Organisms as Agents of Evolution

April, 2023

By Philip Ball

JOHN TEMPLETON FOUNDATION

Inspiring Awe & Wonder

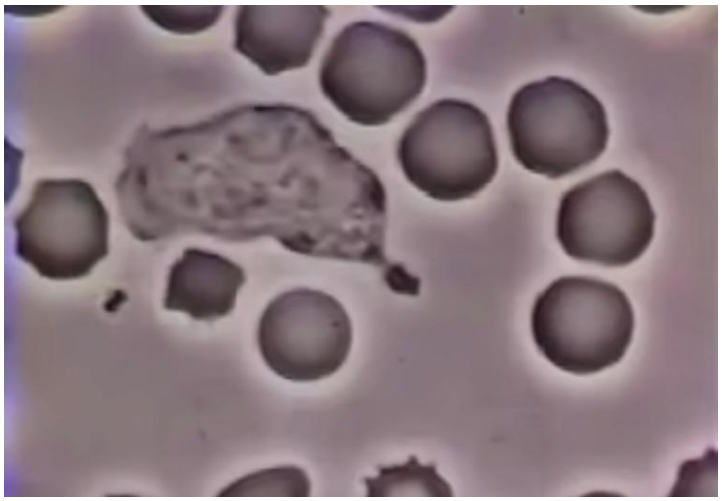
Table of Contents

Introduction 3
I. What is Agency? 6
II.Agency as Goal-Directed Autonomy.....9
III. How to Theorize About Goal-Directedness and Agency.....12
IV. Models of Agency.....16
V.Thermodynamic Origins of Agency.....21
VI. Agency and Evolution 25
VII. Collective and Multicellular Agency.....29
VIII. Agency, Engineering, and Ourselves 31
Conclusion.....33
Bibliography.....34

Introduction

Agency – the capacity to make goal-directed changes to one's self and environment – seems to be a real and general characteristic of living organisms. Yet unlike other general features such as replication and metabolism, we lack widely accepted models or theories of what agency is and how it arises. Do modern biology and evolutionary theory need them? If so, what might they look like?

In the late 1950s, biologist David Rogers of Vanderbilt University recorded a movie using an optical microscope in which a human immune cell called a neutrophil crawls amid red blood cells in pursuit of a single Staphylococcus aureum bacterium. After chasing the zigzagging bacterium for some time, the relentless neutrophil catches its prey, engulfs and devours it.



Agency in action? A neutrophil (large amorphous cell) "chases" a bacterium (small dark dumbbell) amongst red blood cells (circular cells). [https://routledgetextbooks.com/textbooks/9780815344506/videos.php]

This, at least, is how observers typically interpret what they are seeing. The immune cell is presumably sensing and following a chemical trail of some kind exuded from the bacterium, but it is nigh impossible to watch the movie and not frame it mentally with the narrative of a predator and its prey, each trying to out-maneuver the other. The movie seems to validate what Austrian biologist Karl Ludwig von Bertalanffy (a founder of the discipline of general systems biology, which drew on ideas from thermodynamics and cybernetics) said in 1969: "you cannot even think of an organism… without taking into account what variously and rather loosely is called adaptiveness, purposiveness, goal-seeking and the like." (Bertalanffy 1969).

Should we even try to do so? The narrative the mind imposes on Rogers' movie feels dangerously anthropomorphic, seeming to attribute to mere cells the kinds of capacities and behavioural drivers we might normally associate with the world of large and cognitively complex animals. But it is not obvious why, if we are prepared to recognize purpose, goals and agency in our own behaviour, we should make it inadmissible for "simpler" organisms. Perhaps, rather than wondering if we are reading too much purpose and agency into the neutrophil chase, a better question is to ask how agency might manifest both similarly and differently at different scales in time and space.

The predominant tendency in modern biology is, however (and pace Bertalanffy), to deny any need to invoke agency at all – to suppose that cells and bacteria, and perhaps the vast majority of living things, can be regarded as sophisticated machines. As Walsh (2015) says, "Organisms are fundamentally purposive entities, and [yet] biologists have an animadversion to purpose."

The reasons for this aversion are complex. The "reductionistic turn" of a great deal of modern biology, in which explanations for phenomena are ultimately traced to the properties and interactions of molecules, does not obviously generate any prescription for agential concepts. If the causes of all biological phenomena flow from the bottom up, and one cannot reasonably attribute agency or purpose to molecules, how then can these attributes at higher levels be anything more than illusory? But such a case, once it is so explicitly stated, is easily demolished. Atoms and molecules do not (most would agree) possess consciousness or emotions either, but we do not in consequence feel compelled to dismiss these as real organismal features, relegating them to mere "as if" appearances.

Partly, too, the problem is that, in the absence of any widely used tools or concepts for describing properties such as agency and purposiveness, they are often said to be "emergent" in a somewhat ad hoc and vague manner. Without a theoretical framework for handling agency (indeed, without an agreed definition of what it is), that facility seems bound to appear rather tenuous and otiose.

There is also a common suspicion that talk of goals and purpose renders biology teleological, thereby raising notions of design and of some over-arching "plan". Invoking agency as a real biological phenomenon seems to some to open the door to forms of mysticism, such as the old notion that biological systems possess a vitalistic essence that sets them apart from abiotic systems. Worse, notions of purpose and goals might invite quasi-religious explanations of phenomena in the manner of intelligent design: to impute a kind of cosmic agency.

Traditionally, no role has been assigned to or deemed necessary for agency in what is often regarded as biology's unifying schema: Darwinian evolution. Yet despite the successes of the gene-centred view of evolution, as encapsulated in the Modern Synthesis of genetics and Darwinism anointed by Julian Huxley in 1942, there remains disagreement about its causal structure or the scope of its explanatory power. Lewontin (1974) argued that

"To concentrate only on genetic change, without attempting to relate it to the kinds of physiological, morphogenetic, and behavioural evolution that are manifest in the fossil record and the diversity of extant organisms and communities, is to forget entirely what it is we are trying to explain in the first place."

In the Neodarwinian Modern Synthesis, organismal behaviour and morphology become epiphenomena of the competition among gene variants for replicative success. In this view, there is no compelling reason to recognize agency as anything more than a collection of adaptive behaviours. Yet attempts to describe and explain agency as a real biological phenomenon – akin, say, to immunity, metabolism, or cognition – need not obviously pose a threat to the Modern Synthesis (although some argue that such an account will demand that this picture be modified or extended). They simply refocus the biological lens so that the primary question becomes how living entities (including neutrophils) do what we see them do. As Lewontin implies, this demands a shift from a gene-centred to an organism-centred view.

In this survey I shall discuss some of the approaches that have been taken to naturalize agency: to accept it as a real phenomenon that can be understood without appeal to extra-physical influences. I shall also examine how notions of biological causation and evolutionary change might look different if agency is admitted as a real property of living things.

Back to Table of Contents

I. What Is Agency?

Agency is defined by Webster's dictionary as "the capacity to act or exert power", and in robotics and AI research a system that can act in any way in response to environmental stimuli is sometimes considered agential. But in biology, typically something more is demanded. The definition offered by Sultan et al. (2022) is typical: they say biological agency is "the capacity of a system to participate in its own persistence, maintenance and function by regulating its own structures and activities in response to the conditions it encounters." The several definitions listed by Moreno (2018) are similar, and many mention the goal-directedness of agents and their interactions with their environment. It may be that we should seek no harder than this: that a cluster of overlapping definitions enables a more fruitful and inclusive investigation than a premature attempt to impose rigid boundaries.

In the broadest view, agency might be seen as one of the defining characteristics of living entities. Whereas typical definitions of life tend to invoke capabilities such as metabolism, replication and evolution, the notion of agency describes the ends to which such capabilities are put. Agency frames the living entity as a doing thing. At the same time, understanding agency must place the focus not on the what of doing, but the how. "Agency is thus not about all of the many and varied things that organisms do—from building anthills to caching nuts—but rather about how they do them", says Tomasello (2022). "Individuals acting as agents direct and control their own actions."[1] While the existence of biological agency seems intuitive, does the notion truly add anything to biology that is not explained by a reductive, mechanistic account of how its parts interact? What about organisms is not understood that a theory of agency is needed to explain? How would biology look different if it recognized agents as real entities?

To find answers, we first need to be clear about which entities possess types or degrees of agency. Some insist that it obtains only if the agents display deliberate intention, perhaps even consciousness. But here Aristotle's injunction is surely still worth heeding: "It is absurd to suppose that purpose is not present because we do not observe the agent deliberating." [Physics II: 8])

The anthropocentric viewpoint that humans, if not uniquely agential, have a special variety of it, has been the traditional one historically (Riskin 2016). Aristotle distinguished humans from other living beings by the fact that we alone possess a rational soul: the capacity to reason. For Descartes, meanwhile, agency as a behaviour distinguished from machine-like stimulus-response (even if that included feelings and emotions) was exhibited only by humans, by virtue of our immortal soul: this supplied the theologically necessary capacity to *choose*.

Such distinctions are less apparent (if not absent entirely) today, when human behaviour and attributes are considered particular cases of more general features of living things. Many behavioural and cognitive scientists even now regard consciousness as a matter of degree, shared by at least some other

[1] As Love (2023) points out, this then forces us to ask for definitions of "individual" – far from trivial in biology – as well as "direct and control".

metazoan species. In any event, biological agency is rarely considered now to be contingent on conscious intention.

What is less clear is whether life is "agency all the way down". Might it, for example, demand a nervous system capable of formulating and seeking to attain goals? Even bacteria can be considered decision-making entities (Ben-Jacob *et al.* 2014) whose behaviour depends not just on external circumstances but on their own internal state, informed by external data gathered through sensory systems. Some consider that this mode of operation be best regarded within the framework of cognition rather than mere mechanism. Agency too might be regarded as a matter of degree, extending in some measure even to the simplest forms of life.

Perhaps, however, it goes no further. Moreno (2018) argues that non-living systems such as cellular automata (computational simulations in which the states of cells on a grid depend on those of their neighbours) and active matter (non-living particles that have some means of propulsion) do not display true agency because the individual component "particles" cannot be considered to possess goals. Viruses, meanwhile, lack the autonomy that characterizes true agency. Whether true agency (as opposed to a simulacrum of it imposed by the designer) can be designed into systems such as artificial intelligence remains to be seen – but to assess the prospects clearly we will need a better understanding of wherein this property resides.

One answer to the question of what a theoretical framework for conceptualizing agency might bring to biology is that it might foster more predictive capability. The distinction between physics and biology is sometimes illustrated via the thought experiment of repeating Galileo's (almost certainly apocryphal) Tower of Pisa experiment by dropping a cannonball and a pigeon. The trajectory of the cannonball is wholly predicted by the Newtonian laws of motion; the same cannot be said for the pigeon, even though it does not violate any physical laws. To the extent that we can predict what the pigeon will do at all, we implicitly invoke its agency. To explain why it does not simply plummet, it is not enough to invoke aerodynamics; we must also in effect allow that the pigeon does not *want* to plummet. It manifests its agency by virtue of having goals.

Walsh (2015) distinguishes *object theories*, which describe the behaviour of objects according to laws external to the system (typically Newton's laws), and *agent theories*, in which actions are events "that occur as a consequence of agents' pursuit of their own purposes" and are internal to the system. Even for a relatively simple and well-studied biological phenomenon such as bacterial chemotaxis – the movement of bacteria in response to a gradient in chemical concentration – the exact function and mechanism of such agential, goal-directed behaviour is not fully understood, nor is it wholly predictable on the basis of the stimuli alone (Neilson *et al.* 2011; Samanta *et al.* 2017).

The challenge of identifying and explaining agential behaviour is not a mere academic exercise. If, for example, we try to treat a cancer cell as an object that behaves as it does because of causal laws governed by genetic mutations, then judging from present experience we may not get very far towards finding a cure. If our theory incorporates some notion of the cancer cell as an agent interacting with its environment (including other cancer cells), with goals that involve not just survival and replication but

also development, and with all the responsiveness and adaptation of its internal state that this entails, we might do better (Sonnenschein & Soto 2020). As Barbara McClintock expressed it in her 1983 Nobel lecture, "A goal for the future would be to determine the extent of knowledge the cell has of itself and how it utilizes this knowledge in a "thoughtful" manner when challenged."

