The Meaning and Origin of Goal-Directedness

Heylighen, Francis

Published in:
Biological Journal of the Linnean Society

DOI:
10.1093/biolinnean/blac060

Publication date:
2023

License:
Unspecified

Document Version:
Accepted author manuscript

The meaning and origin of goal-directedness: a dynamical systems perspective

FRANCIS HEYLIGHEN

Center Leo Apostel, Vrije Universiteit Brussel, *Pleinlaan 2, 1050 Brussels, Belgium*

This paper attempts to clarify the notion of goal-directedness, which is often misunderstood as being inconsistent with standard causal mechanisms. We first note that goal-directedness does not presuppose any mysterious forces, such as intelligent design, vitalism, conscious intention or backward causation. We then review attempts at defining goal-directedness by means of more operational characteristics: equifinality, plasticity, persistence, concerted action and negative feedback. We show that all these features can be explained by interpreting a goal as a far-from-equilibrium attractor of a dynamical system. This implies that perturbations that make the system deviate from its goal-directed trajectory are automatically compensated—at least as long as the system stays within the same basin of attraction. We argue that attractors and basins with the necessary degree of resilience tend to self-organize in complex reaction networks, thus producing self-maintaining "organizations". These can be seen as an abstract model of the first goal-directed systems, and thus of the origin of life.

ADDITIONAL KEY WORDS: equifinality – plasticity – persistence – concerted action – negative feedback – attractors – basins – resilience – self-maintenance – origin of life.

INTRODUCTION

There is a long-standing controversy over whether the notion of purpose or goal properly belongs in a scientific theory (Deacon & Sherman, 2007). The standard ontology of science is causal: it assumes that the present behavior of a system is fully determined by causes that lie in the past, including the system's previous state and any forces or inputs that act on the system's state. Therefore, there does not seem to be any room for a goal, which lies in the future, to affect the behavior here and now. Moreover, the application of goal-directedness to biological systems has gotten an ill repute because it has been associated with a number of explanations that are incompatible with our present understanding of life, including creator-imposed purpose, intelligent design, a mysterious "life force", and the assumption that goal-directed behavior requires conscious intention.

Yet, in practice both scientists and laypeople liberally use the notion of goal-directedness, because it provides a simple and useful explanation for common phenomena. If you see a person assembling ingredients in the kitchen, then you can safely assume that the purpose is to prepare a meal. If a cheetah runs after a gazelle, its goal is clearly to kill and eat that gazelle. All the actions that the cheetah undertakes during its hunt, such as accelerating, jumping on the gazelle's back or biting the gazelle's throat, can be understood by assuming that they are *directed* at

this specific goal. That interpretation provides a concrete, falsifiable prediction of what will happen when the cheetah has finally caught its prey.

Aristotle summarized this type of explanations with his notion of *final cause*. The cause or reason for the cheetah running after the gazelle is that this will result in the cheetah eating the gazelle. The cause is "final" in the sense that it forms the ultimate outcome of the process to be explained: once the cheetah has eaten the gazelle, the process has reached its "end". Such explanation is called teleological, from the Greek *telos*, which means "end". However, this Aristotelian interpretation seems to contradict our present notion of causality, in which causes necessarily *precede* their effects. Therefore, scientists typically avoid any notion of teleology in their theories.

Still, most features of biological, mental and social organization are intuitively understood as having a function, goal or purpose, i.e. as being there "for" the achievement of some future state. For example, the cheetah's claws and canines clearly seem to be there in order to help the cheetah kill its prey, and they are present long before the cheetah cub has learned to hunt. Thus, it seems as if this future state of killing prey shapes the anatomy and behaviour of the young cheetah here and now. To explain this feature without needing to rely on final causes, biologists have introduced the concept of *teleonomy* (Corning, 2022), so as to emphasize that this purposefulness is intrinsic to the organism, rather than imposed by some external cause. This inherent purpose can be understood as a product of evolution. The idea is that natural selection has eliminated all cheetahs that were poor at hunting. Therefore, the remaining ones are born with the features necessary to hunt, including the instinctive desire to seek for, run after and kill prey.

However, this notion does not really explain the precise mechanisms underlying goal-directedness. Moreover, some confusion remains about what precisely distinguishes *teleonomy* from *teleology*--the term most scientists would rather avoid (Thompson, 1987). A common understanding (or misunderstanding) is captured by the definition of teleonomy in the Merriam-Webster Dictionary as "apparent purposefulness of structure or function in living organisms due to evolutionary adaptation". Thus, some scientists merely seem to use the term 'teleonomy' when they would actually like to speak about goal-directedness, but do not dare to do so explicitly, covering themselves by noting that the observed purposefulness is merely "apparent". Such an interpretation does not explain what distinguishes a purposeful from a non-purposeful organization, nor how a process can be effectively directed towards an as yet non-existing goal state.

Therefore, in the following I will mostly avoid the term "teleonomy", replacing it by the more explicit concept of *goal-directedness*. By this I mean that the behavior of a system is consistently directed towards a particular state that we can interpret as the system's goal. More concretely, that means that if we extrapolate the effects of that behavior towards the future, then the resulting trajectory will end up in the goal state, and this even if various intermediate disturbances make that movement deviate from its initial trajectory. Thus, the statement that a system is goal-directed is in principle testable: it is sufficient to perturb the system's behavior (e.g. by moving the cheetah to a different position within its broad field of vision), and check whether its new trajectory will still converge on the same goal state (e.g. eating the gazelle).

What we call the goal of course depends on which variables the observer uses to model the system's state. The goal of "eating the gazelle" may alternatively be formulated as "satisfying the cheetah's hunger" or "reducing the deficiency in the cheetah's energy metabolism", depending on the level of description. But these

descriptions are operationally equivalent insofar that they make the same predictions about the cheetah's behavior leading up to this goal state.

The present paper intends to further clarify this concept of goal-directedness and show that it is perfectly compatible with causal mechanisms. I will first briefly review the many improper or poorly understood uses of this notion, and explain why scientists reject them. I will then review a number of attempts to come to a more operational understanding of goal-directedness, and propose a definition that integrates them. This will show that goal-directedness understood in this way is easily modeled and explained by interpreting a goal as a far-from-equilibrium attractor of a dynamical system. I will then briefly show how such attractors can self-organize in networks of reactions, producing the equivalent of a self-maintaining or autopoietic system. Such a system is equivalent to an autonomous agent, or a rudimentary living system. Thus, I intend to show that there is no contradiction between goal-directedness and causality, and in particular that goal-directed organization can emerge from simple causal processes. This, I hope, will set the stage for developing a true science of purposive systems that will help us to understand life, mind and behavior.

PROBLEMATIC INTERPRETATIONS OF GOAL-DIRECTEDNESS

Before I introduce a coherent notion of goal-directedness, it is worth reviewing why scientists have been so reluctant to accept this concept as an adequate description of natural, and in particular biological, phenomena.

A first reason is that in the past the notion of purpose has been closely associated with supernatural conceptions, in which the world and our actions in it are explained by some kind of divine plan or *intelligent design*. This plan imposes a trajectory for the evolution of the universe that is directed at some desired end. Here, everything that occurs is explained by assuming that God (or some other, unspecified intelligent force) intended it to happen in this way. For example, God gave wings to birds because He wanted them to fly. By flying, these birds are fulfilling God's purpose. And so are we humans, when we worship, work hard, and support our family.

Scientists correctly noted that anything can be explained by attributing some goal or intention to an invisible power whose full motives remain inscrutable. If a village is destroyed by an earthquake, then it is always possible to come up with an assumed purpose for this event: perhaps the villagers were punished by God for transgressing some norm, or perhaps God was merely testing their faith. Such explanations cannot be confirmed or falsified, and have zero predictive power. Therefore, they do not belong to the realm of science.

A potentially more scientific explanation of biological goal-directness was proposed by *vitalism* (Walsh, 2018). Vitalists assume that living organisms are driven to act towards particular targets by some inherent "life force" that is absent in non-living systems. That force would explain how organisms grow, develop and self-organize, while producing highly specific ordered structures and behaviors. This is what sets them apart from purely physical systems that are subject to the second law of thermodynamics, which implies loss of order, dissipation and diffusion rather than targeted build-up of organization.

One reason vitalism is no longer accepted as a viable scientific hypothesis is that no such vital force has ever been observed. More precisely, through manipulations at the level of molecular biology, we have blurred the line between living and non-living

systems, without finding a point where this vital force becomes operative. Another reason is that we can now satisfactorily explain how self-organizing systems build order by exporting entropy to the environment, i.e. while still obeying the second law. Finally, postulating some mysterious force that no one as yet understands suffers from the same shortcomings as an explanation based on intelligent design: without some mechanism specifying how and under what conditions that force produces its assumed effects, the theory lacks predictive power.

A third reason scientists have been skeptical about goal-directed explanations is their apparent anthropomorphism. Most people see human action as being directed by conscious intentions, in which a person conceives of some desired goal state, and then plans a course of action to realize that as yet purely mental conception. It seems unlikely that lower animals, plants, bacteria or embryos would have an explicit representation of the goal state they are targeting. It is even more implausible that they would consciously reflect on how to achieve it. Yet, as I will show further, it is not necessary for a system to have an explicit internal representation of an end state in order to dependably steer towards that state (cf. Brooks, 1991; Trestman, 2012). There is in particular no need for anything resembling human intelligence or consciousness.

