The Deliberation Model of Organismic Agency

Hugh Desmond

Do not cite -- Draft of July 7, 2023

Organismic agency is often understood as the capacity to produce goal-directed behavior. This paper proposes a new way of modelling agency, namely as a naturalized *deliberation*. Deliberative action is not directed towards a particular goal, but involves a process of weighing multiple goals and a choice for a particular combination of these. The underlying causal model is symmetry breaking, where the organism breaks symmetries present in the selective environment. Deliberation is illustrated by means of bacterial chemotaxis.

Keywords: Agency – Organism – Mechanism – Deliberation – Intentionality

1. Introduction

The concept of agency has been emerging as one of the main contenders to the machine metaphor of organisms. The claim is that organisms cannot always be adequately conceptualized as a complex system of interacting functional mechanisms. Instead, under certain empirical and/or explanatory conditions, organisms must be conceptualized as *agents* (Arnellos & Moreno, 2015; Desmond & Huneman, 2020; Fulda, 2017; Gambarotto & Nahas, 2023; Liljeholm, 2021; Nadolski & Moczek, 2023; Paolo, 2005; Sultan et al., 2022; Tomasello, 2022; Walsh, 2015).

What is agency though? In areas of philosophy of action, ethics, or politics, where the agency concept is least controversial, an agent is a person who acts, and is not merely acted upon. In some contexts this is construed in terms of having intentional representations of future states of affairs; in others as being relatively free of social sources of oppression. Given that humans are the *paradigmatic* agents, the inevitable implication is that organismic agency is *in some sense analogous* to human agency. The question then becomes how this analogy should be understood: as a loose metaphor, or as a heuristic – or perhaps as referring to a grounding explanatory or causal structure?

As flawed as the machine metaphor may be (Nicholson, 2019), it has seen off many contenders over the centuries, from monads to élan vital. Will the agency concept fare differently? This is still uncertain. Yet, certain areas of scientific practice do suggest that the agency concept is fulfilling a real scientific function, most clearly concerning mammal and bird behavior, it where agential language continues to be routinely used, ranging from attributing a "sense of justice" to apes (Brosnan, 2023) to a "sense of beauty" to birds (Prum, 2017).

There are a number of challenges facing the agency concept, though these vary somewhat according to how precisely agency is understood. Some have understood agency in the narrow sense as some mental capacity, such as the capacity for intentional representation (Allen & Bekoff, 1997; Sterelny, 2000) or utilitarian calculation (Okasha, 2018). Then challenge becomes to specify the empirical conditions under which this capacity can be ascribed to non-human species, as well as to guard against the category mistake of attributing mental terms to causal processes (Ryle, 1949/2009). Others have adopted a broad understanding of agency, as the capacity of an entity to cause its own behavior (Desmond & Huneman, 2020; Moreno & Mossio, 2015; Walsh, 2015). On this understanding, agency is a domain-general causal-explanatory concept, applicable to bacteria, humans, or artificial intelligences. However, when goes down this path, the challenge that arises is to make sense of what precisely it means for an organism to cause its *own* behavior. Self-causation is a mysterious concept.

This paper starts from a broad understanding of agency, and aims to clarify what self-causation means in an intuitive, simple way. Ironically, this will be done via a (different) narrow understanding of agency: by reanalyzing a specific human mental function (deliberation) in very domain-general terms, and showing how this identifies a particular type of causal structure (symmetry breaking) that is a good candidate for clarifying what self-causation means.

Why the challenge of clarifying self-causation has not yet been met is because of the limiting assumption that agency must be a teleological notion (or so I will argue). On what I call the *teleological approach*, an organism is not to be viewed as a complex machine because it is capable of directing its behavior towards beneficial goals. The teleological approach may have important virtues: it can be formalized through the mathematical apparatus of attractor dynamics, suggests an intuitive analogy between organismic agency and human intentionality, and allows a dividing line to be drawn between agents and non-agents.