Sultan *et al.* (2021) argue that an understanding of agency would allow the role of the environment to be better incorporated into biology. Currently, environment tends to feature in evolutionary theory as a "given" to which an organism must respond. Of course, it is understood that biological agents may alter their surroundings – but there is no systematic way of theorizing about that process. Agents might deplete resources, but also enrich them, for example through excretion of nutrients. They might restructure the environment in more profound ways, altering local climate or geomorphology. Some of this is incorporated in a rather ad hoc fashion into the evolutionary concept of niche construction or the notion of an "extended phenotype" (Dawkins 1982). But there is no systematic way of predicting or describing such interactions or the principles underpinning them.

For example, an agent intent on self-preservation and maintenance might conceivably respond to environmental change (a rise in temperature or salinity, say) in several ways:

- By developing a capacity to buffer its internal states against external fluctuations or shifts.
- By activity that restores the previous environmental conditions in a homeostatic manner (Dyke & Weaver 2013).
- By migrating to a different environment with more amenable conditions (as in chemotaxis).

Which of these strategies is preferred in a given circumstance? There is no general framework for answering that question.

Walsh (2015) argues that, in the end, recognizing and explaining agency is not just an instrumental desideratum; ultimately it is a part of the scientific quest to understand. "If agency is a real, natural phenomenon, and our scientific theories cannot countenance it, then our understanding of the world is destined to be impoverished".

Back to Table of Contents

II. Agency as Goal-Directed Autonomy

An agent does more than just alter its environment. After all, many non-living dynamical structures do that much: the Sun warms the planets, a cyclone wreaks destruction. Both, like living organisms, are entities that persist out of equilibrium (Schrödinger 1944; Nicolis & Prigogine 1977). But while many non-living far-from-equilibrium systems, such as patterns formed from convection currents in the ocean and atmosphere, sustain ordered structures and dissipate energy, only biological agents seem capable of relatively open-ended and improvisational behaviour governed by goal-directedness. They alter with intent, apparently informed both by learning and by the capacity to innovate. Programming a robot, in contrast, can imbue it only with a kind of pseudo-agency (much as AI may display a form of pseudo-intelligence): the goal is imposed by the (agential) designer rather than being internally generated.

Thus an agent is generally considered to have autonomy – it is self-determining, in particular by actively maintaining its very identity. This self-sustaining nature is, indeed, often considered the agent's ultimate goal (Veloz 2021), and has been implicated as a fundamental and necessary characteristic of life (Maturana & Varela 1980; Muñuzuri & Pérez-Mercader 2022). That feature in turn implies the existence of a boundary that separates self from environment: an agent must be bounded.[2] Moreno (2018) argues that a true agent does not simply persist in an environment but does so by altering the environment *for that purpose*. Certainly it is hard to imagine how an agent could exist as such otherwise. As a non-equilibrium system, it must absorb energy from its surroundings and dissipate it, and typically this flux of energy will be accompanied too by a flux of matter, as agents take in material resources (such as energy-rich food) and discharge waste products.

To the extent that goals and purpose have been admitted into biology at all, apparently purposive behaviour has typically been attributed to genetically prescribed mechanisms of self-preservation. Such a wholly mechanistic account seems to demand no more genuine purpose than does the motion of pistons in an engine. Rather, purpose then appears a "stance", akin to Dennett's "intentional stance" (Dennett 1987): it need be no more than a convenient manner of speaking.

Dennett's agential stance itself remains agnostic about the ontological status of goals, but merely asks if invoking them is predictively useful:

"First you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs."

If we assume that the pigeon has the goal of not falling to strike the ground, we can at the very least predict that it will not share the cannonball's trajectory. Whether (or how) this goal is literally

[2] The possibility of collective (for example, symbiotic) agency complicates that criterion, and is discussed later.

represented in the pigeon's brain is a separate question; the point is to understand why treating some entities as agential in this manner offers a predictive advantage that is not apparent for others.

Biology has not necessarily rejected purposiveness per se. According to Monod (1971), it is "one of the fundamental characteristics common to all living beings without exception… that [they are] *objects endowed with a purpose or project*." Mayr proposes that biological entities "owe [their] goal-directedness to the influence of an evolved program" – typically equated with the genome, although Mayr admits other possible sources too. The problem is that such "programs", when defined so broadly, say nothing more than that organisms have goals and means by which the organism might try to attain them. They do not uniquely specify a mechanistic cause-and-effect sequence of events.

To talk about goals and purpose is, however, usually to invoke a teleological picture. This has generally been seen as a step too far. In Mayr's view (2004), the word teleology implies an Aristotelian "final causation" that, in an evolutionary context, suggests a drive towards improvement or perfection: a directionality seemingly refuted by Darwinian natural selection.

To avoid the discomfort provoked by teleology, Pittendrigh (1958) proposed the concept of "teleonomy", which imputes only a kind of mechanistic lawfulness to apparently purposive behaviour. The term was adopted in the 1960s and 70s by biologists such as Ernst Mayr and Jacques Monod, but never really caught on – in the view of Dresow and Love (2023) for good reason, since they say it was never clearly defined in any event, introduces more confusion than clarity, and was motivated only by a "teleophobic" refusal to accept goals and purposes as real factors in biology.

That discomfort around purpose extends also to the concept of function. It might seem relatively uncontroversial to say that, for example, the function of the enzyme alcohol dehydrogenase is to remove a hydrogen atom from alcohols, or that of a heat shock protein is to protect an organism from temperature fluctuations in the environment. It seems reasonable to say too that the function of the respective genes is to encode these proteins. But even the notion of molecular functionality can be controversial, as shown for example in arguments about the revelation by the international ENCODE project that a large amount – perhaps as much as 80% (although later estimates offer figures of more like 30%) – of the human genome is transcribed to RNA, even though only around 2% codes for proteins (ENCODE 2012). Some argued that a molecule can only be awarded genuine functionality by evolution: it has a function only if it is evolutionarily conserved, which many of the transcribed sequences are not.

Mayr points out, however, that "function" may have two rather different meanings in biology: one descriptive, the other teleological. Alcohol dehydrogenase might be said to have the function of catalysing the conversion of alcohols to ketones or aldehydes; or it might be said to have the function of detoxifying alcohols, a description that keeps in view the wider goal of organismal survival and evolutionary roles. The latter account is inherently teleological, but evidently serves a useful explanatory purpose (without imputing any such "goal" to the enzyme itself). The ENCODE dispute highlighted the lack of consensus on what qualifies a genetic element, or more generally, any molecule in the cell, as functional. As Guttinger and Love (2023) say, one can ask either "what a part is doing

presently in a system" (a biochemical or developmental question) or "why a part is present in the system" (an evolutionary question). The two need not deliver the same answer, for they draw on different methodologies. The same consideration applies to any component of living systems.

Much of the discussion of goals, purpose and teleology in biology has happened within the context of evolution. Development and behaviour feature far less in the debate, for example in terms of how morphogenesis seems to unfold as a directional process with an "ideal" end plan (which might or might not be actually realized). Descriptions of developmental processes are apt to draw on complex genetic wiring diagrams or schematic illustrations of signalling pathways, but with no real sense of what the generic features are that characterize and enable the reliable unfolding of the process.

Are there genuine developmental goals? Large, complex animals tend to acquire a particular body plan during development which is robust in the face of, say, low-level molecular randomness and chance disruptions from the environment. Such robustness is perhaps most spectacularly displayed by the planarian flatworms, which will regenerate an entire body from a dissected fragment (Reddien & Sánchez Alvarado 2004). Experiments by embryologist Gerhard Fankhauser in the 1940s on development of the amphibian pronephric duct (the progenitor of the kidney) in organisms with cells enlarged to various degrees due to an excess of chromosomes showed that the tubular morphology adapted to variation in cell size (Kirschner *et al.* 2000; Kirschner & Gerhart 2005). It is as if the cells collectively "know" what their target structure is and adjust their individual behaviour accordingly (Wolpert 2010; Levin *et al.* 2019).

Such resilience is generally not perfect. Genetic anomalies, environmental disturbances, and chance fluctuations of the developmental process itself can lead to growth defects. Even here, however, the outcomes tend not to be arbitrary: rather, in general development is *canalized* – a concept introduced by Waddington in the 1940s, who described it in terms of a morphological landscape (Waddington 1957). Such a dynamical evolution conveys robustness of outcomes in the face of fluctuations – but does this qualify as a genuine agential goal, or something more akin to a thermodynamic imperative? I return to that question below.

Back to Table of Contents

III. How to Theorize About Goal-Directedness and Agency

An aversion to admitting goals and purposes as more than just a "stance" has stymied an understanding of how to incorporate them into biology. Jaeger (2021) asserts that

"Because of the widespread mechanistic distrust concerning the notion of purposiveness, we do not possess the conceptual and mathematical tools required to appropriately incorporate true organismic agency into models of evolutionary dynamics. This is why we'd rather pretend the phenomenon does not exist, rather than taking it seriously."

If we are willing to take goal-directed agency seriously in biology, where does it come from? How are goals selected? And what "machinery" is needed to enable such agential behaviour?

Efforts to theorize how goals and intentionality arise – to "operationalize" these notions – are still in their infancy (Atlan 1998, 2007; Barandiaran *et al.* 2009; Veloz 2021). Typically they consider goals as emergent properties of generic complex systems, such as the ability of reactive networks to develop self-contained and self-sustained dynamical states. Kauffman (2000) claims that "An autonomous agent must be an autocatalytic cycle able to reproduce and able to perform one or more thermodynamic work cycles" – loosely paraphrased, it must be able to do something and return to its original state afterwards.

The "goal" of such a system is often considered to be simply to keep existing. Systems capable of such emergent behaviour are by definition ones that persist, and the question becomes about what properties enable such self-sustaining dynamics. But these approaches remain too abstracted from biological entities, or from quantitatively testable hypotheses, to offer much insight into how living things set their own goals. They do not obviously suggest how such goals arise from the interaction of a historically situated organism with its environment.

It is central to the notion of agency that a particular organismal goal does not by itself determine the route or mechanism by which it is attained. This is why Mayr's notion of a controlling program, however broadly conceived, does not really work: a mechanistic step-by-step set of instructions is too fragile to error. The whole point about agency is that it can be versatile, adaptive and improvisational. Agency evolves precisely because living organisms are liable to encounter challenges that evolution itself is too slow to adapt to. Here, then, is the fundamental tension between an agential view of biology and the traditional Neodarwinian view. In the latter, chance predominates: if the environment changes, only those organisms that happen to have a beneficial adaptation survive. But agents have, as it were, some say in their own fitness. Indeed, one can argue that natural selection seems bound to produce that kind of agency: an organism with the capacities to adapt behaviour to a range of environmental changes seems sure to be more fit than one that must hope to be rescued from stress or danger by pure good fortune. If agency is something that indeed pertains even to prokaryotes, we might wonder if it *necessarily* outperforms strictly mechanistic and prescriptive stimulus-response rules for sustaining life on Earth.

To open the black box of agency itself, we first need to be more precise about what characterizes an agent. We have seen already that agents must be separable from their environment, with a boundary demarcating the division. At the same time, this boundary must be permeable – in the case of a cell membrane, literally so. An agent must be able to control its interior composition and state – in general, to make it *thermodynamically* as well as compositionally distinct from its surroundings – but also to permit the flux of matter and energy that enables it to maintain a non-equilibrium state. This division supplies an agent with a definition of self.

An agent must persist for some meaningful duration of time. This seemingly obvious requirement reminds us that what characterizes a living organism is not the atoms it contains but its pattern of spatiotemporal organization at a higher level. Agency simply has no meaning as an instantaneous property: it unfolds in time and space, each with characteristic scales. There are time and space horizons – a certain *perspective* – within which an agent has the capacity to act: for all our own impressive, self-conscious agency, we typically have to rely on that of our immune cells to combat microbial pathogens, for example. There are also limits imposed by sensory modality: one might argue that vision is more valuable than smell to our own agency. Such modalities constrain the production of meaning: smell is not just more acute for dogs but *means* more. Viscosity means something different to a fish than to a bacterium: it offers the two organisms very different affordances (see below).