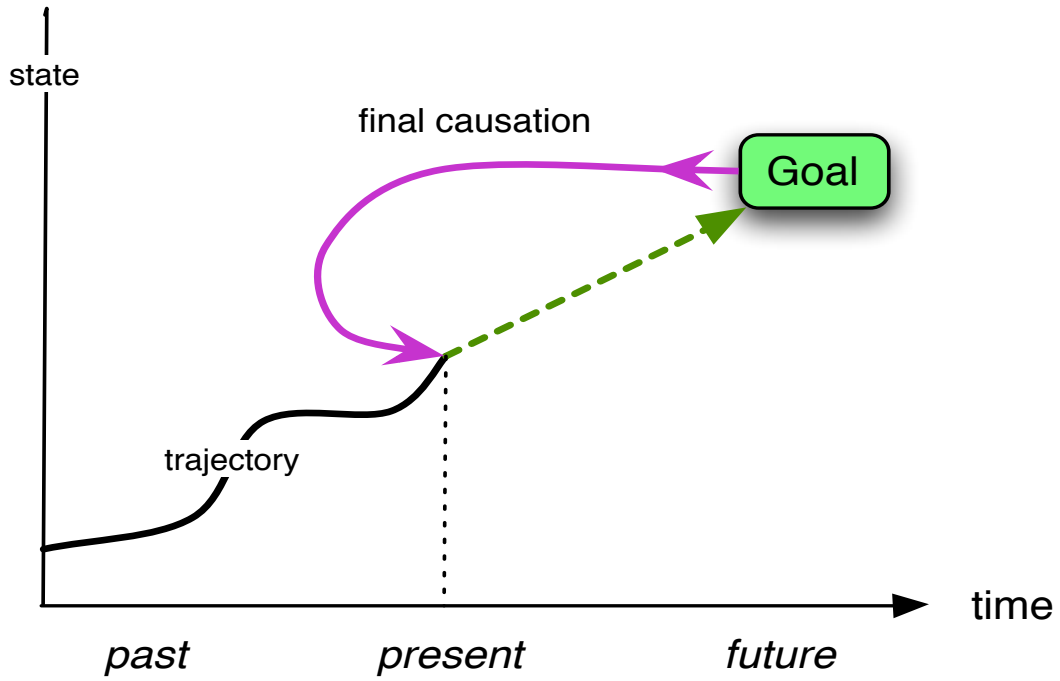

Fig. 1: final causation interpreted as an influence (purple arrow) going backward in time from the future goal to the present state, an interpretation incompatible with our understanding of causality.

Another reason for skepticism is the Aristotelian interpretation of goals as "final causes", i.e. causes that lie in the future. Science assumes that the present, actual state of a system determines it future states, including any "end" state. But such a causal influence does not extend in the opposite direction. A future state cannot reach back to its own past in order to guide the present towards itself (see Fig. 1). Otherwise, we would be confronted with the classic paradox of the time machine (Heylighen, 1990).

Such a device would allow you to travel back to the past and causally affect it. If this were possible, you could for example prevent your mother from meeting your father, thus invalidating your own existence. Therefore, teleological explanations interpreted as time-reversed causality are not compatible with our understanding of space and time.

Yet another criticism derives from the deterministic interpretation of causality that is at the basis of classical, Newtonian mechanics. According to this philosophy, which was developed most explicitly in physics but has been implicitly assumed by most other sciences, the initial conditions (causes) of a process *completely* specify its further evolution (effects). Thus, if you know the initial state of a system, then you can in principle perfectly predict its future trajectory and eventual end state. The argument then is that if the behavior of a system is already fully determined by its present state, then there is no room for some putative goal to affect that behavior. That implies that goal-directed explanations of a system's behavior are either inconsistent with the more fundamental causal explanation, or at best superfluous, by merely confirming what could already be derived from the causal explanation (cf. Walsh, 2012).

This criticism of goal-directedness, however, overlooks two important observations. First, even if we would assume a Newtonian dynamics with completely specified trajectories, then the argument of the initial state fully determining the final state can be turned upside-down. When there is only a single possible trajectory going through these two states, then the initial state is just as much determined by the final state as vice-versa. Newtonian dynamics is fully reversible: it does not care about the arrow of time (Heylighen, 1989). Therefore, you can in principle reconstruct the past from the future in the same way that you would infer the future from the past. That means that you could as well explain the present behavior of a system by starting from its predicted end state or goal. Thus, in a truly Newtonian universe, final causes are in an abstract sense equivalent to initial causes. Here, there is no paradox of the time machine that you use to prevent your mother from meeting your father: in a deterministic universe, where everything is preordained, you simply cannot prevent anything from happening, neither in the future, nor in the past.

A more fundamental observation is that modern physics has long abandoned determinism. The Heisenberg uncertainty principle has shown that microscopic events are fundamentally indeterminate. The fact that most physical systems are non-linear moreover implies that microscopic fluctuations too small to observe can grow into macroscopic differences large enough to change the course of history—a property known as "sensitive dependence on initial conditions" or more colloquially "the butterfly effect" (Hilborn, 2004). Therefore, as testified by the science of complex systems (Heylighen, 2009), in the real world most processes are largely unpredictable. Biologists, ecologists, psychologists and sociologists of course know that they can never hope to build deterministic models of the behavior they observe. Yet, some might still harbor the illusion that at the level of atoms and particles, processes are determined. Nevertheless, physics has conclusively demonstrated that that is not the case (Barrow, 1998). Therefore, causal models are necessarily incomplete. That creates room for goal-directed models to extend them in order to make predictions that strictly causal models cannot make.

ATTEMPTS AT DEFINING GOAL-DIRECTEDNESS

Because of the conceptual difficulties and misinterpretations we reviewed, the notion of goal-directedness or purpose has long been considered to be outside the realm of science (Deacon & Sherman, 2007; Deacon, 2011). Yet, the world is full of systems, such as bacteria, cheetahs, people and corporations, which behave as if they are striving to achieve some as yet distant goal state. We will now briefly review different proposals for defining goal-directedness or teleonomy in a more operational manner, i.e. without implying some mysterious design, conscious intention or backward causation.

The biologist Ernst Mayr interpreted teleonomy as the operation of a system following an internal program that leads to a specified result (Mayr, 1974). For example, an embryo develops into a particular complex morphology by following the instructions in its genetic program, and the cheetah runs after the gazelle because of its genetically programmed instincts. However, a program is merely a fixed sequence of deterministic steps. Therefore, it is not clear what distinguishes it from an ordinary cause-and-effect mechanism.

A more recent proposal by McShea situates goal-directedness not inside the organism, but in an external "field" that directs the behavior of the system (Babcock & McShea, 2021; McShea, 2012). For example, bacteria hone in on a food source by sensing the concentration of food molecules across space, and then adjusting their movement so as to follow the gradient that leads to the highest concentration (chemotaxis) (Sourjik & Wingreen, 2012; Wadhams & Armitage, 2004). The problem here is that this field acts like a physical force, similar to gravitation or electromagnetism, which deterministically controls the trajectory of the system, just like in Newtonian mechanics. Neither of these proposals seems to pay due respect to the autonomy or agency that characterizes a goal-directed system, i.e. its ability to control its trajectory independently of (internal or external) forces.

Perhaps the simplest operational definition of goal-directedness is *equifinality*. This is the notion that different initial states of a system all lead to the same final state (Lyman, 2004). For example, the bacterium will end up at the food source and the hunting cheetah at the gazelle no matter which position they started out from within that goal state's broad neighborhood. The biologist and systems theorist Ludwig von Bertalanffy proposed that equifinality is what distinguishes living systems from mechanical, physical ones (Bertalanffy, 1969). The biologist Hans Driesch even interpreted this property as an argument for vitalism, following his experiments in which he manipulated embryos in different ways while observing that this did not stop them from developing the same final anatomy. Indeed, in a purely Newtonian dynamic, distinct initial states (causes) necessarily produce distinct future states (effects)—a property which in earlier work I have called "distinction conservation" (Heylighen, 1989).

However, this is no longer true for the irreversible dynamics implied by thermodynamics. Here, initial differences tend to be erased by the inexorable increase of entropy. For example, a ball that rolls down into a bowl will come to rest in the lowest point at the bottom of the bowl, having dissipated its energy of motion through friction. It does not matter where in the bowl (initial state) the ball started its trajectory: the end point will always be the same equilibrium state of minimal potential energy. This type of equifinality is not what we would intuitively see as goal-directedness, though.

This problem can be avoided by demanding that the final state would be a steady state "far-from-equilibrium", i.e. a state that requires active intervention and a continuing flow of energy to reach and maintain. Indeed, life is a far-from-equilibrium condition, and organisms cannot survive without a constant source of energy to keep their metabolism going. This also captures our intuitive notion of agency as being *active*, i.e. mobilizing energy or resources in order to combat physical forces such as dissipation. The ball in the bowl, on the other hand, is merely passively obeying the external forces of gravity and friction. That is why von Bertalanffy restricted equifinality to open systems that exchange resources with their environment (Bertalanffy, 1969; Lyman, 2004).

The philosopher Nagel tried to define goal-directedness in a similar vein (Nagel, 1977). Following (Sommerhoff, 1950), he distinguished two requirements for goal-directedness: *plasticity* and *persistence*. Plasticity is roughly equivalent to equifinality. The agent is "plastic" in the sense that it can adapt to a wide variety of initial conditions and still find a way to end up at the same goal. It is moreover "persistent" in the sense that if that movement is perturbed, so that the agent is made to deviate from its anticipated trajectory, it will establish a new trajectory pointing towards the same target, thus persisting in its movement towards the goal. Note that it can be argued that persistence is not an independent condition, but merely an aspect of equifinality: the deviation merely pushes the agent to a different initial state, from which it will now again hone in on the same end state. For a goal-directed trajectory, it does not really matter whether the initial state was the product of a perturbation or not.

As illustrated by the work of Lee & McShea (2020), plasticity and persistence can be operationalized to assess whether the behavior of some concrete system, such as a bacterium or a cheetah, is goal-directed: you just need to measure to what degree that behavior compensates for perturbations or for changes in initial state. Such an operationalization may provide a more precise, quantitative test of a model that assumes goal-directedness.

Nagel (1977) added a third, less well-defined criterion, of which different versions have been echoed by different authors. The idea is that agents make use of a repertoire of complementary actions, components or functions that work together in a *coordinated* manner in order to achieve the goal. This is clearly inspired by biological organisms, where a variety of organs, metabolic processes and functions play their role in goal-directed behavior. For example, the cheetah hunting a gazelle relies at least on eyes, legs, claws and teeth, in addition to various internal organs, instincts and processes, such as respiration and blood circulation, in order to perform the actions necessary to capture and kill the gazelle. Trestman (2012) calls this requirement "dynamic coherence". A perhaps better term is what I will henceforth call *concerted action*: the activities of the different components of the system work together in a driven, coordinated manner, so as to efficiently mobilize the resources needed for the achievement of the goal state.