Nonetheless, the teleological approach is often taken as specifying what agency *means*. To call an organism an "agent" simply *means* that the organism has some capacity for producing goal-directed behavior. This is a limiting assumption, because it does not by itself clarify the distinction between *agential* and *functional* behavior. Agential behavior is caused by the whole organism; functional behavior is caused by a part of the organism – usually a part that has been shaped by natural selection. So if the bacteria swims up nutrient gradients in an apparently goal-directed way, this in and of itself is not sufficient ground for adopting an agential perspective, because chemotaxis can in principle be accounted for as a mechanism that has been selected for in virtue of some ecological function. And in such a case, it is not really the organism that is "causing" its own behavior, any more than a robot lawnmower causes its own behavior – rather, the *vera causa* here is the external designing principle, natural selection. Clarifying what self-causation means, must entail clarifying how the organism itself – and not some part of the organism (like a gene complex, or developmental program) or some external causal principle (especially natural selection) – should be considered as the cause of its own behavior.

This paper proposes that *deliberation* provides the right model for organismic agency. To give a brief preview: deliberation is a form of decision-making process whereby multiple courses of action are weighed, according to certain principles or goals, followed by a decision for a particular course of action. Genuine deliberation is not directed towards a single goal; moreover, despite the general principles involved, it is a very individual and idiosyncratic process, since the token-level features of the situation must be taken into consideration in the decision.

This model, translated to the realm of organismic behavior, refers to situations where natural selection no longer can "pre-decide" what the organism must do when confronted with a particular type of stimulus; instead, the organism must decide itself what to do. Agency is situated at the level of idiosyncratic individual action: this is why agential phenomena can appear as "noise" from the perspective of type-level generalizations. An organism faces uncertainty, since the evolutionary goals it has inherited underdetermine its course of action, and the organism itself acts to break the indecision. If the physicalist analogue of goal-directedness is attractor dynamics, deliberation corresponds to *symmetry breaking,* where the individual itself acts to break the symmetry between courses of action.

In the following two sections I introduce the teleological approach, and argue why it can only be saved with sophistications and epicycles that undermine its intuitiveness. Then, in section 4 I revisit different models of human agency, introducing the distinction between intentionality, autonomy, and deliberation, and in section 5 I formalize the deliberation model in terms of symmetry breaking. Section 6 applies the deliberation model to a chemotaxis in *E. coli*: a well-studied and very basic interaction between organism and environment.

2. The Teleological Approach to Agency

The teleological approach can be stated simply: organismic agency is the capacity of generating goal-directed behavior. The approach does not in itself represent an empirically verifiable definition of agency, because it leaves the core term of “capacity” and “goal” undefined. It does not denote any specific dividing line between agents and non-agents. Rather, the teleological approach to agency denotes a *logic* or style of thinking about agency. It is a way of defining agency with “goal” as a dominant definiens. As such, the teleological approach cannot be falsified by counterexamples (e.g., by pointing to the goal-directedness of thermostats): counterexamples can be taken as providing motivation for more sophisticated teleological definitions of agency.

The roots of the teleological approach trace back at least to the founding work in cybernetics. In one of the founding papers of cybernetics (Rosenblueth et al., 1943), “teleological behavior” was defined as behavior directed towards a goal through a process of negative feedback or “control”, where signals about proximity to the goal are used to modulate behavior. Hence the term cybernetics” was formed as an anglicization of χυβερνήτης (kubernḗtēs) – “steersman” (Wiener, 1948/2019, p. 18). Cybernetics represents organisms as agents insofar they actively steer their activities towards certain goals.

Cybernetics was but one strand in a collective effort to further extend the explanatory reach of statistical physics to include biological phenomena, and to find *laws* that could describe organisms and evolution. Recall that the 1940s also saw Schrödinger proposing the extremal principle that life feeds on “negative entropy”

(Schrödinger, 1944/1992), Prigogine investigating the thermodynamics of dissipative structures as a model for the origins of life (Prigogine, 1947), and Shannon introducing the information-theoretic formulation of entropy (Shannon, 1948). All of these developments were integrated in cybnernetics, which allowed a relatively precise definition of what it mean for a system to "control" its behavior.