In addition to these features, Potter and Mitchell (2023) argue that an agent must have endogenous activity, meaning that it does things "for its own reasons", and not just in a stimulus-response manner. A piston has no endogenous activity: it just responds passively and predictably to a change in gas pressure in the chamber. To put it another way, what goes on *inside* an agent is influenced but not fully determined by what happens *outside*. Cells are not just bags of inert molecules until some signal arrives at the cell surface to prompt them into action; those molecules are constantly interacting and reacting to maintain the cell's integrity, and external signals just nudge that activity.

A corollary of this criterion is that agents must have some internal complexity. A gene can't have real agency because it lacks this "inner life": it simply does not have enough internal degrees of freedom. Proteins might be better candidates for a weak form of agency because they have more complex structure and dynamics that are pertinent to function: their binding and catalytic action may depend on molecular vibrations and interactions among parts of the peptide chain, for example that may convey allosteric activity. A genome has a still better claim to a degree of agency, being intimately connected to a highly complex and regulated three-dimensional structure. But it has little real autonomy: its behaviour is closely coupled to and dependent on events in the rest of the cell. It is really only at the whole-cell level that biological agency fires up in earnest.

Potter and Mitchell add that agents must also show "holistic integration": they are more than the sum of their parts. We can usefully take an organism apart and look at the components – but we can't truly understand what it *does* unless we put them back together again. As Potter and Mitchell say, there's a distinction to be drawn here with a machine that is, so to speak, just "pushed around by its own component parts". It might sometimes look as though this is the case for living organisms (a gene is activated and the cell state changes, say), but in fact the molecular-scale parts are themselves also altered and governed by higher levels of the system. (For example, which protein is produced by a

given gene, through the editing of the intermediary RNA transcript by the so-called spliceosome, may depend on the state of the whole cell, differing in different tissues.) "You cannot horizontally reduce such a system to identify a particular part (or set of parts) that is determining the system's next state", say Potter and Mitchell, "because the activity of that part is, itself, being determined by all the other parts in the whole." The agent thus exhibits a kind of organizational closure (Mossio & Moreno 2010).

What most distinguishes agents is that they have *reasons* for actions, which in turn elicit value judgements: a primitive notion of meaning. These words might sound anthropomorphic – how can there be reasons without true reasoning, or meaning without emotions? But therein lies the challenge to operationalize these terms. When an organism has integrated its inputs in the light of its own internal state, including perhaps its internal representations of the environment, and from this process has selected a response from the palette of actions available, *reason* seems a defensible word to attach to that decision, regardless of whether any awareness is involved. And the choice of which stimuli to attend to, and the weighting attributed to them in motivating a response, might indeed be regarded as the production of *meaning*.

The agent will (again without any necessary attribution of conscious intention) ascribe value and valence to those aspects of its environment that can serve its purposes, or what Barandiaran *et al*. (2009) call *sense-making*. A chemotactic bacterium moves towards higher concentrations of nutrient but will ignore concentration gradients of substances that have no nutritional value (so long as they do not threaten survival). In this process, Ginsburg and Jablonka (2008) suggest, a valenced "proto-feeling" might have accompanied such evaluation even in the earliest multicellular organisms. Bray (2009) claims that a primitive *awareness* of the environment was an essential ingredient in the very origins of life. He calls cells "touchstones of human mentation": a kind of minimal model of what cognition can and should mean.

Such selectivity towards environmental signals can be imprinted by evolution, or it can be learnt. Some plants, for instance, exhibit habituation, whereby a stimulus that initially provokes a response is later ignored when it proves to pose no threat (Gagliano 2017). The stimulus is still present, but for the plant it has lost its meaning. Habituation exemplifies how higher-level reasons, derived from purposes and goals, may free an agent from automatic stimulus-response behaviour. Low-level events with the potential to trigger some response might be ignored if that response conflicts with the higher-level goal. By the same token, identical stimuli might produce different results owing to the *internal* state of the agent. Thus, say Potter and Mitchell,

"Any attempt to understand or explain the causes of an organism's behaviour is doomed to fail if it takes a purely instantaneous view of the physical system. It is not enough to account for how an organism behaves upon detecting some external stimulus or physiological state of affairs – the 'triggering cause'. We must also understand why the system is configured such that it behaves in that way – the 'structuring causes'."

In other words, for agents *history matters*. Agents hold a memory of salient aspects of past events (with associated retention timescales) that may determine future actions. History has configured and primed them, embodying within them both goals and a pragmatic kind of "knowledge of the world" that directs their action. By doing so it has invested them with a causal power that cannot be reduced to the

sum of its parts. Agents experience the world as a genuine web of meaning, which might be best expressed in terms of *affordances* (Gibson 1979; Walsh 2015): how, given this state of affairs, might I best achieve my goal? What transformations of both self and environment might I effect to that end? What is useful to me in that quest?

The criteria of agenthood adduced by Potter and Mitchell might be regarded as an elaboration of those proposed by Barandiaran *et al.* (2009), who condense them to just three:

- Individuality: a physical and thermodynamic distinction between self and environment.
- Interactional asymmetry: the agent is to some degree self-determining, being the genuine cause of its actions, and is not merely "pushed around" by its environment.
- Normativity: the agent has goals.

Interactional asymmetry typically entails the management of fluxes of energy and matter, for example, in the way a cell maintains its chemiosmotic state by actively pumping ions against a concentration gradient. Barandiaran *et al.* admit that interpreting such processes in causal terms can be problematic – do the minimal movements of a bird's wing truly cause its gliding trajectory, when they merely perturb the aerodynamics? – but they present the asymmetry instead in terms of an ability to purposively *modulate* couplings to the environment. An agent can (within certain spatiotemporal horizons) exploit what the environment can offer it by "steering".

Listing criteria for agency runs the risk of becoming like a list of criteria for life: they can seem arbitrary, and one is never quite sure if the list is comprehensive, necessary or sufficient. But Barandiaran *et al.* suggest that their three characteristics are interrelated by the fact that they "share in common an essential role played by the inner organization of the agent", and that they are compliant and consistent with the minimal features commonly adduced for living systems. A living organism remains so *only* if it acts constantly to maintain itself, for it is "permanently precarious" (Di Paolo 2009). It does not first exist and subsequently acquire a goal; having the goal of self-maintenance is a condition of its existence, its individuality. And that goal can *only* be accomplished through the property of interactional asymmetry – not just to resist disruption from perturbations, but to avoid equilibrating with the environment. Thus, say Barandiaran *et al.*, "Minimal life forms already come to satisfy the necessary and sufficient conditions for agency."

Back to Table of Contents

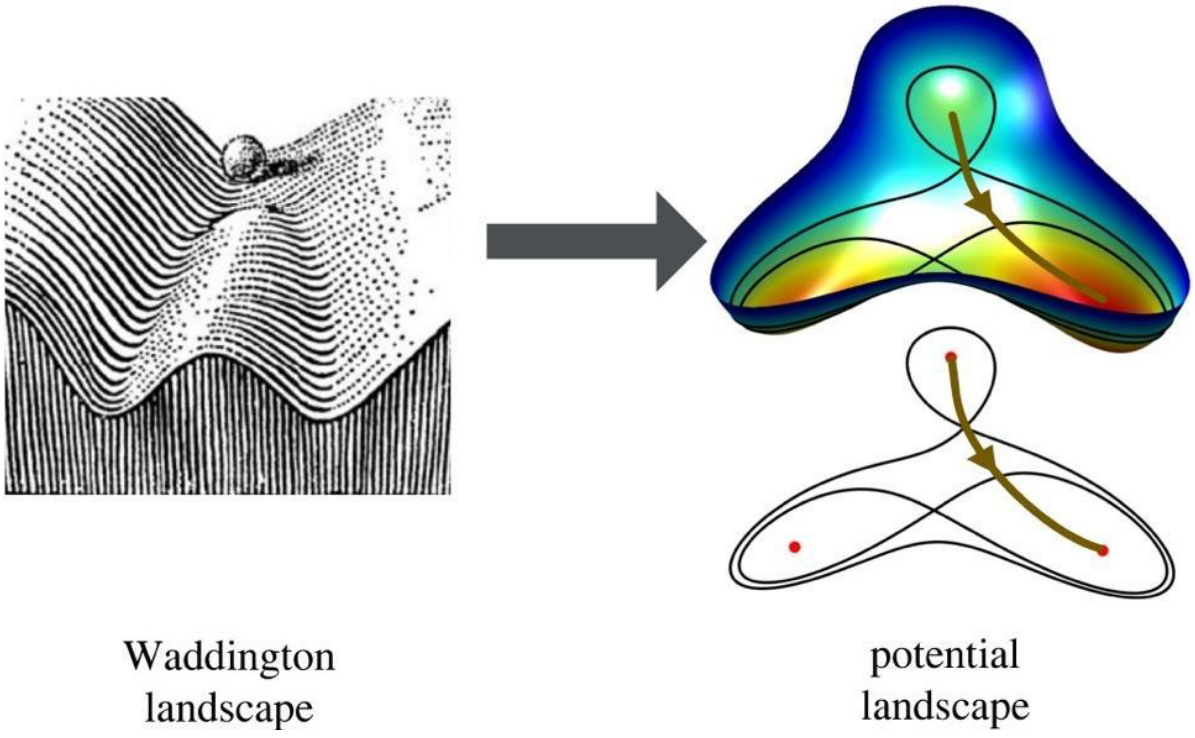
IV. Models of Agency

The problem of agency poses two central questions: how are goals determined, and how are they targeted by goal-directed behaviour? The second question has received the most attention so far. Attempts to operationalize goal-directedness date back at least to the suggestion of cyberneticists in the 1940s (Rosenblueth *et al.* 1943) that it demands negative feedbacks to keep the goal-directed entity "on course". Nagel (1979) suggested that it can be decomposed into two independent tendencies: persistence and plasticity. As elaborated by Lee and McShea (2020), persistence refers to a tendency to return to the goal-directed trajectory after a perturbation, while plasticity is the tendency to find a path to a given goal from many different starting points. While they do not claim this as a unique partitioning of goal-directedness, the metrics can be quantified for some goal-directed behaviours (such as bacterial chemotaxis), most obviously ones where the goal is spatial and the behaviour is motional (and involves attraction to the goal).

One of the challenges for theories of goal-directedness, say Lee and McShea, is to constrain definitions so as to include all we might intuitively want to include and exclude in like fashion. For example, there is some persistence in the motions of the planets: a planetary orbit is predicted to recover its original form if temporarily perturbed by, say, a passing comet. This might look teleological – as though the system "wants" to stay in its original state – but is simply a consequence of there being stable *attractor states* of the dynamical system; it is basically the same as a ball displaced from the centre of a bowl rolling back towards it.

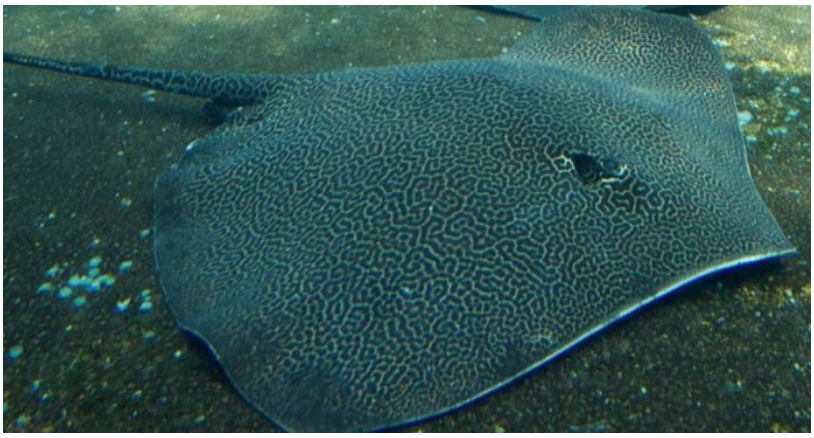
Something analogous governs the developmental trajectories of metazoan cells – an idea foreshadowed by Waddington's landscape metaphor (see above) but which can now be formulated in quantitative terms as a multidimensional space of gene expression levels (Saéz *et al.* 2022). It appears that in general this abstract space collapses to a low-dimensional landscape in which just a few genes dominate the topography, with broad basins (attractors) that correspond to particular cell types.