The problem with this criterion is that it is more difficult to operationalize: how many components are needed, and in what way do they need to interact for the observer to conclude that there is concerted action towards the goal? In my own model of the origin of goal-directedness, I will further suggest how such coordinated activity could be formalized. Here we can already note that this requirement is compatible with the one of a far-from equilibrium end state, because reaching such a state requires the mobilization (i.e. coordinated, driven, non-entropic deployment) of energy-providing resources.

A final, and perhaps most elegant, way to operationally define goal-directedness is the one proposed by cybernetics (Heylighen & Joslyn, 2003). The cybernetic conception assumes equifinality and persistence, but proposes a precise mechanism through which these are realized: regulation through *negative feedback*. This mechanism moreover clarifies the paradox of final causation, as first suggested in a classic paper entitled "Behavior, Purpose and Teleology" (Rosenblueth *et al*., 1943). Cybernetics introduced the notion of *circular causality* to explain how a goal can causally affect an initial state without contradicting the fact that causes precede effects.

Although apparently paradoxical, circular causality can be implemented simply by feeding the output (effect, result) of some system or process back to its input (cause, initial condition). This is also known as *re-entry*, and is the basis of what in mathematics and computer science is known as recursion. As I will illustrate further, circular organization, in the sense of the system producing (part of) its own input, makes that system to some degree *autonomous*, i.e. able to maintain its own dynamics without being at the mercy of external inputs that it does not control. This crucial feature of self-determination can also be seen as a form of *bootstrapping*: it is as if the system lifts itself up (movement towards the goal) by pulling (output action) on its own bootstraps (input component).

Note that this circularity occurs in space, not in time: the output is redirected to affect the same component or variable of the system that was affected by the input, i.e. where the process started. But that feedback of course cannot arrive *before* the initial input—though it can be nearly simultaneous, allowing for a practically instantaneous regulation of the input.

For the simplest form of cybernetic regulation, you merely need a feedback that is *negative*. That means that the effect of the re-entered output is opposite to the effect of the input that caused it. In practice, this means that if some perturbation (input) has made the system deviate from its goal-directed course, then the reaction of the system (resulting output fed back to become a new input) will counteract or compensate that deviation, thus putting the system back on track towards achieving the goal (see Fig. 2).

The classic simple example of a goal-directed system in cybernetics is a thermostat. If the temperature T of a thermostatically controlled room (sensed input to the system) goes down below the goal temperature T_0, then the thermostat reacts by switching on the heating (output caused by this input). This makes the temperature go up again, thus neutralizing the effect of the perturbing input (negative feedback). If the temperature T would increase further above the goal (new input), the thermostat will switch off the heating (new output), thus making the temperature go down again (negative feedback in the opposite direction). In summary, the thermostat has been designed to obey the following condition-action rules:
if $T < T_0$ (condition), *then* switch on heating (action);
if $T \geq T_0$, *then* switch off heating.

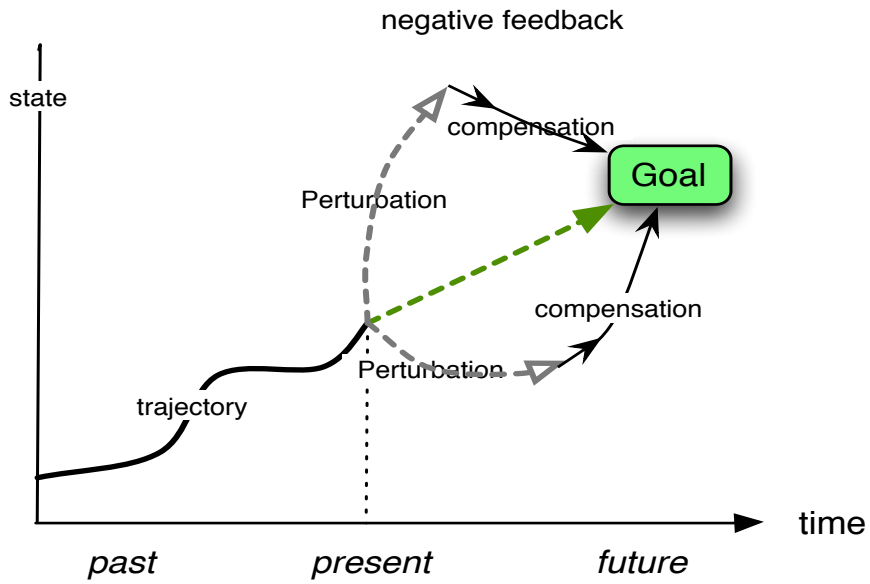

Figure 2: goal-directedness achieved through negative feedback: deviations from the goal-directed trajectory (green broken arrow) caused by perturbations (grey broken arrows) are compensated by counteractions.

The overall result of this in-built dynamics is that the actual temperature T will remain very close to the goal T_0. That is because the output actions, by opposing the effects of perturbing inputs, will suppress any deviations from the goal. The result is equifinality or persistence: the goal is consistently achieved, whatever the size or direction of the disturbances that drive it away from this state. A more sophisticated example of such a feedback-controlled system is a heat-seeking missile, which continuously monitors the difference between its position and the one of the heat source it is targeting. By compensating for sensed deviations (e.g. steering more to the east if the target deviates eastward from the present trajectory), it adjusts its course so as to reduce that difference and thus eventually hit the target. This is essentially similar to the cheetah adjusting its course so as to compensate for the sideway jumps of the gazelle.

Early cybernetics developed a mathematical theory inspired by such control technologies, which included thermostats, servomechanisms and anti-aircraft artillery. However, these systems were artificial, meaning that their goals were programmed into the system by their human designers. That may remind you of the "anthropomorphic" interpretation of goal-directedness, according to which goals need to be explicitly represented or consciously chosen for a system to know which state it should target. Nevertheless, negative feedback relationships, which are common in nature, will simply counteract deviations from *whichever state their dynamics does not oppose*. They can thus be conceived as directed towards that final state—without therefore needing any explicit representation of that state.

For example, a simple kind of thermostat used in fish tanks consists of two parallel metallic strips, one of which is asymmetric in such a way that it bends when the metal contracts because of cooling. As the temperature goes down, it bends further and further towards the other strip, until it touches that second strip. This contact closes an

electric circuit that switches on a heating element in the tank. As the temperature now goes up, the metal expands and its bending diminishes, eventually breaking the contact and thus interrupting the heating again. The precise temperature at which the two strips come into contact acts here as the goal state to which the system will persistently return. Yet, nowhere in the system is there an explicit representation of that "desired" temperature. That temperature depends only on the bending constant and the initial distance between the strips. It can be easily adjusted by increasing or decreasing that distance. And of course, there is no process in the system that could be construed as a conscious reflection about what action to take to achieve that goal.

Cybernetics further extended its reach to describe naturally goal-directed systems, such as organisms, brains and social systems (Ashby, 1960; Luhmann, 1986). Here the goal is not determined from the outside, by a designer, but intrinsic to the system. These systems are truly autonomous or self-steering (Heylighen *et al*., 1990), and thus can be said to exhibit *agency*. For systems that are the product of evolution, the implicit goal is survival and reproduction. The reason is that systems that did not effectively target and achieve this goal have been eliminated by natural selection. However, to reach this overall goal, evolved systems have to aim at several more concrete, subordinate goals, such as seeking food when hungry, warmth when cold, or safety when threatened. I will discuss in the last section how such a network of mutually dependent control actions and goals could have originated.

A DYNAMICAL SYSTEMS INTERPRETATION OF GOAL-DIRECTEDNESS

I will now synthesize and formalize the different conceptions of goal-directedness that I reviewed with the help of *dynamical systems theory* (R. D. Beer, 1995; Sternberg, 2010; Strogatz, 2000). This is a mathematical framework that generalizes and extends Newtonian mechanics, so as to be applicable to complex, non-linear systems with many components. Thus, it provides a general formalism for describing a causal evolution from initial states to final states. Yet, I am here not going to delve into the mathematics of dynamical equations, but merely review some of the associated qualitative notions that will help us to understand goal-directed behavior.

A dynamical system is described by a set of variables $\{s_1, s_2, s_3, \ldots\}$. Variables represent properties that can take on different values, such as positions, velocities, temperatures, or the concentrations of different types of molecules in a solution. These properties typically characterize aspects or components of some concrete system—such as a pendulum, a car, an organism, or an economy. Note that when modeling goal-directed systems, such as organisms, "the system" will here refer to the properties and behaviors of the organism or agent itself, not to properties of its environment (such as presence or absence of food). The dynamics of the system describes how these properties change over time. This is done by introducing the notion of the system state, $s(t)$, defined as the list of the values of all the variables that characterize the system at time t:

$$s(t) = (s_1(t), s_2(t), s_3(t), \ldots)$$

Thus, the state is supposed to provide a complete description of the system at a particular instant—at least insofar that its further evolution is concerned. For example, for a mixture of chemicals in a reaction vessel, the state would include the concentrations of all the different molecules present. Each conceivable combination

of values of the variables defines a potential state for the system. The set of all potential states defines the *state space* (sometimes also called phase space) of the system.

The evolution is then defined as the path or trajectory that the system describes through its state space, visiting different states at subsequent points in time: $s(t)$, $s(t + 1)$, $s(t + 2)$, $s(t + 3)$, … This evolution is assumed to be determined by an equation that expresses the causal relation between the state at the initial time t (cause or input) and the subsequent time $t+1$ (effect or output):

$$f(s(t)) = s(t+1)$$

For simplicity, I here use a "difference" equation, which describes states undergoing discontinuous transitions to subsequent states in discrete time. In the more complex case of continuous evolution, the dynamics is more commonly expressed through a differential equation. The state transition function f maps any state in the state space to the subsequent state to which it would normally evolve given the dynamics of the system. The dynamics represents the whole of "forces" acting within the system so as to make it change.