In cybernetics, control is directed towards goals – and these goals were formalized through attractor dynamics. The *telos* for a cybernetical is thus, more precisely, an attractor state.[1] Attractor dynamics describes a dynamics that exhibits path-independence as long as the system's state remains in the basin of attraction. Regardless of its starting point, the system's state tends to evolve towards the attractor state, and even if perturbed along the way, the trajectory but not its end-point will be modified. Negative feedback – the "steering" in cybernetics – is but one particular way in which attractor dynamics can be implemented, as it ensures an interaction between an system's state and the goal-state so that the system's state does not overshoot. The result is that, if an system's behavior is beholden to attractor dynamics, then mechanistic explanation becomes a rather poor explanation of how a system behaves, and needs to to invoke the end-state – the *telos* – as an explanans.

The basic ideas underlying cybernetics still shape the contours of today's literature on organismic agency. For instance, in Michael Tomasello's recent account of the evolution of agency, he defines agency as the capacity not just for goal-directed behavior, but also for being able to *control* behavior.

> an agent does not just "aim and shoot" at its goals ballistically but rather flexibly controls (or even executively self-regulates) its actions by making informed decisions about what will work best at various points in a dynamically unfolding situation. (Tomasello 2022, p. 11)

Alternatively, in Daniel McShea's analysis of teleology, a distinction is made between two metrics of *persistence* and *plasticity* (Lee & McShea, 2020). Persistence refers the tendency of an object or entity to converge on an end-state given a variety of perturbations to its trajectory. Plasticity refers to the tendency to converge on an end-

[1] For purposes here I will identify the teleological approach with an attractor-based approach, but in general teleology does not imply an attractor dynamics. For instance, when Aristotle spoke of *eudaemonia* as the *telos* of humans he did not have in mind that all human lives automatically converge on eudaemonia.

state given a variety of starting positions. Both metrics describe different basic properties of attractor dynamics.

Denis Walsh's influential work on organismic agency can also be considered as teleological in the cybernetic sense of the term, as "the capacity to pursue a goal-state and sustain that state despite perturbations" (Walsh, 2015, p. 195), where goal-states are stable end-states (i.e., attractors). Or more fully:

> "[Agency] consists in a capacity of the system to pursue goals, to respond to the conditions of its environment and its internal constitution in ways that promote the attainment, and maintenance of its goal states." (Walsh, 2015, p. 210)

There are other elements present in Walsh's account of agency, not least het conceptualization of the goals of an organism as "affordances" (drawing from ecological psychology: Gibson, 1979/2014). These refer to courses of action that the organism is interested in *and* that are allowed for by the environment. This view be viewed as a certain gestalt-shift on the relation between organism and environment – and serves to add further ecological detail to his teleological approach to agency.

The body of work on *autopoiesis* (following Varela, 1979) is a closely allied teleological approach to agency. *Autopoiesis* is a concept that, literally, means "self-making" and is intended to capture just how an organism is organized as to ensure persistence and self-maintenance. At a very intuitive level, it can be thought of as capturing just how organisms are not complex machines that are controlled from outside by natural selection (in the way engineers might design a plane) or by internal parts such as genes (in the way pilots might control a plane). Importantly for our purposes, autopoiesis is a teleological concept in the sense that it identifies the goal of self-maintenance as the overarching directing goal of organismic behavior. So whereas the goals highlighted in Walsh's account are ephemeral ecological goals (e.g., the goal of capturing this particular prey, or of growing towards the sunlight), the "goals" highlighted by the autopoietic approach are general and persistent goal.

The work of Samir Okasha, to mention just one last prominent view on organismic agency, also can be read as an instantiation of the teleological approach. Sometimes the teleology is quite explicit:

> "In [agential thinking about organisms], the telos belongs to an evolved organism (in the paradigm case); the point of treating the organism as agent-like is to capture the fact that its evolved traits, including its behaviour, are adaptive, hence conduce towards the goal of survival and reproduction." (Okasha, 2018, pp. 15–16)

So while Okasha views organismic agency largely as a useful heuristic to represent processes of fitness maximization (the dominant approach in behavioral ecology: Grafen, 2002, 2014), there is still the teleological approach to agency, as the capacity to pursue one general and persistent goal, namely fitness maximization.

In sum, the teleological approach is a very general way of thinking about agency that characterizes very disparate accounts, ranging agency-as-control to agency as responding to affordances in the environment, seeking self-maintenance, or seeking maximal fitness. Some of these are intended as defenses of the reality of agency, whereas others rather as clarifying why agency can be a helpful heuristic. All think of agency as directedness towards some goal.