Here, then, the developmental "goals" are the attractors: stable dynamical states of the gene regulatory networks. This creates some robustness that enables cell fates to persist in the face of stochastic variability in expression rates and external perturbations, and also for the cell to find its fate without having to start from a particular state. There are often multiple developmental routes to a given attractor, as revealed for example by single-cell RNA sequencing (a method for looking at all RNA molecules instantaneously transcribed at a given moment) during development (Farrell *et al.* 2018).



Waddington's developmental landscape and its modern expression as dynamical attractor states of cell fates. (From Saéz et al. 2022) Waddington described the states of cells as valleys down which a ball rolls. Every so often it comes to a Y junction – a bifurcation – where it needs to "make a decision". Bifurcations are also features of theories that describe cell-fate decisions in terms of trajectories in dynamical landscapes of gene expression levels.

It remains an unresolved question to what extent these attractors have been selected, in the Darwinian sense, from a wide palette of options, and to what extent they are constrained and determined by fundamental physical principles governing gene interactions in the networks. The same considerations apply to body plans and tissue structures. For example, those structures and patterns formed by reaction-diffusion dynamics along the lines described by Turing (1952) probably have particular morphologies by virtue of the intrinsic dynamics of the self-organizing system, in effect limiting evolution to a few stable options. The outcomes can be fine-tuned by selection only within limits – as we see, for instance, in the array of animal marking patterns believed to be Turing-like in origin, which tend to have generic forms (such as spots and stripes) in many different species (Meinhardt, 1982).



Pigmentation markings on the reticulate whipray and the yellow-banded poison-dart frog come from the same palette of patterns provided by reaction-diffusion systems.

The stability of cell fates thus exhibit both the persistence and the plasticity identified by Lee and McShea. And after all, it would seem strange if evolution were *not* to harness the robustness and organization available "for free" from such physical principles. We should not expect biological agency to have to invent itself from scratch, and perhaps we might wonder if, for such complex systems, that would even be feasible. There is thus likely to be no clear dividing line between "goals" as thermodynamic or dynamical imperatives (Beer 1995) and goals as the outcomes of internal "deliberation" of the organism.

A general approach to goal-directed behaviour with a clear flavour of a thermodynamic imperative is the Free Energy Principle (FEP) developed by Friston and coworkers (Friston 2010). The theory posits that life-like properties, including agency, are an inevitable and emergent property of any dynamical, non-equilibrium system that is ergodic (that is, it explores all of its available configurations) and possesses a "Markov blanket": in effect, a zone that insulates the system from its environment, so that the two are not directly coupled (Kiverstein *et al.* 2022; Friston, 2010).

The FEP supposes that organisms have the goal of keeping themselves in their expected phenotypic and ontogenetic states, and act to minimize the discrepancy between the desired or predicted state of affairs and the one they experience: in a strictly technical sense, the system aims to minimize "surprise". This process is called active inference, and Friston says that it "furnishes an account of (basic or biotic) sentient behaviour that places agency centre stage." The theory has been formulated as a way to understand the behaviour of cognitive, brain-based agents that make inferences encoded in neural states (Buckley *et al.* 2017).

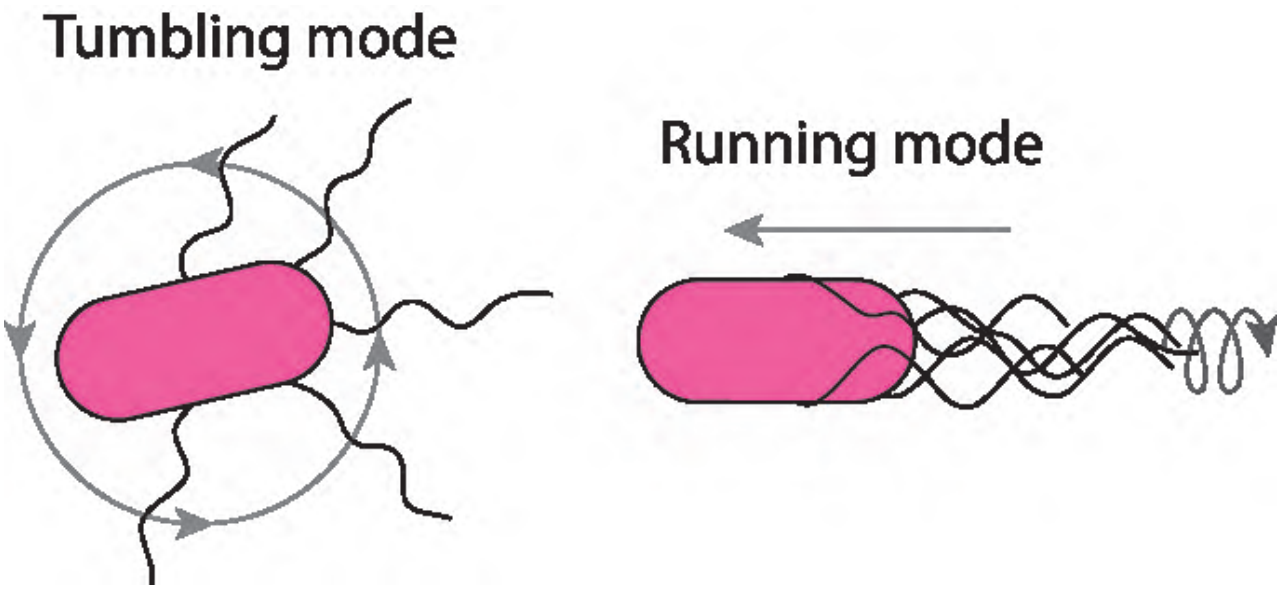
There is continuity here with genuine thermodynamic explanations of change: with how a closed system tends towards an equilibrium state in which entropy is maximized – or equivalently, in which the free energy is minimized. According to the FEP, a non-equilibrium system with a target state (such as the maintenance of self) will seek the most efficient route, typically one defined by considerations of "gradient descent": the path through the available configuration space follows the steepest route to the minimum. In this view, it is not clear whether we should view agency as a real or an apparent property.

As a description of biological entities, the FEP has been criticized for having to trade off generality against realism: it demands too great a degree of abstraction to be able to answer specific biological questions (Colombo & Palacios 2021). Given a particular scenario presented to a somatic cell in development, say, what options does the cell have for regulating transcription so as to satisfy the demands and constraints it experiences? It is not clear that the FEP is able to furnish a means of formulating, let alone answering, that question.

There is some conceptual overlap of the FEP with the qualitative description of agency offered by Walsh (2015) in terms of affordances (Gibson 1979). Walsh argues that, given a goal, an agent will

experience its environment as a menu of affordances: what, in the surroundings, will and will not be useful for attaining that goal? The environment–agent interaction might then be regarded as a landscape that must be navigated to the given end: in effect an optimization problem in which one might expect efficiency to play a role in determining the trajectory.

One possible strategy for an agent navigating its surroundings to its own benefit is to execute random action followed by some "evaluative" process: a biased random walk through the landscape of possibilities that searches stochastically for an optimum. Something like this is exemplified in bacterial chemotaxis. Typically, a bacterial cell follows a gradient of increasing nutrient concentration by exhibiting random tumbling to reset a direction followed by movement ("running") in that direction and sensing of concentration. Tumbling is executed by a burst of uncoordinated "thrashing" of the whiplike appendages called flagellae, orienting the cell in a random direction. This is followed by a period of directional motion in which the flagellae exhibit coordinated movement like a corkscrew-like propellor. During that phase, sensors in the bacterial cell wall measure the concentration of nutrient and determine whether it has increased from one moment to the next. If not, another spell of tumbling is induced. The genetic and protein networks responsible for this pattern of behaviour are now rather well understood (Bray 2009).



This random-walk method is not terribly efficient, but it demands only fairly simple capabilities for motion and local sensing of an environmental variable. Barandiaran and Egbert (2014) suggest that the coupling of a metabolic system to gradient-descent chemotaxis offers a minimal model for the establishment and the active pursuit of norms: the emergence of goal-directed behaviour, in which behaviour becomes positively correlated with the environmental conditions (the "normative field") required to keep the system viable.

The internal resources required for such behaviour are, however, not to be underestimated. In particular, the cells must be able to hold a *memory* of one sensing event so that it might be compared with a subsequent one to evaluate the gradient of nutrient concentration. In this way, cells may formulate a simple (but selective) representation of their environment: according to Bray (2009), "every

cell in your body carries with it an abstraction of its local surroundings in constellations of atoms." The issue of how such representations are encoded, and even how to define them, is currently unresolved, but is likely to be central to a better understanding of both primitive cognition and agency.

Bacteria must attend to more than nutrient concentration alone. Their survival also depends on avoiding extremes of temperature or conditions of low water activity, and other chemical ingredients that might be harmful. Thus the cells must collect other information too, and integrate these inputs to decide on a course of action. It is because of this contingent nature of the response that bacterial agency warrants being considered as a kind of cognition. Implicit in this cognitive aspect of sensing and representation is an ability *not* to act. Information received from the environment does not always necessarily *compel* an action; the outcome depends also on the internal state of the agent. That is illustrated by the case of the waggle dance of the honeybee, by means of which a bee that has located a food source while foraging conveys its location to the other members of the hive (Frisch 1953; Menzel 2019). The information so imparted may or may not induce other bees to seek the source: each bee evaluates the data in the light of its own experience. For agents, sensory input might be best viewed as a suggestion rather than a command.

The notion of a Markov blanket in the FEP offers some sense of how this decoupling of an agent from its environment can occur: the agent has internal states that are *informed* by the environment but, being insulated from it, are not compelled by them. All the same, in the standard formulation of the FEP the dynamics of the internal states are driven by the environment and constrained to mirror it. In contrast, Biehl and Virgo (2022) have developed a model of agency in which the internal states can be used to derive "beliefs" (predictions) about environments that may be totally different from those that actually inform those states. As in the FEP, the agent develops its beliefs using Bayesian inference – the standard principle for updating probabilities as new information becomes available. The agent's beliefs may be "mistaken" – poorly predictive – but it nonetheless acts in a manner that is consistent with those beliefs (and the associated goals).

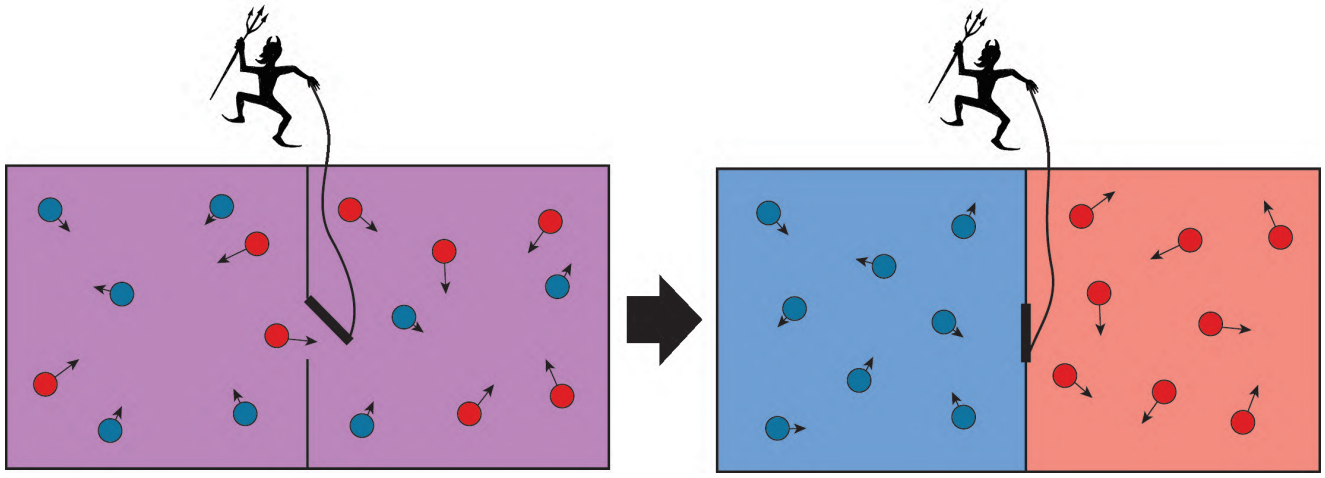
Efficient use of information by agents also typically demands that this information be filtered, for example by insensitivity to noise: a threshold for which a signal is considered meaningful. Mitchell (2023) calls this *causal buffering* and suggests that it demands a hierarchy of representations in which higher levels are insensitive to lower-level fluctuations.

