correct? We argue that it is not. Consider a concrete case. Most species that ever lived are now extinct. Now when a species goes extinct, there is some last living member of that species. When that member dies, it dies as a whole biological organism, not as a mere aggregate of its (physical and chemical) parts. But when it dies,

the specific molecules, proteins, genes, RNA, and so forth that were particular to that species vanish from the further unfolding of the biosphere. That is, the causal account of the extinction of the species requires one to employ the laws of population genetics and ecology, and perhaps also to give a historical account of the conditions of the abiotic environment. By virtue of the death of that organism, its unique molecules are lost from the future unfolding of the biosphere. Hence, the extinction of species alters the molecular makeup of the biosphere.

This case provides a simple, concrete instance of downward causation. An unknown diversity of genes, proteins, and small molecules are no longer present in the biosphere because the organisms of species carrying them no longer proliferate. Notice also that, in order to explain this case of extinction, we need the concept of species and, for sexual species, also the notion of the reproductive isolation of species. Yet, at the molecular level of detail, there are *indefinitely many ways* that species can be reproductively isolated from one another. Here again we confront the multiple platforms issue. We cannot finitely pre-state the mechanisms of reproductive isolation, and thus we cannot carry out the reduction of species-and organism-level language and explanation to a lower level of laws and entities. The mechanisms of species isolation from one another, and the population-ecological level explanations of extinction, are just not reducible in this sense. In living and dying, organisms exercise a causal influence on their parts – an influence that is well described as 'downward causation,' in what we take to be a completely non-mysterious sense of the term.

Species extinction pertains to whole organisms, yet it changes the molecular future of the biosphere. The laws that govern species evolution, including extinction, are not identical to the laws of physics or to explanations at the physical level alone. This particular form of downward causation from biology to physics is a necessary condition for the stronger forms of downward causation debated by philosophers of mind, but one can accept the former without accepting the latter.

Areas for further research and reflection

Teleonomic vs. teleological ascriptions

It is obvious that our minimal molecular autonomous agent fulfills the teleonomic requirements for 'acting as if' and 'referring' and 'choosing'; and if it does, real cells do as well. One need merely add selection and heritable variation to our account to achieve a model of what real cells do. But teleonomic talk is not yet teleological. It is one thing to have the appearance of goal-directed behavior; it is something else to actually engage in goal-directed behavior.

We suggest that minimal molecular autonomous agents are the minimal systems to which one can justifiably attribute referring, choosing, and acting. We are hesitant to make this claim as a metaphysical assertion, since it's not clear how one would 'prove' the appropriateness of playing the agency language game. But the minimal molecular autonomous agent seems to be about as minimal as one can get and still plausibly use the language of agency. Obviously, agency starts at some level of organization within the biosphere. Humans are agents – we plan and carry out actions consciously – and rocks are not; they are involved in cause-and-effect interactions but do not carry out actions. We take agency to be a matter of degree: some minimum conditions must be met for one to ascribe it at all, and then it increases (roughly, as a function of the increase in organizational complexity through evolution) until one encounters full, robust, conscious agency.

At least two paths exist for defending the minimal ascription of agency, and both have already been introduced. On the one hand, an agent is an organism that is faced with a choice situation (it may act this or that way, to achieve this or that end) and then acts to achieve an end (e.g., it moves toward the food source or away from the poison). On the other hand, an agent is an interpretant in a system of semiotic signaling, that is, a system that includes signs, interpretations of the signs, and actions in accordance with the interpretations.[5]

A general biology

Astrobiology is a major growth field within the biological sciences. Astrobiologists attempt to understand biospheres anywhere in the cosmos. In *Investigations* Kauffman (2000) rather boldly proclaimed four candidate laws for any biosphere. The first candidate law arises out of work on parallel processing non-linear dynamical systems, such as random Boolean networks and their continuous variable cousins, which shows that such systems behave in two broad regimes, ordered and chaotic, with a phase transition, sometimes called the 'edge of chaos,' separating the two regimes.

It is not important for present purposes whether all four candidate laws are actually true. Let us assume for the sake of argument that at least the first is true and ask whether, if so, it is reducible to physics. Kauffman and others (Kauffman 1993) have argued that the ordered vicinity at the edge of chaos is a favorable location for the dynamics of autonomous agents. Here the most complex coordinated behavior can occur. Thus it is an attractive hypothesis

[5] We cannot here summarize the recent applications of semiotic theory to biological systems, which, despite some critical reservations, we find intriguing and potentially empirically fruitful. For an introduction to the field see Weber (2003), Emmeche et al. (2002), Taborsky (1999), and Hoffmeyer (1996). Of course, for cells it is natural selection that produces the object-sign-interpretant relationship rather than the socially much more complex way in which humans achieve reference to objects using signs.