Note that this mapping function on the state space implicitly implements re-entry: the output of the first transition, $s(t + 1)$, becomes the input for the next transition:

$$s(t + 2) = f(s(t + 1)) = f(f(s(t))$$

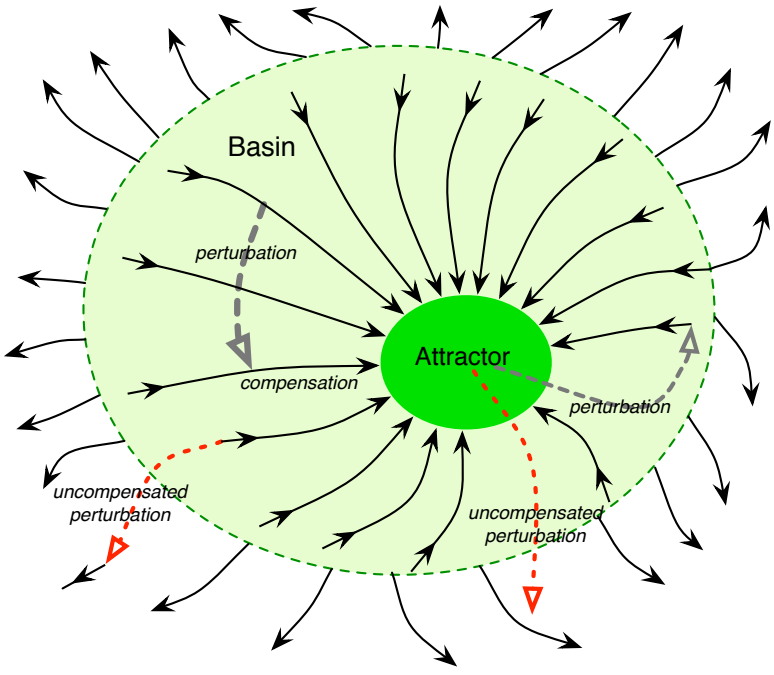

Figure 3: a phase portrait of a dynamical system with an attractor (dark green) surrounded by its basin of attraction (light green). All trajectories (solid arrows) originating in the basin converge to the attractor, unlike those originating outside the basin. Perturbations (broken arrows) are compensated by a new attractor-directed trajectory if their outcome remains within the basin, otherwise they are not.

This recurrent application of *f* to its own results is also known as an "iterated map". It can be used to generate all the subsequent states on the trajectory starting from the initial state $s(t)$. By doing this for initial states throughout the state space, we can create an elegant visual representation—called a "phase portrait"—of the possible evolutions of the system. Here, the state space is represented as a two-dimensional plane. From each point (state) in that space an arrow is drawn that points to the subsequent state that the system would encounter on its trajectory through that initial state (see Fig. 3). These arrows graphically represent the inherent *directionality* of the dynamics, i.e. the direction in which the system will move next, starting from any initial state. The chains of subsequent arrows represent possible trajectories through the state space.

Now we can distinguish two basic modes of operation for such a system. In the first mode, which characterizes reversible, Newtonian systems, the trajectories remain parallel: they do not join, converge or intersect. Here, distinct initial states always lead to distinct final states: no equifinality. In the second mode, which typifies irreversible, non-linear systems, certain trajectories do converge. The end to which they converge is called an *attractor*. In the theory of dynamic systems, an attractor is defined as a region in state space that the system can enter, but cannot leave—and that contains no smaller such region (Milnor, 2006; Sternberg, 2010). An attractor is surrounded by a *basin of attraction*. This consists of all the states whose trajectories converge to the attractor. That means that the dynamics of the system is such that if you calculate the trajectory starting from any initial state in the basin, then that trajectory will eventually reach the attractor and stay there. The attractor functions as the final destination of all trajectories that start in its basin (Fig. 3). Under conditions that I will further specify, it can therefore be understood as the "goal" or "end" of the system.

The concept of attractor demonstrates mathematically how the phenomenon of equifinality is perfectly compatible with standard causal dynamics, where the trajectory of a system depends only on its present state, not on some putative future end state. So, one might wonder why we need to consider that final state at all: why not calculate each trajectory individually just from its initial state? The reason is that the trajectories can be determined only in the ideal case where the system follows its own deterministic dynamical rules, without undergoing any unknown interferences.

In practice, however, a real-life system, such as an organism, an ecosystem or an economy, is subjected to various unpredictable perturbations that make it deviate from its initial trajectory. These perturbations result from events that the system does not control, i.e. that are not part of its dynamics as defined by the function *f*. They typically originate in the system's environment. Examples are encounters with other organisms, accidents or changes in the weather.

Because of such external disturbances, the precise trajectory of a complex system, such as an organism, is in practice impossible to predict. Nevertheless, if we know the attractors and their basins, then the final destination becomes predictable again: as long as the perturbations do not push the system outside of its present basin, its dynamics will sooner or later lead it back to that basin's attractor. This is illustrated in Fig. 3, which depicts two types of perturbations. The ones that stay within the basin are automatically compensated by the dynamics, while those that move out of the basin are not. Therefore, for complex, open systems an attractor-based model is more useful, flexible, and realistic than a fully deterministic dynamical model.

The term “attractor” may be misleading, though. It is not the attractor that exerts some pulling force, attracting the system towards itself. It is the system, because of its own, intrinsic dynamics, that is consistently moving towards the attractor, and that will persist doing so even if pushed away from its initial trajectory by a perturbation (that does not leave the basin). It must be emphasized that the phase portrait with its attractors and basins describes the system’s autonomous way of functioning, i.e. its in-built procedures or condition-action rules. It does not represent an external field such as a food gradient that would guide the system towards its goal, like in the model of McShea (2012). As I will explain further with the example of chemotaxis, if the system ends up at a food source, it is because its condition-action rules only allow it to stop moving when it senses the presence of food, not because of some force emanating from the food that pulls on it. This is similar to the thermostat interrupting the heating only when its sensor registers the set goal temperature. Clearly, that temperature does not attract the heating system towards it...

If we make the additional assumption that the attractor is not an equilibrium state of minimal potential energy, like the ball at the bottom of the bowl, then this attractor-directed movement requires the on-going mobilization of energy-providing resources. This is necessary to climb up any potential energy gradients and to compensate for energy lost through dissipation. Therefore, the system’s moves towards the attractor and against deviations can be interpreted as goal-directed actions. Such a system exhibits a basic form of *agency*: it actively resists forces pulling and pushing it, rather than passively being subjected to them, like the ball being pulled down by the force of gravity.

Its actions have the effect of compensating, neutralizing or suppressing perturbations. But because the system’s dynamics is causal, such a counteraction can only be started after the state of the system has deviated in some way from the original, “intended” one. Its new initial state now determines a new, adjusted trajectory, albeit still directed at the same attractor (Fig. 3). Thus, any change, whether the effect (“output”) of some move by the system, some outside perturbation, or some combination of the two, feeds back into the present state or “cause”, acting as a new input to the process. The feedback is negative if it opposes deviations away from the attractor—thus making the system consistently return to the attractor.

CONCERTED ACTION

The dynamical systems model can further help us to understand the aspect that I called *concerted action*. The state of a complex system includes multiple variables along which the components of the system can vary. For example, the movement of a simple system like a ball can be described by just the coordinates of the ball’s center of gravity. But to describe the more complex movement of a human body, you will need to include at least the degrees of freedom that characterize all the different limbs and joints of the body. For example, the angle separating upper and lower bones of an elbow defines a single, one-dimensional variable, while the orientation of an arm relative to the shoulder requires two variables, and the rotation of the head around the vertical axis requires yet another variable. The state space of a dynamical system representing human movement is therefore many-dimensional.

A specific movement, such as climbing a staircase, can be described as a particular trajectory through that state space, leading towards a certain end state, which here is the position standing at the top of the staircase. To describe that trajectory, the

different body parts must move in a coordinated manner. For example, while one leg stretches to push the body up towards the next stair, the other leg must bend to position itself on that stair. That coordination is expressed in the dynamical system representation by the fact that the different variables that make up the state of the system change in a concerted manner. For example, while the angle of the knee in the stretching leg increases, the angle of the knee in the other leg decreases. Perturbations that interfere with this coordination must be counteracted to keep the system on track. For example, if the bending leg bends too much, its foot will dangle above the stair without support, until a compensating stretch restores the balance.

This shows that the dynamical systems representation implicitly encompasses the notion that goal-directed behavior requires the coordinated activity of a broad repertoire of actions, aspects and components of the system. This coordination can be quite complex, including different types of dependencies between the actions performed by different components (e.g. body parts) and the sensed variables that determine the state of the system (e.g. positions of body parts). Such dependencies can be classified in the following categories (Heylighen, 2013; Trestman, 2012):

- *sequential* (an action must be followed by a specific, subsequent action),
- *parallel* (two or more actions must be performed simultaneously),
- *recurrent* or *circular* (the result of an action is fed back to become the condition that triggers the same kind of action),
- *hierarchical* (a complex of actions functions together as a single higher order action), or more generally
- *networked* (several actions depend on each other via a network of connections)

While such complex, goal-directed action can in principle be described as the activity of a dynamical system, this description raises a fundamental question: how could such a coordinated system have arisen without some form of intelligent design of this complex network of dependencies? The theory of evolution by natural selection explains how an accumulation of small variations can make an organism increasingly complex and adaptive in its organization. For example, variation through multiplication and mutation of DNA strings followed by selection produces genomic networks that regulate the different metabolic processes in the cells, and thus specify the living system's dynamics. Epigenetic processes on these networks further produce the different functional organs of multicellular organisms during ontogenetic development.

The functional elements of this regulatory (epi)genetic network can be conceived as *condition-action rules* (also known as production rules: Anderson, 2014; Heylighen, 2016). These specify which action should be taken under which condition. For example, a rule may state that in the presence of a particular type of molecule a particular enzyme should be produced (by reading and expressing the DNA sequence that codes for this enzyme). That enzyme will then react with that molecule to transform it into a different molecule (e.g. to neutralize a toxin or digest a food molecule). Thus, *a condition-action rule expresses an actual causal process*, executed by the organism's metabolism, that transforms one physical condition (e.g. presence of molecule) into another one (e.g. neutralization of that molecule by an enzyme). The whole of condition-action rules, which act on the different relevant variables of the system (such as the concentrations of different types of molecules in the cell), defines the dynamic system that governs the changes in the state.