Back to Table of Contents

V. Thermodynamic Origins of Agency

A conceptual approach that connects agency and goal-directedness to cognition, information and thermodynamics can be found in the scenario of Maxwell's demon (Rex 2017). As posited by James Clerk Maxwell in 1867, this hypothetical, microscopic being is able to harness random molecular motions to undermine the second law of thermodynamics – the tendency for all change in a closed system to result in a net increase in its total entropy.

For reasons ultimately theological, Maxwell proposed that this seemingly inexorable law might be subverted by a demon that operates a trapdoor connecting two compartments containing a gas at uniform temperature. By opening the trapdoor to let more-energetic molecules in the statistical distribution pass in one direction, and less-energetic ("cooler") molecules to pass in the other, the demon can accumulate hot and cool molecules in distinct compartments, thereby creating a temperature gradient that can be harnessed to do work. If the trapdoor is frictionless, no energy need be expended in this process: all that is required is information gathered by the demon about the trajectories and energies of the molecules. An energy source has seemingly been produced "from nothing".



By observing molecular motions and selectively operating a trapdoor, Maxwell's demon creates a temperature gradient from an initially uniform gas.

It was pointed out by Landauer, and clarified by Bennett (Bennett 1982; Rex 2017), that the demon cannot subvert the second law indefinitely because the information it gathers must eventually be erased from its finite memory to allow for more. But information erasure incurs a minimal entropic cost per bit (the Landauer limit), compensating for the entropy decrease during the demon's manipulations. However, as now conceived, Maxwell's demon illustrates a deep connection between information and thermodynamics: in effect, information itself becomes the energy source. That this conversion of information to energy is possible has now been demonstrated experimentally by the manipulation of microscopic objects.

Maxwell's demon offers a simple model for exploring Schrödinger's assertion (1944) that living organisms operate at the molecular scale to reduce their own entropy and sustain organization – to

"feed off negative entropy" – during their lifetime. The key point is that the movements of gas molecules, apparently random and uniform on average, become *meaningful* if we can collect information about individual particles. Such concepts underpin the understanding of "Brownian ratchets" in biology (Oster 2002), such as the way movement of cells is enabled by correlating the growth of actin filaments in the actomyosin cytoskeleton with fluctuations of the cell membrane against which it pushes.

For conceptualizing agency, Maxwell's demon illustrates several key factors. First, the demon has a goal – to create a temperature gradient – that determines its behaviour. Second, to achieve this goal it must *correlate* its behaviour with changes in its environment: specifically, it opens the trapdoor only when molecules with the right energies approach from the respective directions. The agent is not so much determining events as steering them. This process shows that agency demands a congruence of timescales as well as spatial scales: if, say, the responsiveness of the agent has a much slower rate than the rate of relevant environmental change, that change cannot be harnessed to achieve a goal. This can be seen as a problem of impedance-matching.

Third, the agent must be able to sense and record relevant information: the demon needs to measure the energies and trajectories of the particles, and to retain that information for at least as long as it takes to open the trapdoor and let the particles pass. Memory is thus essential. And the demon must be capable of distinguishing information relevant to its goal from that which is not. It does not matter, for the sake of creating a temperature gradient, whether the molecules in an air-filled compartment are oxygen or nitrogen. The demon could also do work by marking this distinction so as to create a gradient of chemical potential (in which case particle energies would then not matter). So the goal determines the *meaning* of information, distinguishing what is meaningful *for the agent* from what is not.

A correlation between the state of an organism and that of its environment implies that they *share information in common*. Kolchinsky and Wolpert (2021) say that it is this shared information that helps the organism stay out of equilibrium — because, like Maxwell's demon, it can then tailor its behaviour to extract work from fluctuations in its surroundings. If it did not acquire this information, the organism would gradually revert to equilibrium: it would die, buffeted mercilessly by random fate.

So living organisms can be regarded as entities that attune to (correlate with) their environment by using information to harvest energy and evade equilibrium. Life can then be considered as a computation that aims to optimize the acquisition, storage and use of such meaningful information. And life turns out to be extremely good at it. The best computers today dissipate many orders of magnitude more energy than Landauer's lower limit. But Wolpert estimates that the thermodynamic efficiency of the total computation done by a cell is only ten or so times greater than the Landauer limit. Biology is able to minimize the amount of computation an organism does.

This picture of agents adapting to a fluctuating environment allows us to deduce something about the way they store information. Still *et al.* (2012) show that, so long as such entities are compelled to use the available energy efficiently, they are likely to become "prediction machines": they must be able to anticipate incipient change in their surroundings so as to be able to work most efficiently. Such a device therefore needs to possess a *memory* of some kind, along with some capacity to retain information

about the past environment that can be used for predictive purposes. This information can be used for prediction when the agent contains an "implicit model" – a representation – of the environment. Still *et al.* show that the efficiency of such a proto-agent is contingent on its ability to distinguish information relevant to its efficient operation from that which is not: collecting information indiscriminately that is of no use to its goal decreases the efficiency.

How does such a proto-agent arise? One possibility is via conventional Darwinian evolution: these entities might, for example, develop sensory mechanisms for all manner of environmental signals, but those that sense *relevant* signals are more efficient and thus more fit. Such reliance on random variation is not the only possibility however. For example, Perunov *et al.* (2016) claim that self-organizing systems out of equilibrium have a thermodynamically driven tendency to "adapt" to their environment – in effect to develop correlations with the fluctuations in their surroundings. Such adapted entities, they say, are better at maximizing entropy production: at absorbing energy from the environment and dissipating it. They will generally be *selected* from all possible states the system could adopt. "When highly ordered, [dynamically] stable structures form far from equilibrium", they say, "it must be because they achieved reliably high levels of work absorption and dissipation during their process of formation".

In other words, systems that are complex, versatile and sensitive enough to respond to fluctuations in their environment can display a kind of evolutionary adaptation even if they are not self-replicating and do not undergo Darwinian evolution. There is no conflict between this physical process of adaptation and the Darwinian one; in fact, the latter can be seen as a particular case of the former. If these complex systems *can* replicate, we would expect certain states to emerge that are best adapted to taking in and dissipating energy, *by virtue of their own orderliness* – just as Schrödinger envisaged. In this view, Perunov *et al.* say, "the Darwinian account of adaptation and the thermodynamic one become one and the same."

Similarly, Egbert *et al.* (2016; 2022) describe how "ante-organisms" – dissipative entities not yet truly living – might actively regulate their environment to support their own persistence ("viability-based behaviour") due to simple feedback effects even before they begin to undergo Darwinian evolution. The researchers argue that such systems may have relatively facile access to more diverse forms with improved viability: they can harness variation without undergoing replication and mutation. The model challenges the prevailing idea that evolution per se was a precondition for the development of increasingly organism-like entities; the reverse could be true.

Such efficient energy-absorbing, highly dissipative states don't in themselves necessarily display agency. But they do appear to have a kind of thermodynamic quasi-goal – to maximize the use of energy in the environment – which is enough by itself to give the system some degree of structure. Morowitz and Smith (2007) argue that for this reason life (and thus agency) is highly likely to arise, purely on thermodynamic grounds, in any environment that has the necessary chemical ingredients along with concentrated reservoirs of energy.

Furthermore, Adam *et al.* (2018) argue that systems with many-tiered, hierarchical levels of structure are best suited for this process of converting and dissipating energy in the environment – especially high-energy inputs such as gamma rays that intensely disrupt just a few subatomic components of the

system – into far-from-equilibrium, dissipative organized structures and high levels of dynamical complexity. Such complexity is no guarantee of agency in itself, but it might help promote the degree of internal structure that seems to be a precondition for it.

Again, then, the question arises of to what extent agency can harness universal physical laws and to what extent it relies on bespoke strategies discovered by random variation and natural selection. There is reason to suppose that, while the agency observed in living organisms is qualitatively distinct from simple thermodynamic gradient-descent to the most stable steady state, it was not "invented de novo" during the onset of Darwinian evolution, but drew on pre-existing tendencies for adaptive, self-sustaining behaviour in complex ante-organisms, much as complex and ultimately self-aware cognitive mechanisms surely arose from pre-existing cell-cell interactions.

At what point such behaviours qualify as agency perhaps has no more definitive an answer than does the question of whether viruses (or when putative prebiotic protocells) qualify as life. Concepts like these do not need to have sharp boundaries to be useful. By the same token, *models* of agency can afford to be agnostic – to take what we might call an agential stance. They can incorporate agency or not, depending on whether this adds to their predictive or explanatory power. "Agent-based" models of road traffic and pedestrians (Helbing *et al.* 2001; Kerner 2004), for example, may need to assume nothing more than that vehicles behave as collision-avoiding active particles interacting mechanically via repulsive interactions, lacking in true agency. That does not deny the genuine agency of drivers or pedestrians, but recognizes that it might not be needed to model these situations adequately.

Back to Table of Contents

VI. Agency and Evolution

The Modern Synthesis explicitly omits any consideration of how organisms work as autonomous agents, and emphatically denies any teleology in evolution itself (a position that is arguably distinct from Darwin's own (Lennox 1993)[3]. If agency and purpose are to be readmitted to biology, the question is whether they can be reinstated as an addendum to conventional evolutionary theory, or whether their influence is more profound and disruptive.

When agency is excluded from evolutionary biology, organisms appear to be merely pushed around by random mutations and environmental influences. But while there is good reason to think that DNA mutations occur randomly (albeit at rates that are themselves subject to various levels of control within the organism; see Melamed *et al.* 2023), the question as yet unresolved is whether phenotypic variation follows suit. For example, genetic mutations might produce variations in the developmental patterns generated by self-organizing diffusional or reaction-diffusion processes of morphogens – but those variations are highly constrained by the intrinsic dynamics of these at a higher-level, which seems to have a distinct palette of possible outcomes (Newman 1992; 2019). The same may be true for cell states, for which the dynamical landscape of gene expression appears to be highly structured with attractor states (Saéz *et al.* 2022). Development thus seems both to buffer and to canalize random genotypic variation. This gives development a conservative tendency while also permitting evolutionary change to be potentially transformative: nothing changes, one might say, until everything does (Kirschner & Gerhart 2015).

Because there is as yet no true integration of evolutionary genetics with developmental morphology, we cannot say for sure how important these issues are for evolution. The distinction is between organisms as the *products* of evolution versus organisms as the *agents* of evolution. Some believe that this distinction calls for a substantially new theory – an "extended evolutionary synthesis" (Laland *et al.* 2014). Others argue that many of the features such a theory claims to provide are already incorporated into conventional modern evolutionary theory (Laland *et al.* 2015).

Not only has the agency of organisms been considered largely irrelevant to the way evolution has progressed, but even organisms themselves have been largely written out of the picture (Wagner 2015). In the Neodarwinian Modern Synthesis that regards evolution in terms of changes in population allele frequencies, organisms are typically portrayed as mere vehicles for genes. This has led organisms to be regarded as a "paradox" for evolution (Dawkins 1990): how do they exist at all when they seem to require collaboration rather than competition between genes? As Dawkins put it,

[3] Reviewing Darwin's legacy in *Nature* in 1874, the American botanist Asa Gray wrote "Let us recognize Darwin's great service to Natural Science in bringing back to it Teleology; so that, instead of having Morphology *versus* Teleology, we shall have Morphology wedded to Teleology." (Gray 1874)

"The paradox of the organism is that it is not torn apart by its conflicting [gene] replicators but stays together and works as a purposeful entity, apparently on behalf of all of them. Not only is it not torn apart; it functions as such a convincingly unified whole that biologists in general have not seen that there is a paradox at all!"