3. Agency versus Natural Selection

Agency arrives on the scene with a tight-knit family of established concepts that work together to explain goal-directedness in organisms: function, natural selection, plastic reaction norms, and genetic change. Whether the goal-directed behavior concerns fitness, or self-maintenance, or one of the many more specific ecological goals (seeking shelter or nutrition, seeking mates or cooperators, avoiding predators or competitors), there is in principle a candidate explanation available for such behaviors that makes no reference to agency.

This is a problem for teleological accounts of agency. The general reason is simple enough: if natural selection can explain goal-directed behavior, and if what can be explained as resulting from natural selection need not be explained as resulting from agency, then some goal-directed behavior need not be considered as agential.

The key claim here is that agency and natural selection *compete* as explanations for goal-directedness. This is clearly a challenge for the broad understanding of agency: if a behavior can be causally explained as resulting from a mechanism shaped by natural selection, it need not be explained as resulting from the "whole organism". However, it is also a problem if one adopts a narrow understanding of agency, i.e., one where agency is identified with a mental function (e.g., intentional representation, rationality, utility maximization) and then attributed beyond the human realm. For such accounts typically identify *certain conditions* under which organisms may be considered as agents. For what types of goal-directedness is agency a good heuristic, and for what types of goal-directedness is a straightforward selectionist explanation sufficient? Okasha, for instance, seeks an answer in the idea that that the goals of the different parts of an organism must cohere, and form some overall purpose attributable to the whole organism. The goals of the parts can be explained through natural selection; the goal of the whole *may* be identified as agential.

However, one could attempt to go further. It is one matter to argue that it is *possible* to adopt the gestalt where organisms are viewed as goal-directed agents, but quite another to argue that there is a *necessity* for doing so. Why should the agency concept be adopted by the scientist who is quite content with the mechanistic metaphor? Such a scientist may acknowledge that the machine metaphor is imperfect in many respects (Nicholson, 2019), but may nonetheless not see any compelling reason to add the concept of agency to their conceptual toolkit, much less to think it *true* that the nature of organisms is agential. Let us try to go as far as we can in reducing accounts of agency to the family of function, natural selection, plastic reaction norms, and genetic change.

3.1 Reducing Agency to Natural Selection?

The cybernetic idea of convergence on attractor states mediated by negative feedback provides the template for such a reduction. Negative feedback was originally described as the "signals from the goal are used to restrict outputs which would otherwise go beyond the goal" (Rosenblueth et al., 1943, p. 19). Translated into slightly different terms, negative feedback describes a behavioral program, containing *commands* for how organisms need to respond to types of environmental input. The program would take as input "distance between organismal state and goal state", and would then would modulate behavior as to move the organismal state closer to the goal. There is nothing in a process of negative feedback that cannot be programmed into an algorithm.

This can be applied to affordance-seeking. Behavior directed towards the goals of capturing this particular prey, or of growing towards the sunlight, are in principle explainable as the expression of a functional program that takes in certain types of input, and certain types of behavior as output. Similarly, autopoiesis could analyzed as the type of autonomy of a sophisticated *automaton:* the behaviors that generate autopoiesis are a collection of various functional capabilities, each of which have been shaped by natural selection. From this perspective, one could wonder if unity-of-purpose is *sufficient* for viewing an organism as an agent, since well-designed machines – from robot lawnmowers to bottle openers and cars – are characterized by unity-of-purpose. Having an overarching purpose – whether fitness maximization or self-maintenance – does not imply that the organism may be entirely explainable as having been shaped by natural selection.

Let us formalize this argument somewhat. If we assume that one particular stimulus vector (i.e., a combination of various sensory inputs) generates one particular behavioral state, we get a functional relationship between the behavior variable B and the input variable $\mathbf{\Gamma}$

$$B = f(\mathbf{\Gamma})$$

If further assume that the function is computable, then an algorithm can be designed that where all the behavioral responses in response to sensory input have been "pre-decided". The link with natural selection can be understood as follows: since natural selection occurs at the level of types (i.e., types of organism, responding to types of recurring environmental conditions, see e.g. overview in Abrams, 2014), selected behavioral types can be viewed as commands that connect types of environment to types of behavior ("if the environmental input is such-and-such, then produce a behavior that is so-and-so").