We can assume that these condition-action rules have been selected for their ability to lead the organism towards an attractor (goal) that ensures its fitness. Organisms

whose attractor states are unfit simply do not survive. For example, a bacterium's condition-action rules will be such that it moves towards substances it considers as food and away from substances it considers as toxins (chemotaxis, see Sourjik & Wingreen, 2012; Wadhams & Armitage, 2004). This is implemented by the iterated application of the following simple rule:

if the concentration of food molecules decreases,
then change direction of movement

The result is that the bacterium will be continually moving towards (or within) the zone where the food concentration is highest. Remaining in this zone then plays the role of the attractor of this very simple dynamics. However, note that the bacterium does not have any conception of this zone as something existing external to itself: for the bacterial dynamical system the attractor is merely that part of its (internal) state space where the sensed concentration of food does not decrease. The bacterium's behavior is neither determined by its internal program, like in Mayr's (1974) conception of teleonomy, nor by the external field of food concentrations, like in McShea's (2012) conception, but by the bacterium's (partly programmed) reactions to (unprogrammed) external perturbations (in this example, decreasing food concentrations). This is similar to the behavior of the thermostat, whose (internal) rules tell it to switch on the heating each time the temperature dips below the goal temperature (external perturbations). Neither thermostat nor bacterium need to know (i.e. have an internal representation of) the external situation that corresponds to their internal goal state: they only need to know when to act (to compensate a perturbation), or stop acting (when having reached the goal).

"Food" here is defined as something the bacterium can digest to extract the resources that are necessary for its survival, while "toxin" is defined as something that interferes with its metabolism, and therefore endangers survival. Bacteria whose internal dynamics were directed at different attractors (e.g. moving towards toxins, or ignoring food) have been eliminated by natural selection. The same applies to bacteria whose actions were insufficiently energetic or too poorly coordinated to effectively achieve the right goals in the face of common perturbations. Thus, variation and selection adequately explain the development and fine-tuning of a concerted action dynamics directed at fitness-supporting goal states.

Still, this scenario assumes that there already is an organism whose overall, implicit goal is survival, i.e. maintaining its living organization (Mossio & Bich, 2017). More concrete, subsidiary goals, such as evading toxins or catching a gazelle, can then be easily explained as means that have evolved in the service of this overall end. But where did the primary goal of survival come from? This is another way of formulating the *origin of life* problem: how did an assemblage of physico-chemical processes that were not goal-directed give rise to an organization that was efficiently directed at the goal of self-maintenance (Deacon & Sherman, 2007)? And how did such an organization develop the necessary plasticity and persistence to dependably achieve that goal in a highly variable environment bombarding it with a wide variety of perturbations?

As a first step towards clarifying these issues, I will review some general properties of attractors and basins. In the next section, I will then introduce the formalism of reaction networks to demonstrate how the right kind of dynamics can self-organize from the equivalent of chemical reactions.

ATTRACTORS AND BASINS

An attractor as a cohesive region in a multidimensional state space can have a wide variety of sizes, shapes, and dimensions, including the chaotic or fractal shapes known as "strange attractors". The simplest one is a single state s_0, which is known as a *fixpoint* of the transition function f: any further transition merely returns to the same point.

$f(s_0) = s_0$.

A point-like attractor can be seen as a model of *homeostasis*: the fixpoint represents the "ideal" state for the system from which it will not move (homeostasis = staying the same). In practice, for organisms, certain variables are allowed to change within a limited range around their ideal values, for as long as these variations do not endanger the essential organization. For example, the temperature of the human body can vary within a small range around 37° C. That means that the attractor covers a small, continuous region rather than a point in the state space.

Another common type of attractor is a *limit cycle*. Here the endpoint of the process is not a single state but a continuously repeating sequence of states (see Fig. 4). For example, here is a sequence cycling through the three states s_1, s_2, s_3:

$f(s_1) = s_2$,
$f(s_2) = s_3$,
$f(s_3) = s_1$

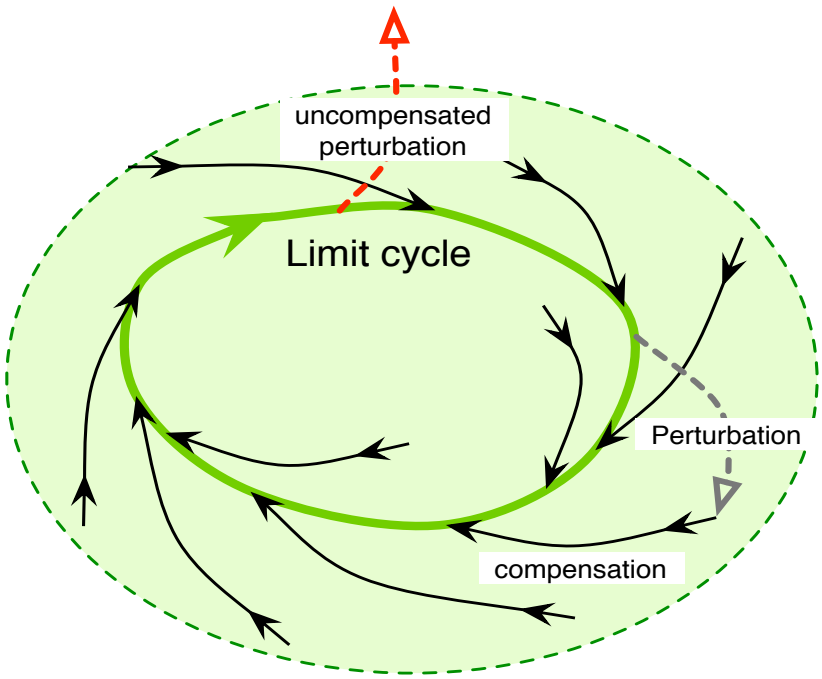

Fig. 4: a limit cycle attractor (green ellipse) surrounded by its basin (light green). Solid arrows represent trajectories converging to the limit cycle, broken arrows represent perturbations

A classic example of this type of attractor can be found in the cyclical increases and decreases in the populations of predators and prey, as modeled by the Lotka-Volterra equation. But a limit cycle could also describe various metabolic cycles inside the organism in which certain molecules or structures are periodically broken down and then rebuilt. Such cycles must be kept going, and any deviation from them must be opposed so that the system converges back to its attractor. The goal here is not a state, but a periodic process. For example, in the case of the cheetah, the goal of killing a gazelle is in fact only a stage in a cycle that could be described more precisely as:

hunting → killing → feeding → resting and digesting → hunting → …

A perhaps unconventional attractor regime can be represented by an infinite line (Fig. 5). (Note that the line itself is not strictly an attractor, but a component of an attractor basin that converges toward an attractor lying at infinity.) This can describe processes of on-going growth, where certain variables, such as population, size or experience increase without a priori limit. For example, the development of trees is such that they only increase in diameter as additional rings of wood are added year by year. This growth may be perturbed, e.g. by drought reducing the availability of water necessary for growth so that less wood is added in a difficult year. But the tree as a goal-directed, dynamical system will try to compensate for such deviation from its ideal growth process, e.g. by producing longer roots that reach more deeply for water.

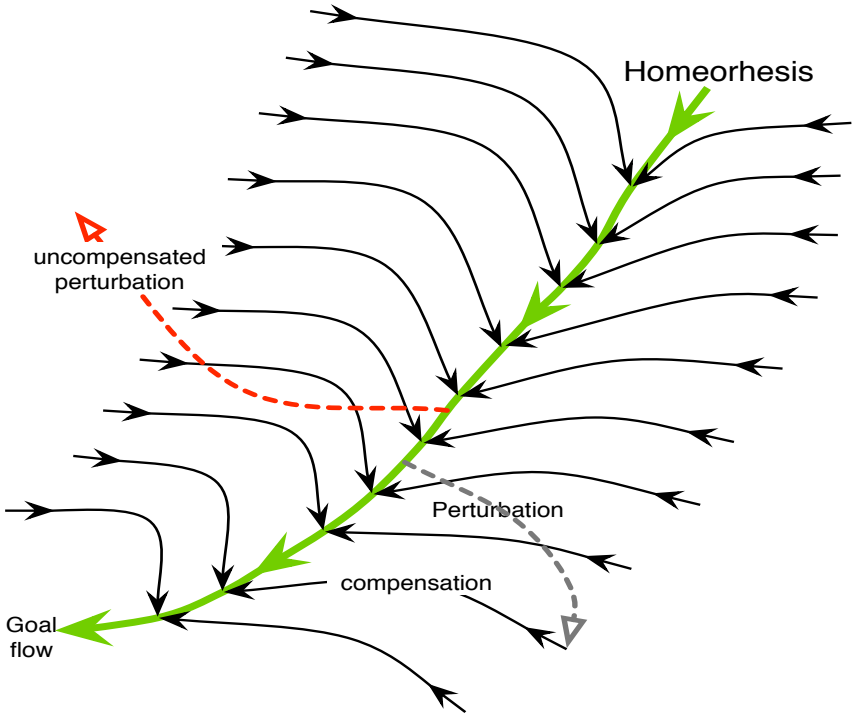

Fig. 5: an attractor regime (green arrow) in the form of a continuing line, illustrating the goal-directed dynamics of homeorhesis.

Such return to an ideal process rather than an ideal state may be called *homeorhesis* ("flowing the same") instead of *homeostasis*. The term homeorhesis (Gare, 2017) seems to better express the notion that living systems are directed towards a trajectory of unlimited development than the more ambiguous concept of allostasis ("staying different") (Day, 2005).

An important issue is that goal-directedness is characterized by sufficient plasticity, persistence and concerted action. All of these imply a *large* basin of attraction surrounding the attractor/goal. The larger the basin, the larger the number of initial states ending up in the attractor (equifinality, plasticity), the larger the deviations from its initial trajectory the system will be able to compensate (persistence, negative feedback), and the larger the number of variables that it will be able to control in a coordinated, goal-directed manner (dynamical coherence, concerted action).