One response to this situation might be to wonder if an explanation for the evolution of Darwin's "endless forms most beautiful" that ends up excluding the explicandum has gone astray, so that the aim should be to go back and identify the mistake. The alternative was to redefine the problem: to make evolution and indeed life itself *all about* genes. In the course of so doing, what was lost with the organism (namely agency) was essentially relocated in the gene (Queller 2019). In conventional narratives of the Modern Synthesis, the gene itself becomes imbued with a kind of quasi-agency, having the goal of propagation and the guise of an active agent capable of replicating and competing with others.

It is often overlooked that the "selfish replicator" version of the Modern Synthesis, at least as presented by Dawkins (1976, 1982), does not in fact hinge on the alleged unique capacity of genes encoded in DNA to replicate. Dawkins (1982) defines a "replicator" in this context as "anything in the universe of which copies are made" – which must then presumably also include all other biomolecules, from proteins to lipids. This does not conform to any usual, or indeed obvious, definition of a replicator.

It might be argued that DNA is nonetheless a special kind of replicator because it replicates *itself* – except that in fact it never does so, except through the intervention of a genuine agent (the cell, or perhaps the technician performing the process of PCR by which DNA is amplified in the lab). Making copies of genes thus depends on agency, but it is not truly an agency possessed by the genes themselves.

Some have argued that, in any event, *evolvability* itself demands more than mere replication: it requires the existence of coherent and agential entities that Griesemer (2006) has called *reproducers*. Such entities might be regarded as the fundamental evolvable unit of all organisms, and we might crudely equate them with the cell itself. They have hierarchical organization that absorbs and adjusts to the unexpected. The reproducer perspective, says Jaeger (2021), offers "an organizational theory of evolution by natural selection, which has the organism (and its struggle for existence) back at its core, as it was in Darwin's original theory."

One might say that the "replicator" model of the Modern Synthesis in fact simply describes the mathematical models useful for evolutionary genetics: it is not "wrong" but serves a limited purpose. The harder question is whether genuine agency significantly alters the picture of how evolution happens by random mutation and natural selection.

Random mutation is after all not the only way in which genomic sequences may change and become fixed by natural selection. It is well attested that viruses can introduce new elements to genomes, for example – and it has been proposed (Prudhomme *et al.* 2005; Shapiro 2013) that such viral modification might have been important for the radiation of mammals since the divergence from marsupials. We also know that genomes actively rearrange themselves, in particular through gene duplications and the activity of transposons ("jumping genes"). It is less clear whether there is anything systematic, let alone goal-directed, in such transformations. It seems possible, for example, that

genomic variability contributes to evolutionary innovation. It has been argued that the widespread transcription of non-coding DNA (ENCODE 2012) might create a reservoir of molecular variation for potential regulatory RNA (and perhaps small peptide molecules) from which evolutionary innovations can come. Gene duplications, meanwhile, allow proteins to acquire new roles without compromising their existing ones.

Do organisms per se exercise top-down influence on any of this variation in sequence, in effect exercising agency over their own genomic evolution? Shapiro (2013) argues that there is evidence that the activity of transposons responds to changes in the environment. Virgo *et al.* (2023) explain how a selective pressure could arise that favours organisms capable of exerting control over their mechanisms of variation, since this can promote lineages with greater evolvability and thus greater fitness in the long term. Moreover, *epigenetic* changes in gene regulation, produced by environmental stresses, may be inherited – this seems well attested in plants (Henderson & Jacobsen 2007), although the evidence for it in animals is weaker, and it is even less clear that such effects are persistent enough across generations to be evolutionarily significant.

Whether or not organisms can influence their own genomes, there is a more general argument that agency impacts evolution. In at least one respect this is hardly news: the capabilities of organisms, which surely may include behavioural flexibility and innovation, of course affect their fitness. It is the agency of humans to construct shelter, clothing and technologies that has enabled us to survive in so many diverse habitats. Tool use in corvids is surely an adaptive agential skill (the role of tools in mediating the agent-environment interaction is a topic ripe for exploration). Some argue that in such ways – for example, in theories of niche construction – agency is already adequately incorporated into evolutionary theory as an adaptation like any other. As with other arguments about the role of agency in evolution, the disagreement here seems to stem from a divergence of opinion about where to locate *causation* (Uller & Helanterä 2019).

Some argue that the exclusion of agency from a Neodarwinian picture in which evolution is viewed as changes in gene frequencies hides the richness of the phenomena that the central organizing theory of biology ought to address. Walsh (2015) puts it starkly:

"The most glaring defect of the Modern Synthesis approach to inheritance is precisely that it makes no provision for the various ways in which organismal development, broadly construed, can contribute to the pattern of resemblance and difference that constitutes inheritance. Organisms, as purposive, adaptive agents, actively participate in the maintenance of this pattern. Their omission has left us with a distorted and devitalised conception of inheritance."

"Assimilating the agency of organisms into evolutionary thinking", Walsh adds, "renders a conception of evolution that, while wholly consistent with Darwinism, puts considerable strains on the Modern Synthesis account of evolution." He proposes a revised model that he calls Situated Darwinism. He claims that "the spontaneous order of the biological world is held in place by the purposiveness of organisms", and that in consequence, "if our scientific methodology fails to countenance purpose, then it renders us blind to a perfectly real, evolutionarily important class of empirical regularities".

If agency does affect the trajectory of evolution, might it generate directionality to the overall process – giving evolution itself some goal or target?[4] Can the agency of organisms, the objects of selection, produce at least the *appearance* of agency in evolution itself? The question is still considered heretical by some, but there need be nothing mystical about it – it is not a backdoor for intelligent design. In some sense it is uncontroversial that evolution possesses something resembling goals, for we see it in the well attested phenomenon of convergent evolution, where different evolutionary lineages independently find their way to the same solution: eyes, brains, wings. It is generally believed that this happens because those properties or structures are good "engineering" solutions to common problems: how to make good use of information conveyed by light, how to fly, and so on. We might then usefully regard it as another example of biological *attractors*: evolution is channelled into attractor states created by the environment along with the principles of physical law. By the same token, might the operation of agency conceivably create evolutionary attractors shaped by the internal nature of evolution itself? These are open questions.

Back to Table of Contents

[4] Agency is just one putative source of directionality in evolution; the latter need not rely on it.

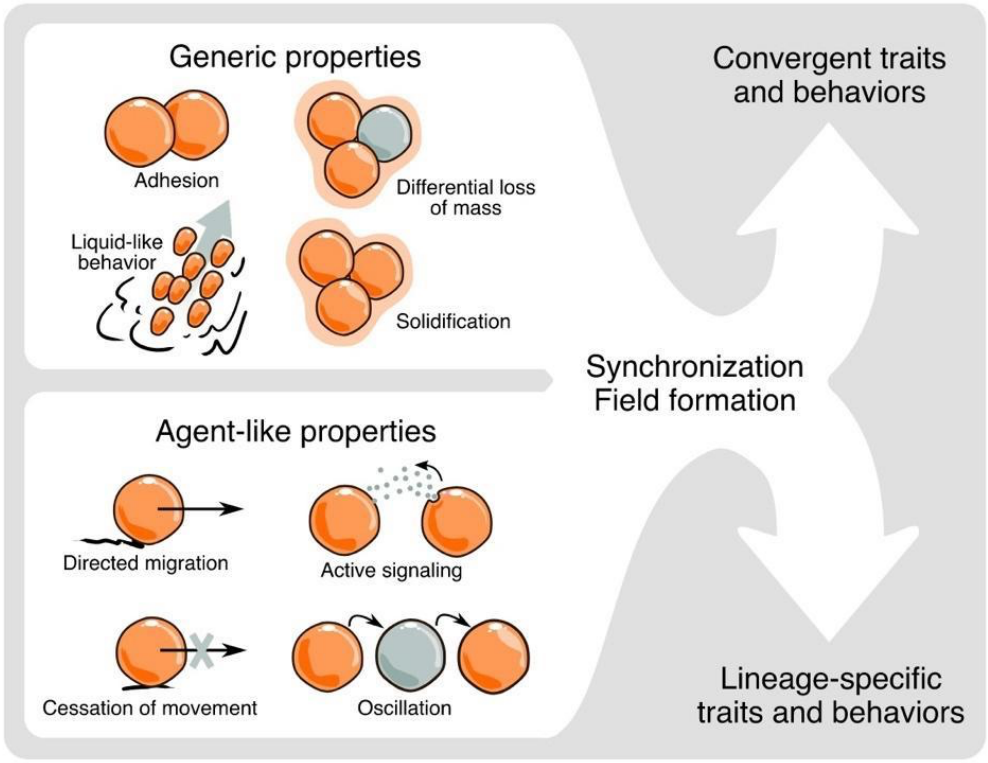
VII. Collective and Multicellular Agency

Although organisms do not have to become "paradoxes", nonetheless there are genuine questions about how the agency of individual cells in a multicellular organism (such as neutrophils) conforms to the needs of the organism as a whole. One way of regarding cancer, and perhaps also autoimmune conditions, is as a breakdown of this kind of collaborative agency.

Such conditions remind us that multicellular agency is in some respects precarious. The conventional view is that cellular autonomy is constrained in multicellular organisms via a complex series of checks and balances, including the ability of the immune system to learn to distinguish self from other, the innate tendency of genetically damaged cells to "commit suicide" (the process called apoptosis), multiple levels of cell-cycle regulation, the existence of tumour-suppressor genes, and the rapid passage of stem cells into canalized, differentiated states during development to minimize the risk of aberrant trajectories.

All this is instrumental to the attainment of the goal of organismal viability, but it does not really address the question of how that collective goal arises in the first place. It seems clear that multicellularity can be adaptive – it appeared several times in evolution (Parfrey & Lahr 2013) (although only thrice led to the complex multicellular organisms of animals, plants, and fungi), and can be engineered in single-celled eukaryotes by applying selective pressure (Ratcliff & Travisano 2014; Herron *et al.* 2019). But how do agential cells align their goals with those of the collective? Arnellos & Moreno (2015) suggest that unleashing complex organismal behaviour, particularly in the context of mobility and sensorimotor function, requires a degree of global oversight and coordination from a nervous system, rather than relying on spontaneous self-organization. In a sense, the nervous system thus creates a kind of unified self. Godfrey-Smith (2020) suggests that a global nervous system might have come about in evolution by the merging of two separate nerve networks that initially worked independently for internal coordination (particularly of motion) and external sensing.

Arias Del Angel *et al.* (2020) argue that the morphological and behavioural characteristics of "simple" multicellular aggregates are dictated more by the way generic physical properties couple to the agential nature of cells than by a selective alignment of their goals. They consider the tendency of both Myxobacteria (prokaryotes) and dictyostelid amoebae (eukaryotes) to form structures called fruiting bodies in response to nutrient depletion. In both cases the cells become developmentally distinct. The researchers say that the formation of streams and of rippling waves in these aggregates arises from the liquid-like properties that result from cell-cell adhesion, as well as a subsequent liquid-solid transition, coupled to agent-like behaviours such as directed cell migration (chemotaxis, say), changes of internal state (such as quiescence), and oscillatory internal dynamics triggered by external signals. They argue that morphological evolution, including the emergence of metazoans, must therefore take account of the higher-level implications of cellular agency as well as the genetic origins of body patterning in highly conserved systems such as *hox* genes. Here, then, an understanding of agency and its consequences might contribute to a well-posed biological question: what are the factors that influenced the evolution of form?



Generic physical effects and agent-like behaviours that contribute to multicellular development in aggregative forms. From Arias Del Angel et al. 2020.