A more sophisticated set of commands is entailed by the *reaction norm* of a trait. The reaction norm is a functional mapping of environmental states to phenotypic states (Pigliucci, 2001), and is in this sense a relatively simple program that contains a multitude of commands on how to react in different possible environments. Thus, in "heterogeneous" environments, traits with a flexible reaction norm profile will be selected for (see models in Godfrey-Smith, 1996; Moran, 1992), but if the organism exhibits plastic behavior in response to changes in the environment, this need not be seen as an action of the "whole organism", but simply as the plastic trait carrying out the commands entailed by the reaction norm.

These are two big assumptions – namely that the relation between stimulus and response is a function (i.e., a many-to-one mapping), and that the function is computable – and we will relax them in later sections. However, this analysis can clarify the problem: is "agency" fulfilling an indispensable role in explaining the behavioral function *f*? If the behavioral function can be reduced to a set of elementary commands ("if you see/feel this, then do that"), then there does not seem to be any *a priori* reason why this could not have been shaped by natural selection.

On a more intuitive level, evolution by natural selection here can be thought of as a long process of *pre-decision*. What evolution by natural selection does is decide *beforehand* how an organism will behave: based on regularly recurring patterns in the selective environment, certain behavioral mechanisms will be passed down through the generations more frequently than others, until certain reflexive behaviors are universal.

To what extent is it *realistic* to describe real organismic behavior as a complex set of commands? That is a different question, and the jury is still out on the success of reductionist scientific programs in e.g. computer science or psychology (Carruthers, 2006). The point here is merely that addressing the possibility of such a reduction is a genuine challenge for the concept of agency – for the teleological approach gives reason why such a reduction cannot succeed – and if the reduction can succeed, then "agency" is a concept of *mere* cognitive convenience.

3.2 Saving Teleology at the Cost of Clarity and Simplicity

The teleological approach has as number of responses at its disposal. All involve defining "goal" in a more sophisticated way, such that the type of goal-directedness that natural selection can explain is distinguished from the type of goal-directedness that only agency can explain.

For instance, one possible response would be to question the assumption of *computability*. Not all functions are computable, and there is no good general reason to assume that organismic behavior can be described by computable functions. This means that there may indeed be a functional relationship between sensory inputs and behavior outputs – i.e., each distinct input determines the output (no indeterminacy) – but there still may not be a set of commands (or "algorithm") on how to generate output from input. It is well established that slightly more sophisticated decision theory frameworks – such as in partially observable Markov decision processes, where the organism's information about its environment to be incomplete and imperfectly reliable (Kochenderfer, 2015) – generate non-computable functions. I will not develop this possible response further, but just mention it to illustrate how the concept of computability can help distinction between two types of goal-directed behavior: computable and non-computable goal-directedness. Defenders of the teleological approach could then claim that the second type corresponding to agency.

Another good possible response lies in the concept of *environmental novelty*. Organisms often react adaptively in novel environments: environments that have not occurred before in their lineage. However, this means that natural selection could not have shaped an adaptive response, so natural selection cannot provide an explanation of the organism's behavior (see author ms.). In this way, environmental novelty allows for another distinction between two types of goal-directed behavior: goal-directed behavior in environments that occurred in the evolutionary past, and goal-directed behavior in novel environments. Fully developing this response would require meeting several challenges (e.g., what precisely is "novelty", and to what extent do "novel environments" simply consist of old cues in a new context?), but I mention it as another example of how the teleological approach could be shored up.

My main argument here is not that these responses do not work. They may indeed succeed, and in the process shed new light on the nature of agency. However, their success would come at the cost of undermining the main rationale for the teleological approach. If agency is goal-directedness in response to novelty, for instance, then what is distinctive about agency is not the goal-directedness *per se*.

Instead, it would be desirable to have a simple and easily graspable agency concept that is immediately distinguishable from the functionalist-selectionist family of concepts. So perhaps my critique concerns the values of communicative clarity and pragmatic simplicity. Communicative clarity: how can we clarify what is agency about, in such a way that would be graspable by a non-specialist? After all, everyone has some intuitive idea of what "agency" is. Pragmatic simplicity: a scientist who is familiar and content with their conceptual toolbag of function/selection/mechanism, and who is only willing to invest precious mental energy to follow philosophical-conceptual sophistications if it offers dividends for their scientific work, must be offered an easily graspable and scientifically useful concept of agency. These are perhaps the most important motivations to pursue a different way of thinking about organismic agency.