Now, imagine an attractor with a small surrounding basin that extends just a little outside the attractor. The smallest perturbation will be sufficient to push the system across the border of that basin, into the basin of a different attractor. That means that it cannot on its own return to its goal. Goal-directedness here is very weak, because the system is hardly able to resist any perturbation from its preferred configuration. Such a goal configuration is fragile or brittle.

The opposite of brittleness is robustness, or more precisely *resilience*, i.e. the ability to bounce back to the goal configuration after undergoing a more or less intense perturbation. Thus, from the dynamical systems perspective, effective goal-directedness appears equivalent to the resilience of the goal configuration. The literature on resilience in social and ecological systems (Beigi, 2019; Holling, 1973; Walker *et al*., 2004) further clarifies how the properties of the basin affect the persistence of the attractor.

Another question worth raising is how a dynamical system can model subordinate goals (e.g. feeding), which are in reality merely way-stations, stepping stones or means that lead to the overall goal or end (e.g. surviving). An attractor by definition is the ultimate outcome of a process. Any intermediate steps towards this end must therefore be situated in the basin, not in the attractor. Yet, they may still be attractors for a subsystem of the overall dynamic system, i.e. a partial system that lacks certain components or variables.

For example, the goal of feeding can be modelled as an attractor if we consider "hunger" not as a variable, but as a fixed property with the value "high". For hungry cheetahs, feeding is the dominant goal. However, if we turn "hunger" into a variable, the model would need to incorporate an additional dynamics which states that hunger decreases sharply after feeding, that the system in a state of low hunger is directed at "resting and digesting", and that the latter activity slowly increases the value of the variable "hunger", until we get back to a state where "feeding" is the dominant goal. Such an additional variable plays the role of a *control parameter*, which changes the dynamics in such a way that stable solutions (attractors) may shift, bifurcate, or disappear while the parameter value changes.

An example of such an approach can be found in *Perceptual Control Theory* (PCT), a cybernetics-inspired dynamical model of goal-directed behavior in organisms and people (Bourbon, 1995; Powers, 1973, 1995). Here, a goal (reference value to which a negative feedback control loop converges) at a subordinate level is set by the action (output) of a control loop directed at a goal of a higher level. The higher-level control loop takes into account additional perceived variables (such as hunger) to specify which goal (e.g. feeding or resting) the subordinate loop should

target. This method of turning a goal state that is fixed at one level into a variable that is controlled at a higher level allows PCT to develop a complex, multilevel hierarchy of goals for describing behavior on both short-term and long-term time scales.

THE ORIGIN OF GOAL-DIRECTED SYSTEMS

The last question I wish to address in this paper is how such complex dynamic systems with a far-from-equilibrium attractor surrounded by a large basin could have developed. As we saw, biological evolution through variation and selection can explain the continuing fine-tuning of such systems, but not the origin of goal-directedness itself. Implicitly, though, evolutionary theory specifies the most fundamental goal of all: survival. For far-from-equilibrium systems, survival means first of all *active self-maintenance*: the continuous repair and reconstitution of components and structures that would otherwise be consumed by the inevitable production of entropy. This concerns in particular the dynamical structure that distinguishes the "self", i.e. the system to be maintained, from its environment, i.e. the perturbations that affect the system's functioning, but which the system does not control.

It is this perspective that led Maturana and Varela to define living systems as *autopoietic*, i.e. self-producing (Maturana & Varela, 1980; Mingers, 1994; Razeto-Barry, 2012). Autopoiesis is achieved through *organizational closure*: the processes inside the system are organized in a circular manner, so that their output becomes part of their input and dynamics, and the different components and processes reconstitute each other. Self-maintenance can be seen as the ultimate goal, value, or end of a living system (Mossio & Bich, 2017). For such natural systems, "the purpose of the system is what it does", to use a famous expression coined by the cybernetician Stafford Beer (Beer, 2002; Lockton, 2012). The system's goal is not imposed by some intelligent designer, like in artificial or "allopoietic" systems. Its goal is intrinsic or implicit to the way the system functions: to continue functioning in the same way. Thus, a living system exemplifies a structure with *emergent purpose*: its components and processes are merely simple causal mechanisms; yet together they form an autonomous "agent" or "self", i.e. an organizationally (but not thermodynamically) closed whole that acts so as to ensure its continued functioning.

While autopoietic systems are goal-directed in this sense, the theory of autopoiesis does not explain how they could have evolved out of earlier abiotic systems that lacked this feature. Independently of autopoietic theory, many scenarios have been proposed for the origin of life. While much progress has been made in understanding specific physical and chemical conditions that may give rise to certain of the concrete components, such as amino acids, cell membranes or RNA, the overall process remains obscure. The problem is that none of these scenarios explains how such abiotic components could have developed purposeful action (Deacon & Sherman, 2007). That seems intrinsically difficult when the starting elements are linear causal mechanisms and passive physical objects, such as molecules or membranes, rather than actively self-maintaining organizations.

From the proposed scenarios, the one that comes closest to the circular causality demanded by cybernetics and autopoiesis is the self-organization of an *autocatalytic cycle* (Hordijk *et al*., 2010; Kauffman, 1986). However, this scenario has a fundamental shortcoming. Autocatalysis is by definition a form of *positive feedback*: the more there is of a catalyst, the more will be produced. The resulting runaway

growth makes the cycle intrinsically unstable: instead of coming to some "desired" concentration (goal), its reserve of catalysts grows exponentially, until the "food" molecules it needs for further production are exhausted. Unless there is a dependable outside source of food, this entails the interruption of the cycle and therefore the death of the incipient "metabolism" (Deacon & Sherman, 2007).

Inspired by the theory of autopoiesis, Peter Dittrich has proposed a generalization of the notion of autocatalytic cycle, which he called a *chemical organization* (Benkö *et al.*, 2009; Dittrich & Fenizio, 2007; Veloz *et al.*, 2022). Such a generalized cycle can include both positive and negative feedbacks, but is generally stable. Like an autopoietic system, it produces its own components, thus constituting itself as an autonomous system. But unlike the rather obscurely formulated notion of autopoiesis, a chemical organization has a precise mathematical structure that can be investigated analytically and computationally using the formalism of reaction networks.

The elements of this formalism are *molecules* (also termed "(molecular) species") $\{a, b, c, \ldots\}$ and *reactions* $\{r_1, r_2, \ldots\}$. Reactions map combinations of molecules onto different combinations. They have the form:

$$r: a + b + \ldots \rightarrow e + f + \ldots$$

Reactions represent processes that consume certain molecules (the "reactants" or inputs of the reaction: *a, b*, …), while producing other molecules (the "products" or outputs of the reaction: *e, f*, …). They can be catalyzed by specific molecules, but need not be. Reactions can be interpreted as condition-action rules, with the input side representing the necessary conditions (joined presence of the reactant molecules) for the reaction to take place and the output side representing the resulting action (generation of products and removal of reactants). Reactions thus represent the physical forces or interactions, such as chemical reactions, that drive or *cause* the dynamical behavior of the system.

"Molecules" can be, but do not have to be, actual chemical reactants. They can represent any resources that react with other resources to produce further resources. The reaction system formalism of Chemical Organization Theory (COT) makes abstraction of the physical or biochemical nature of the resources so as to better understand the functional relationships between reactions (Veloz & Razeto-Barry, 2017a, 2017b). That makes COT perfectly general, and applicable to a wide range of domains, including physical, biological (Kreyssig *et al*., 2012), ecological (Veloz, 2019), social (Dittrich & Winter, 2008) and computational systems (Matsumaru *et al.*, 2005)

A subset of the molecules and reactions in a reaction network is called an *organization* when it is closed and self-maintaining. *Closed* means that no molecules are produced by the reactions that were not in the initial subset. *Self-maintaining* means that all molecules in the subset that are consumed by reactions are produced again by other reactions in an amount sufficient so that their total concentration does not diminish. Thus, the components that make up the organization are perpetually reconstituted: nothing gets lost, nothing new is created. Note that the formalism allows the input of "food" molecules and evacuation of "waste" so as not to contravene thermodynamic constraints. That means that organizations are typically far-from-equilibrium configurations.

As an illustration, here is a very simple organization representing the Earth's ecosystem at a high level of abstraction:

$\rightarrow$ *sunlight*
plants + *sunlight* + CO_2 + *minerals* $\rightarrow$ *plants* + O_2 + *heat*
plants $\rightarrow$ *waste* + *heat*
plants + *animals* + $O_2 \rightarrow$ *animals* + CO_2 + *waste* + *heat*
waste + *decomposers* + $O_2 \rightarrow$ *decomposers* + CO_2 + *minerals* + *heat*
heat $\rightarrow$

This describes the recycling of oxygen, carbon dioxide and minerals by plants, animals and decomposers. This cycle is fueled by the energy of the sun (which enters the system from the outside, which is why the reaction producing it has no input within the system), while dissipating heat into space. The resulting self-maintaining network can be conceptualized as a planetary-scale autopoietic system, which has been called Gaia (Rubin *et al*., 2021).

Note that an organization directly implements the feature that I called concerted action. The different reactions in the corresponding network are coupled in sequence, in parallel, and in cycles, thus working together to support the maintenance of the whole. Each molecule or reaction plays a particular role or function in the overall autopoietic system (Nunes-Neto *et al*., 2014). In the above illustration, e.g., the function of plants is to produce the oxygen and food needed by the animals and to consume carbon dioxide, while the function of the decomposers is to release the minerals needed by the plants. These functions are essential for the system to persist. Other functions contribute to self-maintenance but are not essential. For example, while animals contribute to the production of the carbon dioxide needed by the plants, a similar function is performed by the decomposers. Therefore, the Gaian organism could in principle survive without the animals. This is of course a highly simplified example, and more typical organizations will contain hundreds of molecules and cyclically coupled reactions. Further research will be needed to disentangle the different functionalities, hierarchies and forms of coordination used in different types of chemical organizations.