that agents will evolve – that they will learn to play the increasingly complex natural games that agents play to earn a living – by tending to move to the ordered vicinity at the edge of chaos. Assume for the moment that this hypothesis is true. Now note that, like Darwinian natural selection, the hypothesis is systematically vague about the physical mechanisms of action of the non-linear variables such that the systems they comprise will lie in the ordered regime near the phase transition. Thus, again, it appears to be impossible in principle to pre-specify the necessary and sufficient physical conditions such that autonomous agents would evolve to the ordered vicinity of chaos. Furthermore, and importantly, the *laws that govern the whole are not to be found in any specific physical realization of such a system*, but rather in the mathematics of this broad class of dynamical systems, *whatever* their material realization. Put differently, the laws exist at the level of the collective variables, of whatever physical character, that comprise the system as a whole. Here too, it appears, we are confronted with a case of ontological emergence.

The need for a theory of the organization of processes

The holism and task closure that occur in the reproduction of a minimal autonomous agent – and hence, to be more concrete, in most living cells – involves an organization of processes. Clearly, if a cell is to be evolutionarily successful, one of its main tasks must be reproduction. Yet as we showed, not all the causal consequences of its parts are implicated in the causal and functional account of the closure of the reproductive task that the cell carries out.

No adequate theory of organization currently exists. The required but not yet available theory of organization will involve more than merely a discussion of entropy. Entropy counts the number of configurations a system occupies, but it is not concerned with constraints. As we saw, central to cell functioning is the use of energy to do work to construct constraints on the release of energy, which is work that propagates to construct more constraints on the release of energy, leading to other construction tasks that finally close upon themselves in the reproduction of an autonomous agent, a living cell.

We may not yet be able to formulate an acceptable theory of organization. Still, it may be helpful to speculate about some of the features theory-candidates should have. Several clues are available. First, in the case of full-fledged human action we know (and need to know) both what *overarching tasks* and *linked tasks* are. We know what it is to plan and prepare dinner, eat dinner, and clean up after dinner. Implicit in this knowledge is a tacit knowledge of tasks and other events that are linked to the overarching tasks, as well as causal factors that are not essential to them. We peal the carrots. How we do so – whether, for example, we do so over the counter or over the sink – is not crucial in specifying what dinner preparation entails. Knowledge of tasks, then, identifies a subset out of the sum of all causal consequences and events; that

subset is essential for specifying the tasks, whereas a myriad of other causal factors are not essential. This fact suggests that any theory of organization must have a way to pick out subsets of causal consequences and events that are relevant for accomplishing biologically essential tasks.

One wants to know, second, which subset of causal consequences is relevant to distinctively biological explanations. We have made the standard biologist's argument that, at the level of parts of organisms, one must pick out those tasks or functions upon which selection has acted. This too is a clue: in biology no theory of organization will be adequate that does not incorporate the core principles of evolutionary theory.

A third clue lies in the concept of the cycle of work: the constrained release of energy, and the use of work to construct those very constraints. Interestingly, work seems to include its own distinction between relevant and irrelevant causal consequences. Consider again the cylinder and piston with hot gas in the cylinder head. During the power stroke the gas does work on the piston, causing transversal motion. But the gas also exerts pressure – momentum transfer – to the (rigid) cylinder. We distinguish the former as relevant and the latter as irrelevant to work done. A theory of the organization of process must be able to make the same sort of distinction.

Undoubtedly there are other clues. Signals are low mass-energy flows that unleash larger mass-energy flows in the organization of cells and human activities. It seems likely that any theory of organization must include signals and the semiotics of signals. If so, this hints that some concept of reference must play a role in a theory of organization. But how? Perhaps it is that the sign–interpretant system itself picks out the relevant from the irrelevant aspects of a situation, object, or feature of the object to be interpreted by the interpretant. How these clues fit together is at present unclear. Dare one say that 'goals' or 'purposes' may enter into the task of distinguishing between relevant and irrelevant causal consequences? At a meta-theoretical level it seems clear that the theory of organization needs to be able to specify goal-directed processes, even if, at present, we don't know what this would mean for actual biological theories at the level of minimal autonomous agents. In full-fledged human design (as in architecture, engineering, and other arenas) the goal or purpose of the artifact distinguishes between the causal consequences that are relevant and irrelevant. In evolution the 'purpose' of the heart is to pump blood; biologists then provide selection-based explanations and interpret the 'purpose' teleonomically. It is therefore possible that future theories of biological organization will need to be restricted to evolved entities and purposive acts and constructions. But this suggestion remains speculative for now. The fact that work itself distinguishes relevant from irrelevant causal consequences may suggest that a natural theory of self-organization can be created using the work concept. Getting clear on this particular issue would be an important step toward a broader theory of organization.