4. **Deliberation and Human Agency**

To outline this thinking, I suggest to first revisit our representation of the paradigmatic form of agency, *human agency*. Acknowledging the centrality of human agency is not an implicit embrace of anthropocentric thinking; rather, it is a step to examine our own reasoning processes regarding a form of organismic agency about which we have a wealth of empirical data scatters across humanities and social science. Reflecting about the narrow concept of agency – i.e., a mental function – can generate lessons on how the broad concept – i.e., self-causation – should be construed. To further motivate this, consider how the teleological approach to agency draw on narrow conceptions of agency.

4.1 Intentionality and Autonomy

In general, the teleological approach exhibits structural similarities to the intentionality model of human agency. Both intentionality model of human agency and the teleological approach to organismic agency seem to rely on a similar cause-effect structure between goal and behavior. The "goal" of an organism plays the same causal-explanatory role as the "intended future state of affairs" of a human: neither mechanistically causes the action or behavior, and both explain the counterfactual robustness by which the action/behavior tend towards a particular end state. Both need to contend with teh problem of what Hofstadter, drawing on an example of Dean Woolridge, once called "sphexishness" (Hofstadter, 1982): the possibility that elaborate, goal-directed behavior is nonetheless entirely mechanically explainable. For this reason, the teleological approach can be seen as a *de facto* "naturalization" of human intentionality.[2] Conversely, intentionality can be viewed as providing a particular answer to the meaning of self-causation: only when a behavior is caused by a human's *intention* can it be considered to be caused by the human *as a whole* (instead of by an automatic cognitive module, or by some other person).

This connection between intentionality and teleology was explicitly present in founding work in cybernetics, where one key goal was to reflect about the causal

[2] I am using the term "naturalization" as co-extensive with "translation into causal-explanatory and observable terms". Intentionality needs to be naturalized because it cannot be directly observed in organisms; only inferred from behavior.

structure of human agency[3] , represent this structure in a generalized and abstract way, and apply it to living beings and "computing machines" (Rosenblueth et al., 1943; Wiener, 1948/2019). However, while not always so explicitly, it is in the background in other teleological accounts. For instance, it has informed efforts by philosophers of action to identify "primitive" forms of agency, e.g. by Tyler Burge who views primitive agency as consisting of goal-directed ("functional") whole-organism behavior (Burge, 2009). The connection between teleology and intentionality can also help explain the wide presence of the former, in virtue of and the dominance of the latter. Whether in the philosophy of action (Schlosser, 2015.), phenomenology (following Husserl, 1913/2014), or philosophy of mind (Dennett, 1989; Searle, 2000), human action and experience are primarily analyzed in terms of intentionality. Intentionality also frames how culpability is conceptualized in jurisprudence: the level of culpability of a person in many jurisdictions depends on *mens rea,* the level of "intent" present in the mind of that person (Dubber, 2015). It seems fair to conclude that agency-as-intentionality is a widely recurring conceptual prism through which human agency is viewed.

Intentionality is not the only way to understand human agency. Autonomy, for instance, autonomy defines agential actions as those that are "freely" guided by moral ideals or convictions, and not determined by sources of "heteronomy". It captures what it means for a human's behavior to be "freely" guided by moral ideals or convictions, and not determined by sources of "heteronomy". These sources can include sensory inclinations (as emphasized by Kant), but also political sources of tyranny, or even ignorance (preventing informed consent). In this sense, autonomy fills in the meaning of "self-causation" in a slightly different way, and is a concept of agency that is especially common in the context of ethics and politics.

I mention autonomy because autopoiesis can most straightforwardly seen as its naturalization. In fact, sometimes autopoiesis and "biological autonomy" are used as near-synonyms (Moreno & Mossio, 2015; Rosslenbroich, 2014). One could wonder how precisely intentionality and autonomy relate in the present context, but that would bring us beyond what is necessary for present purposes. We do not need to make any important difference between "intentionality models" and "autonomy models".