Reaction networks can be modelled as dynamical systems. The concentrations of the different molecules play the role of the variables that define the state of the system. The rate at which reactions take place defines the transition function that determines how the state changes over time. It can be shown that the *attractors* of such a dynamical system are all organizations (Dittrich & Fenizio, 2007; Peter & Dittrich, 2011). That means that sufficiently complex networks tend to spontaneously evolve towards such a self-maintaining configuration. In other words, the configuration will *self-organize* (Heylighen *et al*., 2015).

Thus, according to our dynamical system interpretation of goals, a chemical organization behaves like a goal-directed system that will protect its autopoiesis by compensating for perturbations that may make its trajectory deviate from the goal (Busseniers *et al*., 2021). Still, there are an infinite number of conceivable organizations, from trivially simple to extremely complex. The mathematical formalism of Chemical Organization Theory (COT) provides a simple, effective and computationally tractable tool for identifying and analyzing the different types of potential organizations that exist within a given network of reactions (Centler *et al.*, 2008).

My collaborators have recently developed a software application (Veloz *et al*., 2022) that efficiently computes all the organizations in a given reaction network. Such a network may be derived from empirical data (e.g. about metabolic or ecological networks), created as a toy model for the purpose of demonstration, or randomly

generated. The software can also simulate the dynamics that allows the network to settle into one of these organizations, either starting from a given initial state or after a perturbation. Thus, we have a tool that allows us to explore the self-organization and evolution of self-maintaining systems under a wide range of different conditions.

It is clear why the attractors of such a dynamical system must be (chemical) organizations: a configuration that is not closed and self-maintaining will by definition change into a different configuration—by acquiring new molecules (non-closure) and/or losing existing molecules (non-self-maintenance). Therefore, it cannot be the final stage of the dynamical evolution of a reaction network.

However, it is not evident whether the organizations that *are* final stages will have all the features we associated with goal-directedness, i.e. sufficiently great plasticity, persistence and concerted action. Organizations may have a very small (or even empty) basin, meaning that nearly any perturbation will push the system into a different basin, thus making it end up in a different organization, where in turn it may be perturbed and pushed to yet another organization, and so on. Such a haphazard jumping from attractor to attractor cannot really be interpreted as the controlled self-maintenance that we associate with goal-directed living systems.

Still, it seems plausible to assume that a sufficiently complex reaction network would exhibit a wide variety of attractors, some of which have larger basins than others. An on-going series of random perturbations applied to this system is likely to make it jump out of the fragile, small-basin attractors while allowing it to settle into the most resilient, large-basin attractors, which are resistant to most further perturbations (Heylighen *et al*., 2015). Our preliminary simulation results with randomly generated reaction networks and perturbations seem to confirm this hypothesis. If we could generalize these simulation results to actual networks of chemical reactions, then we have a plausible scenario for the origin of robustly self-maintaining systems that seem to exhibit the fundamental features of goal-directed activity, and thus life, i.e. great plasticity, persistence and concerted action.

While further research is needed to elaborate this scenario in the necessary level of detail, I hope to have shown that the emergence of goal-directed organizations from simple causal processes is neither mysterious nor improbable, but rather the natural outcome of an evolution that homes in on self-maintaining configurations, while selecting the ones with the greatest ability to survive a wide range of perturbations.

CONCLUSION

Within the sciences of life, goal-directedness (or teleonomy) remains a controversial concept that is frequently misunderstood. This paper has tried to clarify what goal-directedness is, how it can be operationalized and modelled, and how it may have originated from causal processes that are not goal-directed. First, I have tried to exorcise the roots of the controversy by pointing out that goal-directed behavior does not presuppose any obscure mechanisms—such as supernatural guidance, intelligent design, vital force, conscious intention, or backward causality—that fall outside our present understanding of the causal mechanisms that govern physical systems.

I then showed that goal-directedness is not only consistent with standard, "forward" causality, but that it complements such causal explanations by allowing us to predict the outcome of processes where complexity, stochasticity or external perturbations make step-by-step causal prediction impossible. For example, the assumption that a cheetah's behavior is directed at the goal of feeding allows you to

predict what will happen when the cheetah catches the gazelle it is chasing—no matter how complex and unpredictable the trajectory described by the running cheetah and gazelle.

This feature of diverse and irregular trajectories predictably ending up in the same final state is known as *equifinality*. Related characteristics of goal-directedness have been formulated as *plasticity* (being able to reach the same goal under a variety of different circumstances), *persistence* (continuing to move towards the same goal in spite of perturbations) and *negative feedback* (neutralizing deviations from the goal by a compensating counteraction). A more complex characteristic is what I have called *concerted action*: the coordination of the activities of the different components of the system towards the goal.

All of these features can be elegantly explained by means of a dynamical systems model of the goal-directed process. The dynamics of such a system commonly exhibits attractors: regions in the system's state space to which all trajectories in the attractors' basin converge, so that the attractor functions as the true "end" of the process. Perturbations are in this case automatically compensated by the dynamics, at least as long as they do not push the system outside of the basin. The attractor/goal does not need to be a fixed state (homeostasis), but can be a complex process that is circular (limit cycle) or ever advancing (homeorhesis).

In order to differentiate a goal-directed system, such as a living organism, from a system that merely settles in a stable equilibrium, such as a ball ending up at the bottom of a bowl, I added the requirement that the attractor should correspond to a far-from-equilibrium configuration, i.e. one that requires input of energy in order to reach and maintain. That fits in with our general intuition of goal-directed agency as actively intervening in the world by mobilizing resources—rather than passively submitting to external forces.

While such dynamical systems may provide a causal model of goal-directed behavior, this invites the question how systems could have originated that have the right configuration of causal processes to produce this kind of dynamics. To elucidate this, I briefly sketched the formal framework of Chemical Organization Theory, which starts from reactions (chemical and other) between resources as elementary causal processes. This framework shows how networks of such reactions tend to self-organize into attractors known as "organizations", which are characterized by self-maintenance and closure. Thus, organizations provide a simple and elegant model of the circular, self-producing dynamics that have been postulated in the theories of cybernetics, autopoiesis and autocatalytic cycles as the essence of autonomous agency, and thus of life. While this model requires a bit of abstract mathematics in order to demonstrate the otherwise counterintuitive notion that self-maintenance can emerge spontaneously, its elements, the reactions, represent concrete physical, chemical or biological processes.

Yet, goal-directedness as I defined it, and life as we know it, require more than a self-maintaining configuration. That configuration should moreover be highly resilient, i.e. it should be able to recover from the wide variety of perturbations it is likely to encounter in a natural environment. That in turn requires that the corresponding attractor should be surrounded by a very extensive basin—a requirement equivalent to wide-ranging plasticity, persistence and concerted action. In further research using software simulations, my collaborators and I hope to elucidate the precise conditions under which such extended-basin organizations may emerge. This would provide us with a formal model of the origin of both life and goal-

directedness (Heylighen *et al.*, 2022),, and thus with the concepts and methods needed to develop a true science of naturally goal-directed, purposive or teleonomic systems.

ACKNOWLEDGMENTS

This research was funded by the John Templeton Foundation as part of the project "The Origins of Goal-Directedness" (grant ID61733), under its research program on "The Science of Purpose". I thank my VUB colleagues collaborating on this project (Veloz *et al*., 2022), and in particular Evo Busseniers, Shima Beigi and Tomas Veloz, for many inspiring discussions on the concepts presented here. I also thank Peter Corning and Richard Vane-Wright for inviting me to the Linnean Society meeting "Evolution 'On Purpose'", where this paper was first presented, and where I discovered a number of related approaches summarized here. My article is thus a contribution to the special issue on *Teleonomy in Living Systems*, guest edited by Vane-Wright and Corning, based on this meeting .

DATA AVAILABILITY

This theoretical article is based on information available in the literature.

REFERENCES

Anderson JR. 2014. *Rules of the mind* (2nd edn). Hove, UK: Psychology Press.

Ashby WR. 1960. *Design for a brain: the origin of adaptive behavior*. New York: Wiley. http://archive.org/details/designforbrain00ashb

Babcock G, McShea DW. 2021. An externalist teleology. *Synthese* **199**: 8755–8780. https://doi.org/10.1007/s11229-021-03181-w

Barrow JD. 1998. Impossibility: the limits of science and the science of limits. New York: Oxford University Press.

Beer RD 1995. A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* **72**(1/2): 173–215.

Beer S. 2002. What is cybernetics? *Kybernetes* **31**(2): 209–219. https://doi.org/10.1108/03684920210417283

Beigi S. 2019. A road map for cross operationalization of resilience. *In* Rattan SIS, Kyriazis M (eds.), *The science of hormesis in health and longevity*. Elsevier: Academic Press, pp. 235–242. https://doi.org/10.1016/B978-0-12-814253-0.00021-8

Benkö G, Centler F, Dittrich P, Flamm C, Stadler BMR, Stadler PF. 2009. A topological approach to chemical organizations. *Artificial Life* **15**(1): 71–88.

Bertalanffy LV. 1969. *General system theory: foundations, development, applications* (revised edn). New York: George Braziller.

Bourbon WT. 1995. Perceptual control theory. *In* Roitblat HL, Meyer JA (eds), *Comparative approaches to cognitive science.* Cambridge MA: MIT Press, 151–172.

Brooks RA. 1991. Intelligence without representation. *Artificial Intelligence* **47**(1): 139–159. https://doi.org/10.1016/0004-3702(91)90053-M

Busseniers E, Veloz T, Heylighen F. 2021. Goal directedness, chemical organizations, and cybernetic mechanisms. *Entropy* **23**(8): 1039. https://doi.org/10.3390/e23081039

Centler F, Kaleta C, di Fenizio PS, Dittrich P. 2008. Computing chemical organizations in biological networks. *Bioinformatics* **24**(14): 1611–1618.

Corning, P. A. 2022 (in press). Teleonomy in evolution: "The Ghost in the Machine." In P. A. Corning (Ed.), *Evolution on Purpose*. MIT Press.

Day TA. 2005. Defining stress as a prelude to mapping its neurocircuitry: no help from allostasis. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **29**(8): 1195–1200. https://doi.org/10.1016/j.pnpbp.2005.08.005

Deacon TW. 2011. *Incomplete nature: how mind emerged from matter*. New York: Norton.

Deacon T, Sherman J. 2007. The physical origins of purposive systems. *Embodiment in Cognition and Culture* **71**: 3.