Summary

We have argued that Darwinian natural selection is agnostic with respect to the indefinitely many mechanisms by which reproduction and heritable variation can arise within this or that member of a small family of similar universes. We suggested that reduction requires formulating descriptions and laws in a lower-level language that are necessary and sufficient for the truth of the higher-level language and explanatory of the phenomena at that higher level. Given what we called *systematic vagueness* – the incapacity in principle to pre-state in physical terms the mechanisms that would fulfill Darwin's requirements – we concluded that neo-Darwinian evolution cannot be formally reduced to physics alone.

We also argued that extinction is a case of downward causation: when the last member of a species dies, it dies as a whole organism; yet at the same time its death alters the molecular makeup of the biosphere. The physicist must appeal to laws and phenomena at the level of populations of organisms if she is to explain extinction. Finally, we showed that one cannot finitely pre-state all possible Darwinian pre-adaptations because (among other reasons) the concept of an environment in which an organism makes a living is systematically vague and hence cannot be reduced, even in principle, to a lower-level language. Physics does not contain what turn out to be the essential theoretical terms for conceptualizing the relationship between organisms and environments, such as fitness and selection pressures.

All of these conclusions appear to fulfill the conditions for ontological emergence: biological entities have causal powers, such as the capacity to evolve, that pre-biological physical entities do not have. We have argued that agency, which is present in full-blown human action, can be meaningfully and fruitfully ascribed to minimal autonomous agents as well. It would follow that living organisms represent a new form of matter, a new instance of the organization of processes, that fulfills Kant's dicta and is thus ontologically emergent. We close by reemphasizing that what is needed to fully understand biological agency has not yet been formulated: an adequate theory of organization. Such a theory may well be the missing link in the contemporary emergence vs. reductionism debate.[6]

References

Anderson P.W. 1972. More is different: broken symmetry and the nature of the hierarchical structure of science. Science 177: 393–396.

Atkins P.W. 1984. The Second Law. Scientific American Books, New York.

Ashkenasy G., Jagasia R., Yadav M. and Ghadiri M.R. 2004. A self-organized synthetic chemical network, submitted.

[6] We are grateful to Kim Sterelny and to an anonymous referee for this Journal for insightful philosophical criticisms that have improved the quality of the final product.

Bishop J. 1983. agent-Causation. Mind N.S. 92: 61–79.
Broad C.D. 1925. The Mind and Its Place in Nature. Routledge & Kegan Paul, London.
Chisholm R.M. 1976. The agent as cause. In: Myles B. and Douglas W. (eds), Action Theory, D. Reidel, Dordrecht.
Clayton P. 2004. Mind and Emergence: From Quantum to Consciousness. Oxford University Press, Oxford.
Daley A.J., Girvin A., Kauffman S.A., Wills P.R. and Yamins D. 2002. Simulation of chemical autonomous agents. Z. Phys. Chem. 216: 41–49.
Emmeche C., Kull K. and Stjernfelt F. 2002. Reading Hoffmeyer, Rethinking Biology. Tartu University Press, Tartu.
Heil J. 2003. Multiply realized properties. In: Sven W. and Heinz-Dieter H. (eds), Physicalism and Mental Causation: The Metaphysics of Mind and Action, Imprint Academic, Exeter.
Hoffmeyer J. 1996. Signs of Meaning in the Universe. Indiana University Press, Bloomington(trans. Barbara J. Haveland).
Kant I. 1987. Critique of Judgment. Hackett, Indianapolis(trans. Werner S. Pluhar).
Kauffman S. 1993. Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, New York.
Kauffman S. 1996. At Home in the Universe: The Search for Laws of Self-Organization and Complexity. Oxford University Press, New York.
Kauffman S. 2000. Investigations. Oxford University Press, New York.
Lee D.H., Granja J.R., Martinez J.A., Severikn K. and Ghadiri M.R. 1996. A self-replicating peptide. Nature 382: 525–528.
Laughlin R. 2005. A Different Universe: Reinventing Physics from the Bottom Down. Basic Books, New York.
McLaughlin B. 1992. The rise and fall of British emergentism. In: Beckerman A., Hans Flohr and Jaegwon Kim (eds), Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism, Walter de Gruyter, New York.
Monod J. 1971. Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology. Wainhouse, Knopf, New York(trans. Austryn).
O'Connor T. (ed), 1995. Agents, Causes, and Events: Essays on Indeterminism and Free Will. Oxford University Press, New York.
Taborsky E. (ed), 1999. Semiosis, Evolution, Energy: Towards a Reconceptualization of the Sign. Shaker Verlag, Aachen, Germany.
Weber A. 2003. Natur als Bedeutung: Versuch einer semiotischen Theorie des Lebendigen. Königshausen & Neumann, Würzburg, Germany.
Weinberg S. 1992. Dreams of a Final Theory. Pantheon, New York.
William W.C. 1994. The ontology of complex systems: levels of organization, perspectives, and causal thickets. Can. J. Philos. 20: 207–274(Supplementary).
von Wright G.H. 1971. Explanation and Understanding. Cornell University Press, Ithaca.