[3] "Now, suppose that I pick up a lead pencil. (…) Our motion proceeds in such a way that we may say roughly that the amount by which the pencil is not yet picked up is decreased at each stage." (Wiener, 1948/2019, p. 12)

For our purposes, autopoiesis singles out certain very general goals (self-maintenance/persistence), and can be seen as a variation on the intentionality model (where the "intention" is self-preservation).

3.2 Deliberation

With this in mind, it may not seem so strange to introduce a different concept of organismic agency by means of a different angle on human agency. There is a third ways of characterizing human agency that is perhaps most commonly found in applied ethics and virtue ethics, but that has hitherto received little attention: deliberation.

To illustrate the phenomenon, consider the judge reflecting on what sanction to hand to the defendant who has just been found guilty. The jury has already decided on the binary question of guilt and innocence, but deciding what precise sanction is appropriate is one with many more possible outcomes. The personality of the defendant for instance, or the number and nature of prior convictions, may constitute what they call "attenuating" or "aggravating" circumstances. Deliberation refers to the nature of the process rather than the outcome. In the mind of the judge, all factors pertaining to the case are synthesized, and influence the final decision – but in coming to the decision, the judge *weighs* the factors in a way that is open-minded or "blind". These three factors – weighing, no pre-determined outcome, and decision – are the three symbolic features of *Justitia* and are a good way to summarize the nature of deliberation.

Deliberation thus gives a qualitatively different answer to the question: when does a person act, as opposed to being acted upon? The answer lies in the presence or absence of intentionality, but rather details of the decision-making process, and the requirement that this process must have a broadly deliberative structure. *Multiple* goals or values are weighed, the outcome must not be pre-decided, but nonetheless there must be a decision: a response must be chosen.

Deliberative judging is to be contrasted with types of overtly "goal-directed judging", which would more commonly be called biased or ideological judging. Genuine judging is aimed at a state of affairs that best corresponds to the abstract ideal of "justice". However, this abstractness of this ideal is such that it does not determine the weight of the sanction given the empirical state of affairs. The defendant may have been found guilty of killing a person, but the severity of the sanction will depend on their intent (e.g., distinguishing between murder and manslaughter, or manslaughter of the second degree and third degree, etc.), and various aggravating and attenuating circumstances. Genuine deliberation does not pre-decide what the outcome should be – and in fact, if the outcome is pre-decided, as by a corrupt or biased judge might, this would even be grounds for doubting the presence of genuine deliberation.

Part of not pre-deciding the outcome means not ruling out any source of evidence beforehand. Think of the phenomenon of a "show trial": there is a semblance of a genuine deliberation, but the defendants are sentenced harshly because such sentencing suits political aims (and not the aim of justice). Show trials are "a foregone conclusion": while there is some type of normative goal-directedness involved in a foregone conclusion, genuine deliberation is no longer present. These judges disregard the particulars of the case and of the defendants, and subsumes both the case and defendants into general typologies. The judge is not making their own decision, based on the exact case before them, but rather the source of bias or ideology is in some sense "making the decision".

Deliberation is also to be contrasted with *mechanistic judging*. Mechanistic judging has a set of rules, and a finite set of criteria (i.e., further rules) for how to categorize particular situations according to the rules. Some forms of mechanistic judging may be bureaucratic: for instance, when three criteria are met, the sanction may be reduced by 25%; when four criteria are met, by 30%; and so on. All the information that is necessary to pass judgment on the sanction is contained by the rules. The mental operation of this mechanistic judge is *calculation* rather than *deliberation*. There is no deliberation involved, no weighing of possible outcomes, because the result is determined by how the calculation was designed. Instead, the lawmaker (or other member of the judiciary) who *designed* these rules (including application criteria) was the genuine deliberative agent.

This difference between calculation and deliberation is significant, because elsewhere in the organismic agency literature, agency is conceptualized as *utility maximization* and organismic agency as its analogue, namely fitness maximization (Okasha, 2018). This reflects the default in economics, and it might be helpful for presentation purposes to rudimentarily distinguish the approach here from utility maximization. However, once utilities have been assigned to the goals, and once a decision rule has been given on how to weigh these goals when they compete, the true "deliberation" is over. When decision-processes are calculations in this sense, behavior has been "pre-decided" by some designing principle. In the case of organismic agency, the pre-deciding designing principle is natural selection, and getting a grip on just how organismic agency is separate from natural selection is the goal of the next section.