A broader question is whether a kind of agency can emerge from the collective dynamics of entities that are themselves not truly agential. At some level, this must be so: the fact that agents (cells) can arise from the interactions of many entities (molecules) that lack it offers a kind of existence proof for the idea that purely artificial objects might, in aggregates of sufficient complexity, be capable of developing it. We can also ask whether collective agency can be qualitatively distinct from that of individual agents that comprise the collective: are companies agents in their own right, for example (Ludwig, 2016)? An understanding of "swarm intelligence" – the harnessing of information flows in groups and aggregates for efficient problem-solving – is now a rather mature and active field of research, but far less attention has been given to the concept of *swarm agency*. Such considerations prompt the question of whether agency might be engineered, and whether agents can be systematically directed towards goals that are not recognized by the agents themselves.

Back to Table of Contents

VIII. Agency, Engineering, and Ourselves

That living cells have individual and collective goals might be considered evident in the structures they can form ex vivo. Organoids made from stem cells (either embryonic or induced by reprogramming somatic cells) (Kim *et al.* 2020) recapitulate the developmental and morphological programs they follow in embryogenesis, albeit modified and adapted to their in vitro circumstances – in particular, because of the lack of signals from other tissues, the organoid structures are generally just approximations of the corresponding structures in the body. These stem-cell-derived artificial structures now include embryo-like bodies, which can be cultured even through gastrulation (the point at which key body axes and structures start to appear) to organogenesis and function: synthetic embryoids have been made with beating primitive hearts (Amadei *et al.* 2022; Tarazi *et al.* 2022).

Tissues grown in this way in vitro can be regarded as *agential materials* (Davies & Levin 2023), in which the building blocks have "plans of their own" – a capability not just to replicate but also to differentiate and self-organize. Engineering with such materials – a key component of the discipline called *synthetic morphology* (Davies 2008) – thus becomes a kind of collaboration between designer and material, in which it becomes imperative to understand the goals of the agents and how these might be best guided towards a desired end. In effect the aim is to align the agency of the engineer with that of the substrate.

There is also an increasing desire to produce wholly artificial materials with "animate" properties such as self-repair, environmental energy harvesting, self-organizing complexity, and adaptiveness (Ball 2021). Although no artificial system to date can lay persuasive claim to possessing agency (Moreno 2018), a better understanding of what that characteristic entails should allow agent-like properties increasingly to be designed into them. Davies and Levin (2023) suggest that such materials will display different degrees of "persuadability" – "the level of micromanagement and the expected degree of sophistication through autonomous behaviour and problem-solving needed to obtain a specific result" – depending on the amount of computational and cognitive complexity they contain.

Artificial agential materials seem likely to first arise in a semi-synthetic manner, incorporating some biological components – perhaps DNA strands that encode information used for self-assembly, read out using the natural molecular "machinery" for transcription and translation. By the same token, biological cells will themselves be increasingly designed and engineered to exhibit non-natural properties – as, for example, in the creation of molecular systems for keeping records of cell-state histories (Farzadfard & Lu, 2018), or of synthetic cell-adhesion molecules for constructing novel tissues (Stevens *et al.* 2023). It is far from clear that even wholly natural cells of complex multicellular organisms display the full gamut of stable cell states accessible to them: reprogramming techniques might release agential potential that evolution has not yet revealed. Indeed, the production of artificial living systems called xenobots from cells freed from normal developmental constraints (Blackiston *et al.* 2021) suggests that the morphologies displayed in nature are not unique outcomes of the rules governing collective cell behaviour.

Our own agency is typically experienced in terms of conscious choices: we explicitly state goals (whether or not these are realistic, or are the real motivators of our behaviour) and formulate explicit plans to achieve them. As we saw, some have made such capabilities the hallmark of true agency, although that now seems too restrictive. In one view, agency at least demands the possibility of the agent's *acting other than it did*. Such counterfactuals then impinge on the contested territory of free will.

The problem of free will has metaphysical roots in issues of determinism. Some argue that if determinism holds, there can be no true causation at all, and so the very notion of agency – of agents as causes of things – evaporates. (Determinism is at some level irreconcilable with the current picture of fundamental randomness in quantum events – but some degree of quantum indeterminism does not obviously speak to questions of agency since it is inaccessible to manipulation.) Others argue that determinism merely asserts cause-and-effect relationships behind events – and agency requires only that agents themselves be bona fide causal entities in their own right (Dennett 1984). In that view, a fully compatibilist account of agency and free will in a deterministic world seems possible.

That debate remains unresolved after more than two millennia of discussion. Perhaps a more scientifically tractable perspective on free will centres on the nature and origins of volitional behaviour as a neurobiological phenomenon (Brembs 2020; Hills 2019). In this perspective, free will manifests in degrees of deliberative volition – like consciousness, it need not be seen as something special and discontinuous to humans, but builds on more general (neuro)biological capacities evident at least throughout the varieties of metazoans and perhaps more widely in nature. We might, for example, consider what we call free will in humans to arise from an ability to create complex mental representations of imagined futures, and to consciously select actions deemed likely to attain those considered desirable. This ability of "mental time travel" (Suddendorf et al. 2009) no longer seems a uniquely human attribute (Emery & Clayton 2004). In this view, agency becomes a precondition of "free will", and the latter is merely a sophisticated form of it (Mitchell 2023).

Back to Table of Contents

Conclusions

There is a strong case for considering agency to be a real property of most if not all living organisms, and perhaps to be a defining feature of life on a par with, for example, metabolism and self-replication. Yet if this is so, it remains unclear how it is to be defined and identified, and to what extent it can be operationalized so that researchers can formulate and answer questions about living systems that cannot be addressed with a purely phenomenological and mechanistic approach. It seems entirely possible that attempts to construct a unique definition and theory of agency will be counterproductive: like cognition and consciousness, it is probably not some single essence that living things contain in different amounts.

Nonetheless, there are many promising directions for studying minimal models of agency. Central to these efforts is the goal of developing a richer understanding of how organisms interact with their environment: how they function as *situated* entities. Such understanding seems likely to have implications for evolutionary theory – for, as Levins and Lewontin (1985) pointed out,

"Natural selection is not a consequence of how well the organism solves a set of fixed problems posed by the environment [as it typically appears in the Modern Synthesis]; on the contrary, the environment and the organisms actively co-determine each other."

At root, a focus on agency in biology would restore the organism at the heart of the life sciences as a self-sustaining, autonomous and goal-directed entity. Recognizing these features, however, demands an acceptance that goals and purpose play roles in biology: a perspective that can justifiably be called teleological. Properly construed, these concepts need give the biologist nothing to fear; there is a good reason to think that they can, as it were, be tamed and instrumentalized.

All the same, there might be limits to a strict calculus of agency and purpose. We might look for the origins of goal-directed behaviour in thermodynamics or information theory, for example, but there is no reason to suppose that, in a highly hierarchical system like a living organism, goals flow from a single source or have a single nature – and the highest-level goals might not be ones that can be expressed in differential equations. As Steane (2018) has put it:

"Perhaps the purpose of a lioness is to dissipate entropy as quickly as possible (I doubt it). Perhaps the purpose of a lioness is to produce more lioness genes (I doubt it). Perhaps the purpose of a lioness is to kill, eat, and copulate (I doubt that, too). Perhaps the purpose of a lioness is to be a lioness (this seems to me to be on the right track)."

There are likely to be benefits in this enterprise for developmental biology, bioengineering and biomedicine, microbiology, robotics and AI, and evolutionary biology. Understanding agency is a highly interdisciplinary endeavour, which will need input from (inter alia) developmental and cell biology, genetics, ethology, biophysics, complex-systems science, non-equilibrium thermodynamics, computational science, as well as the philosophy of biology. The somewhat ad hoc nature of efforts so far has been productive but not always coherent or integrated (Love & Dresow 2022). This is not

necessarily a bad thing: in the early days of a field, plurality can be preferrable to too hasty and narrow a consensus. But it is already possible to discern some common themes and questions that bode well for a productive confluence of ideas in the years ahead.

Back to Table of Contents

References

Z. R. Adam, A. C. Fahrenbach, B. Kacar & M. Aono (2018). Prebiotic geochemical automata at the intersection of radiolytic chemistry, physical complexity, and systems biology. *Complexity* **2018**, 9376183.

M. Aguilera, B. Millidge, A. Tschantz & C. L. Buckley (2022). How particular is the physics of the free energy principle? *Physics of Life Reviews* **40**, 24-50.

C. Allen, M. Bekoff & G. Lauder (eds) (1998). *Nature's Purposes: Analyses of Function and Design in Biology*. Cambridge, Ma.: MIT Press.

G. Amadei *et al.* (022). Synthetic embryos complete gastrulation to neurulation and organogenesis. *Nature* **610**, 143-153.

J. A. Arias Del Angel, V. Nanjundiah, M. Benítez & S. A. Newman (2020). Interplay of mesoscale physics and agent-like behaviors in the parallel evolution of aggregative multicellularity. *EvoDevo* **11**, 21.

A. Arnellos & A. Moreno (2015). Multicellular agency: an organizational view. *Biology & Philosophy* **30**, 333-357.

H. Atlan (1998). Intentional self-organization. Emergence and reduction: towards a physical theory of intentionality. *Thesis Eleven* **52**, 5–34.

P. Ball (2021). Animate materials. *MRS Bulletin* **46**, 553-559.

X. Barandiaran & A. Moreno (2006). On what makes certain dynamical systems cognitive. *Adaptive Behavior* **14**, 171–185.

X. E. Barandiaran, E. Di Paolo & M. Rohde (2009). Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behaviour* **17**, 367-386.

X. E. Barandiaran & M. D. Egbert (2014). Norm-establishing and norm-following in autonomous agency. *Artificial Life* **20**, 5-28.

R. D. Beer (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* **72**, 173-215.

E. Ben-Jacob, M. Lu, D. Schultz & J. N. Onuchic (2014). The physics of bacterial decision making. *Frontiers in Cellular and Infection Miocrobiology* **4**, 154.

C. Bennett (1982). The thermodynamics of computation—a review. *International Journal of Theoretical Physics* **21**, 905–940.

L. von Bertalanffy (1969). *General Systems Theory*. New York: George Braziller.

M. Biehl & N. Virgo (2022). Interpreting systems as solving POMDPs: a step towards a formal understanding of agency. Preprint http://www.arxiv.org/abs/2209.01619.

D. Blackiston, E. Lederer, S. Kriegman, S. Garnier, J. Bongard & M. Levin (2021). A cellular platform for the development of synthetic living machines. *Science Robotics* **6**, abf1571.

D. Bray (2009). *Wetware: A Computer in Every Living Cell*. New Haven: Yale University Press.

B. Brembs (2020). Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in vertebrates. *Proceedings of the Royal Society B* **278**, 930-939.

C. L. Buckley, C. S. Kim, S. McGregor & A. K. Seth (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology* **81**, 55-79.

M. Colombo & P. Palacios (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology & Philosophy* **36**, 41.

P. Corning et al. (eds) (2023). *Evolution "On Purpose": Teleonomy in Living Systems*. Cambridge, Ma.: MIT Press.

J. Davies (2008). Synthetic morphology: prospects for engineered, self-constructing anatomies. *Journal of Anatomy* **212**, 707-719.

J. Davies & M. Levin (2023). Synthetic morphology with agential materials. *Nature Reviews Bioengineering* **1**, 46-59.

R. Dawkins (1976). *The Selfish Gene.* Oxford: Oxford University Press.

R. Dawkins (1982). *The Extended Phenotype: The Gene as the Unit of Selection.* Oxford: Oxford University Press.

R. Dawkins (1990). Parasites, desiderata lists and the paradox of the organism. *Parasitology* **100**, S63-S73.

D. C. Dennett (1984). *Elbow Room*. Oxford: Oxford University Press.

D. C. Dennett (1987). *The Intentional Stance*. Cambridge, Ma.: MIT Press.

E. A. Di Paolo (2005). Autopoesis, adaptivity, teleology, agency. *Phenomenology & the Cognitive Sciences* **4**, 429-452.

E. A. Di Paolo (2009). Extended life. *Topoi* **28**, 9–21.

M. Dresow & A. C. Love (2023). Teleonomy: Revisiting a proposed conceptual replacement for teleology. *Biological Theory* https://doi.org/10.1007/s13752-022-00424-y.