Dittrich P, di Fenizio PS. 2007. Chemical organisation theory. *Bulletin of Mathematical Biology* **69**(4): 1199–1231. https://doi.org/10.1007/s11538-006-9130-8

Dittrich P, Winter L. 2008. Chemical organizations in a toy model of the political system. *Advances in Complex Systems* **11**(4): 609. https://doi.org/10.1142/S0219525908001878

Gare A. 2017. Chreods, homeorhesis and biofields: finding the right path for science through Daoism. *Progress in Biophysics and Molecular Biology* **131**: 61–91. https://doi.org/10.1016/j.pbiomolbio.2017.08.010

Heylighen, F. 1989. Causality as distinction conservation: a theory of predictability, reversibility and time order. *Cybernetics and Systems* **20**(5): 361–384. https://doi.org/10.1080/01969728908902213

Heylighen, F. 1990. *Representation and change: a metarepresentational framework for the foundations of physical and cognitive science*. Ghent: Communication & Cognition. First published 1987; web edition: http:// pcp. vub. ac. be/ books/ Rep&Change. pdf

Heylighen, F. 2009. Complexity and self-organization. *In*: Bates MJ, Maack MN (eds), *Encyclopedia of Library and Information Sciences* (3rd edn). Boca Raton: Taylor & Francis, pp. 1215–1224. http://www.tandfonline.com/doi/abs/10.1081/E-ELIS3-120043869

Heylighen, F. 2013. Self-organization in communicating groups: the emergence of coordination, shared references and collective intelligence. *In*: Massip-Bonet À, Bastardas-Boada A (eds), *Complexity perspectives on language, communication and society*. Berlin: Springer, pp. 117–149. http://pcp.vub.ac.be/Papers/Barcelona-LanguageSO.pdf

Heylighen, F. 2016. Stigmergy as a universal coordination mechanism I: definition and components. *Cognitive Systems Research* **38**: 4–13. https://doi.org/10.1016/j.cogsys.2015.12.002

Heylighen F, Joslyn C. 2003. Cybernetics and second-order cybernetics. *In*: Meyers RA (ed.), *Encyclopedia of physical science and technology* (3rd edn) **4**: 155–169. San Diego: Academic Press. http://pespmc1.vub.ac.be/Papers/Cybernetics-EPST.pdf

Heylighen F, Rosseel E, Demeyere F. 1990. *Self-steering and cognition in complex systems: toward a new cybernetics*. New York: Gordon and Breach Science Publishers.

Heylighen, F., Beigi, S., Busseniers, E., 2022. The Role of Self-Maintaining Resilient Reaction Networks in the Origin and Evolution of Life. *BioSystems* (in press).

Heylighen F, Beigi S, Veloz T. 2015. *Chemical Organization Theory as a modeling framework for self-organization, autopoiesis and resilience*. Belgium: ECCO Working Papers (2015–01): 1–30. http://pespmc1.vub.ac.be/Papers/COT-ApplicationSurvey-submit.pdf

Hilborn RC. 2004. Sea gulls, butterflies, and grasshoppers: a brief history of the butterfly effect in nonlinear dynamics. *American Journal of Physics* **72**(4): 425–427.

Holling CS. 1973. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics* **4**: 1–23.

Hordijk W, Hein J, Steel M. 2010. Autocatalytic sets and the origin of life. *Entropy* **12**(7): 1733–1742.

Kauffman SA. 1986. Autocatalytic sets of proteins. *Journal of Theoretical Biology* **119**(1): 1–24.

Kreyssig P, Escuela G, Reynaert B, Veloz T, Ibrahim B, Dittrich P. 2012. Cycles and the qualitative evolution of chemical systems. *PloS One* **7**(10): e45772.

Lee JG, McShea DW. 2020. Operationalizing goal directedness: an empirical route to advancing a philosophical discussion. *Philosophy, Theory, and Practice in Biology* **12**: Article 005. https://quod.lib.umich.edu/p/ptpbio/16039257.0012.005/--operationalizing-goal-directedness-an-empirical-route?c=ptb;c=ptpbio;g=ptpbiog;rgn=main;view=fulltext;xc=1;q1=Article

Lockton D. 2012. *POSIWID and determinism in design for behaviour change* (SSRN Scholarly Paper ID 2033231). Elsevier: Social Science Research Network. https://doi.org/10.2139/ssrn.2033231

Luhmann N. 1986. The autopoiesis of social systems. *In*: Geyer F, van der Zouwen J (eds), *Sociocybernetic paradoxes* **6**: 172–192. London: Sage. http://cepa.info/2717

Lyman RL. 2004. The concept of equifinality in taphonomy. *Journal of Taphonomy* **2**(1): 15–26.

Matsumaru N, Centler F, di Fenizio PS, Dittrich P. 2005. Chemical organization theory as a theoretical base for chemical computing. *In*: Teuscher C, Adamatzky A (eds), *Proceedings of the 2005 Workshop on Unconventional Computing: from cellular automata to wetware*. UK: Luniver Press, pp. 75–88.

Maturana HR, Varela FJ. 1980. *Autopoiesis and cognition: the realization of the living*. Dordrecht: D Reidel Publishing.

Mayr E. 1974. Teleological and teleonomic, a new analysis. *In*: Cohen RS, Wartofsky MW (eds), *Methodological and historical essays in the natural and social sciences*. The Netherlands: Springer, pp. 91–117. https://doi.org/10.1007/978-94-010-2128-9_6

McShea DW. 2012. Upper-directed systems: a new approach to teleology in biology. *Biology & Philosophy* **27**(5): 663–684.

Milnor JW. 2006. Attractor. *Scholarpedia* **1**(11): 1815. https://doi.org/10.4249/scholarpedia.1815

Mingers J. 1995. *Self-producing systems: implications and applications of autopoiesis*. Berlin: Springer Science & Business Media.

Mossio M, Bich L. 2017. What makes biological organisation teleological? *Synthese* **194**(4): 1089–1114. https://doi.org/10.1007/s11229-014-0594-z

Nagel E. 1977. Goal-directed processes in biology. *The Journal of Philosophy* **74**(5): 261–279. https://doi.org/10.2307/2025745

Nunes-Neto N, Moreno A, El-Hani CN. 2014. Function in ecology: An organizational approach. *Biology & Philosophy* **29**(1): 123–141.

Peter S, Dittrich P. 2011. On the relation between organizations and limit sets in chemical reaction systems. *Advances in Complex Systems* **14**(1): 77–96. https://doi.org/10.1142/S0219525911002895

Powers WT. 1973. *Behavior: the control of perception*. Chicago: Aldine.

Powers WT. 1995. The origins of purpose: the first metasystem transitions. *World Futures* **45**(1/4): 125–137. https://doi.org/10.1080/02604027.1995.9972556

Razeto-Barry P. 2012. Autopoiesis 40 years later. A review and a reformulation. *Origins of Life and Evolution of Biospheres* **42**(6): 543–567. https://doi.org/10.1007/s11084-012-9297-y

Rosenblueth A, Wiener N, Bigelow J. 1943. Behavior, purpose and teleology. *Philosophy of Science* **10**(1): 18–24.

Rubin S, Veloz T, Maldonado P. 2021. Beyond planetary-scale feedback self-regulation: Gaia as an autopoietic system. *Biosystems* **199**: 104314. https://doi.org/10.1016/j.biosystems.2020.104314

Sommerhoff G. 1950. *Analytical biology*. Oxford: Oxford University Press.

Sourjik V, Wingreen NS. 2012. Responding to chemical gradients: bacterial chemotaxis. *Current Opinion in Cell Biology* **24**(2): 262–268. https://doi.org/10.1016/j.ceb.2011.11.008

Sternberg S. 2010. *Dynamical systems*. North Chelmsford, MA: Courier Corporation.

Strogatz SH. 2000. *Nonlinear dynamics and chaos: with application to physics, biology, chemistry, and engineering*. Boulder, CO: Westview Press.

Thompson, N. S. 1987. The Misappropriation of Teleonomy. In P. P. G. Bateson & P. H. Klopfer (Eds.), *Perspectives in Ethology: Volume 7 Alternatives* (pp. 259–274). Springer US. https://doi.org/10.1007/978-1-4613-1815-6_10

Trestman MA. 2012. Implicit and explicit goal-directedness. *Erkenntnis* **77**(2): 207–236. https://doi.org/10.1007/s10670-012-9379-2

Veloz T. 2019 (2020). The complexity–stability debate, chemical organization theory, and the identification of non-classical structures in ecology. *Foundations of Science* **25**: 259–273. https://doi.org/10.1007/s10699-019-09639-y

Veloz T, Razeto-Barry P. 2017a. Reaction networks as a language for systemic modeling: fundamentals and examples. *Systems* **5**(1): 11. https://doi.org/10.3390/systems5010011

Veloz T, Razeto-Barry P. 2017b. Reaction networks as a language for systemic modeling: on the study of structural changes. *Systems* **5**(2): 30. https://doi.org/10.3390/systems5020030

Veloz T, Maldonado P, Busseniers E, Bassi A, Beigi S, Lenartowicz M, Heylighen F. 2022. An analytic framework for systems resilience based on reaction networks. *Complexity*. https://researchportal.vub.be/en/publications/an-analytic-framework-for-systems-resilience-based-on-reaction-ne

Wadhams GH, Armitage JP. 2004. Making sense of it all: bacterial chemotaxis. *Nature Reviews Molecular Cell Biology* **5**(12): 1024–1037. https://doi.org/10.1038/nrm1524

Walker B, Holling CS, Carpenter SR, Kinzig, A. 2004. Resilience, adaptability and transformability in social–ecological systems. *Ecology and Society* **9**(2): 5.

Walsh D[M]. 2012. Mechanism and purpose: a case for natural teleology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* **43**(1): 173–181. https://doi.org/10.1016/j.shpsc.2011.05.016

Walsh DM. 2018. Objectcy and agency: towards a methodological vitalism. *In*: Nicholson DJ, Dupré J (eds), *Everything flows*: towards a processual philosophy of biology. Oxford: Oxford University Press.