J. G. Dyke & I. S.Weaver (2013). The emergence of environmental homeostasis in complex ecosystems. *PLoS Computational Biology* **9**, e1003050.

M. D. Egbert & J. Pérez-Mercader. Adapting to adaptations: behavioural strategies that are robust to mutations and other organisational transformations. *Scientific Reports* **6**, 18963,

M. Egbert *et al*. (2022). Behavior and the origin of organisms. Preprint [details t/k].

N. J. Emery & N. S. Clayton (2004). The mentality of crows: convergent evolution of intelligence in corvids and apes. *Science* **306**, 1903-1907.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74.

J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev & A. F. Schier (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131.

F. Farzadfard & T. K. Lu (2018). Emerging applications for DNA writers and molecular recorders. *Science* **361**, 870-875.

K. Friston (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* **11**, 127-138.

K. Friston (2013). Life as we know it. *Journal of the Royal Society Interface* **10**, 20130473.

M. Gagliano (2017). The minds of plants: thinking the unthinkable. *Communications in Integrated Biology* **10**, e1288333.

J. J. Gibson (1979). *The Theory of Affordances: The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

P. Godfrey-Smith (2020). *Metazoa: Animal Minds and the Birth of Consciousness.* London: William Collins.

A. Gray (1874). Scientific worthies III: Charles Robert Darwin. *Nature* **10**, 79-81.

J. Griesemer (2006). Genetics from an evolutionary process perspective. In E. M. Neumann & C. Rehmann-Sutter (eds), Genes in Development. Durham, NC: Duke University Press.

S. Guttinger & A. C. Love (2023). modENCODE and the elaboration of functional genomic methodology. Preprint.

D. Helbing, P. Molnár, I. J. Farkas & K. Bolay (2001). Self-organizing pedestrian movement. *Environment and Planning B: Planning and Design* **28**, 361–383.

I. R. Henderson & S. E. Jacobsen (2007). Epigenetic inheritance in plants. *Nature* **447**, 418-424.

M. D. Herron *et al.* (2019). *De novo* origins of multicellularity in response to predation. *Scientific Reports* **9**, 2328.

T. T. Hills (2019). Neurocognitive free will. *Proceedings of the Royal Society B* **286**, 20190510.

J. Jaeger (2021). The fourth perspective: evolution and organismal agency. Preprint, 26 February. https://osf.io/2g7fh

G. F. Joyce & J. W. Szostak (2018). Protocells and RNA self-replication. *Cold Spring Harbor Perspectives in Biology* **10**, a034801.

S. Kauffman (2000). *Investigations*. Oxford: Oxford University Press.

B. Kerner (2004). *The Physics of Traffic.* Berlin: Springer.

J. Kim, B.-K. Koo & J. A. Knoblich (2020). Human organoids: model systems for human biology and medicine. *Nature Reviews Molecular Cell Biology* **21**, 571-584.

M. Kirschner, J. Gerhart & T. Mitchison (2000). Molecular "vitalism". *Cell* **100**, 79-88.

M. Kirschner & J. C. Gerhart (2005). *The Plausibility of Life: Resolving Darwin's Dilemma*. New Haven & London: Yale University Press.

J. Kiverstein, M. D. Kirchhoff & T. Froese (2022). The problem of meaning: the free energy principle and artificial agency. *Frontiers in Neurorobotics* **16**, 844773 (2022).

A. Kolchinsky & D. Wolpert (2021). Work, entropy production, and thermodynamics of information under protocol constraints. *Physical Review X* **11**, 041024.

K. N. Laland *et al.* (2015). The extended evolutionary synthesis: its structure, assumptions and predictions. *Proceedings of the Royal Society B* **282**, 20151019.

K. N. Laland *et al.* (2014). Does evolutionary theory need a rethink? *Nature* **514**, 161–164.

J. G. Lennox (1993). Darwin was a teleologist. *Biology and Philosophy* **8**, 409-421.

M. Levin, A. M. Pietak & J. Bischof (2019). Planarian regeneration as a model of anatomical homeostasis: recent progress in biophysical and computational approaches. *Seminars in Cell. and Developmental Biology* **87**, 125-144

R. Levins & R. Lewontin, *The Dialectical Biologist*. Cambridge, Ma.: Harvard University Press.

R. C. Lewontin (1974). *The Genetic Basis of Evolutionary Change*. New York: Columbia University Press.

Y. Louzoun & H. Atlan (2007). The emergence of goals in a self-organizing network: a non-mentalist model of intentional actions. *Neural Networks* **20**, 156–171.

A. C. Love & M. Dresow (2022). Organizing interdisciplinary research on purpose. *BioScience* **72**, 321–323.

K. Ludwig (2016). *From Individual to Plural Agency*. Oxford: Oxford University Press.

H. R. Maturana & F. J. Varela (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: Reidel.

E. Mayr (1997). The objects of selection. *Proceedings of the National Academy of Sciences USA* **91**: 2091-2094.

E. Mayr (2004). *What Makes Biology Unique?* Cambridge: Cambridge University Press.

H. Meinhardt (1982). *Models of Biological Pattern Formation*. New York: Academic Press.

D. Melamed *et al.* (2023). *De novo* mutation rates at the single-mutation resolution in a human *HBB* gene-region associated with adaptation and genetic disease. *Genome Research* 10.1101/gr.276103.121.

R. Menzel (2019). The waggle dance as an intended flight: a cognitive perspective. *Insects* **10**, 424.

K. Mitchell (2023). *Free Agents*. Princeton: Princeton University Press.

A. Moreno (2018). On minimal autonomous agency: natural and artificial. *Complex Systems* **27**, 289.

A. Moreno & M. Mossio (2015). *Biological Autonomy*. Dordrecht: Springer.

H. Morowitz & E. Smith (2007). Energy flow and the organization of life. *Complexity* **13**, 51-59.

M. Mossio & A. Moreno (2010). Organisational closure in biological systems. *History and Philosophy of the Life Sciences* **32**, 269-288.

A. P. Muñuzuri & J. Pérez-Mercader (2022). Unified representation of Life's basic properties by a 3-species stochastic cubic autocatalytic reaction-diffusion system of equations. *Physics of Life Reviews* **41**, 64-83.

E. Nagel (1979). *Teleology Revisited, and Other Essays in the Philosophy and History of Science.* New York: Columbia University Press.

M. P. Neilson, D. M. Veltman, P. J. M. van Haastert, S. D. Webb, J. A. Mackenzie & R. H. Insall (2011). Chemotaxis: a feedback-based computational model robustly predicts multiple aspects of real cell behaviour. *PLoS Biology* **9**, e1000618.

S. A. Newman (1992). Generic physical mechanisms of morphogenesis and pattern formation as determinants in the evolution of multicellular organization. *Journal of Bioscience* **17**, 193-215.

S. A. Newman (2019). Inherency of form and function in animal development and evolution. *Frontiers in Physiology* **10**, 702.

G. Nicolis & I Prigogine (1977). *Self-organization in Nonequilibrium Systems: From Dissipative Structures to Order Through Fluctuations*. London: Wiley-Blackwell.

S. Okasha (2018). *Agents and Goals in Evolution*. Oxford: Oxford University Press.

G. Oster (2002). Brownian ratchets: Darwin's motors. *Nature* **417**, 25.

L. W. Parfrey & D. J. Lahr (2013). Multicellularity arose several times in the evolution of eukaryotes. *BioEssays* **35**, 339-347.

N. Perunov, R. Marsland & J. England (2016). Statistical physics of adaptation. *Physical Review X* **6**, 021036.

M. Pigliucci & Muller (eds) (2010). *The Extended Synthesis*. Cambridge, Ma.: MIT Press.

C. S. Pittendrigh (1958). Adaption, natural selection, and behavior. In A. Roe & G. G. Simpson (eds), *Behavior and Evolution.* New Haven: Yale University Press.

H. D. Potter & K. J. Mitchell (2023). Naturalising agent causation. [Preprint, details t/k]

S. Prudhomme, B. Bonnaud & F. Mallet (2005). Endogenous retroviruses and animal reproduction. *Cytogenetic and Genome Research* **110**, 353–364.

D. C. Queller (2019). The gene's eye view, the Gouldian knot, Fisherian swords and the causes of selection. *Philosophical Transactions of the Royal Society B* **375**, 20190354.

W. C. Ratcliff & M. Travisano (2014). Experimental evolution of multicellular complexity in *Saccharomyces cerevisiae*. *BioScience* **64**, 383-393.

P. W. Reddien & A. Sánchez Alvarado (2004). Fundamentals of planarian regeneration. *Annual Reviews of Cell and Developmental Biology* **20**, 725-757.

A. Rex (2017). Maxwell's demon – a historical review. *Entropy* **19**, 240.

J. Riskin (2016). *The Restless Clock: A History of the Centuries-Long Argument over What Makes Living Things Tick*. Chicago: University of Chicago Press.

M. Sáez, J. Briscoe & D. A. Rand (2022). Dynamical landscapes of cell fate decisions. *Journal of the Royal Society Interface* **12**, 20220002.

S. Samanta, R. Layek, S. Kar, M. K. Raj, S. Mukhopadhyay & S. Chakraborty (2017). Predicting *Escherichia coli*'s chemotactic drift under exponential gradient. *Physical Review E* **96**, 032409.

E. Schrödinger (1944). *What Is Life?* Cambridge: Cambridge University Press.

J. A. Shapiro (2013). How life changes itself: The Read-Write (RW) genome. *Physics of Life Reviews* **10**, 287-323 (2013).

C. Sonnenschein & A. M. Soto (2020). Over a century of cancer research: Inconvenient truths and promising leads. *PLoS Biology* **18**, e3000670.

A. Steane (2018). *Science and Humanity: A Humane Philosophy of Science and Religion*. Oxford: Oxford University Press.

A. J. Stevens *et al.* (2023). Programming multicellular assembly with synthetic cell adhesion molecules. *Nature* **614**, 144-152.

S. Still, D. A. Sivak, A. J. Bell & G. E. Crooks (2012). Thermodynamics of prediction. *Physical Review Letters* **109**, 120604.

T. Suddendorf, D. R. Addis & M. C. Corballis (2009). Mental time travel and the shaping of the human mind. *Philosophical Transactions of the Royal Society London B* **364**, 1317-1324.

S. E. Sultan, A. P. Moczek & D. Walsh (2021). Bridging the explanatory gaps: What can we learn from a biological agency perspective? *BioEssays* **2021**, 2100185.

S. Tarazi *et al.* (2022). Post-gastrulation synthetic embryos generated ex utero from mouse naïve ESCs. *Cell* **185**, 3290-3306.

M. Tomasello (2022). *The Evolution of Agency: Behavioral Organization from Lizards to Humans*. Cambridge, Ma.: MIT Press.

A. M. Turing (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society* **237**, 37-72.

T. Uller & H. Helanterä (2019). Niche construction and conceptual change in evolutionary biology. *British Journal for the Philosophy of Science* **70**, 351-375.

T. Veloz (2021). Goals as emergent autopoietic processes. *Frontiers in Bioengineering and Biotechnology* **9**, 720652.

N. Virgo, E. Agmon & C. Fernando (2023). Lineage selection leads to evolvability at large population sizes. Preprint.

K. von Frisch (1953). *The Dancing Bees: An Account of the Life and Senses of the Honey Bee*. New York: Harvest Books.

C. H. Waddington (1957). *The Strategy of the Genes*. London: Allen &Unwin.

G. P. Wagner (2015). Reinventing the organism: evolvability and homology in post-Dahlem evolutionary biology. In A. C. Love (ed.), *Conceptual Change in Biology*, *Boston Studies in the Philosophy and History of Science* **307**, DOI 10.1007/978-94-017-9412-1_15. Dordrecht: Springer Science+Business Media.

D. M. Walsh (2015). *Organisms, Agency, and Evolution*. Cambridge: Cambridge University Press.

L. Wolpert (2010). Arms and the man: the problem of symmetric growth. *PLoS Biology* **8**, e1000477.

Back to Table of Contents