5. The Symmetry-Breaking Approach to Agency

Whereas the naturalistic formalization of the teleological approach to agency is to be sought in attractor dynamics, that of the deliberation model lies in a process of phase transition, or more generally, *symmetry breaking*.

The transition from paramagnetism to ferromagnism is the classic example of a symmetry breaking. A ferromagnet above the Curie temperature (without an external magnetic field) is characterized by spatial symmetry: there is no preferred orientation of the magnetic spins that constitute the metal. However, once the temperature is lowered (so that the kinetic energy of the particles is no longer sufficient to overcome the magnetic force they exert on each other), different regions of homogenous spin emerge. In these different regions, the spatial symmetry has been broken. One particular orientation has been “chosen”, and no external cause has “pre-decided” this process.

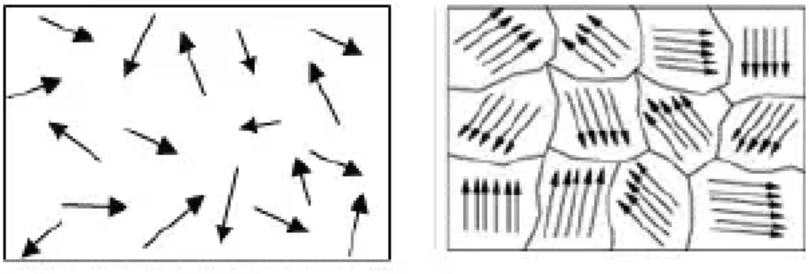

Figure 1: The Breaking of Spatial Symmetries

In general, “symmetries” refer to two different types of outcome that are equally probably given the initial state of the system plus the laws that describe the system’s evolution. Symmetry breaking then refers to a process by which one outcome is preferred by another. Per definition, this process cannot be fully understood as determined by the pre-existing laws. In this way, symmetry breaking is one of the most powerful concepts in physics to explain the appearance of *novelty,* whether novel properties (e.g., superconduction, ferromagnetism) or novel particles, or even novel forces. We do not need to get bogged down in technical details regarding the nature of symmetry breaking (Brading et al., 2021); we only need to note that the concept does not seem to have a universally accepted definition, and that is often viewed as identifying a general explanatory template rather than any rigidly defined theory (Borrelli, 2021). In this sense it can be thought of as a logic or a style of thinking to make sense of temporality, contingency and irreversible changes. With a few exceptions (Longo & Montévil, 2011), it has not featured much in thinking about organisms to the extent that the concepts of equilibrium and approach to equilibrium have. Yet, it is suited to make sense of how an organism can act as an agent.

5.1 Organismic Agency as Breaking Symmetries of Natural Selection

How should it be applied to organisms? What are the dynamics that govern the behavior and development a token organism? The behavior of token organisms is often too noisy to be predicted, and typically biological science is interested in type-level behavior. The possibility of predicting the behavior of token organisms should not be excluded on principle. Some aspects of an organism's behavior may be perfectly explainable in this way. The exact sequence of behaviors by which a wasp burrows its eggs may be explainable by reference to the inheritance of fitness-contributing behaviors.

In general, there may be various goals or "attractors" that shape the behavior of organisms, whether these are ecological goals (competition and cooperation, nutrition and mating, etc.) or goals entailed by inherited developmental plans (Jaeger & Monk, 2014). Goal-directed behavior does not necessarily indicate agency, because the goals may have been shaped over the course of evolutionary history by various "external designing principles". We will especially consider natural selection as one such designing principle, but we need not assume that it is the only one. Developmental biologists have long pointed to the independent explanatory role played by abstract body plans in guiding development its evolution (DiFrisco & Wagner, 2022): in such a view, abstract "body-types" rather than agency may explain certain aspects of an organism's development or behavior.

Also in general, an organism will have *multiple* goals, and may often find itself in an situation where the goals shaped by natural selection (or other external designing principles) *underdetermine* its behavior. Consider an illustrative example: a gazelle may be feeding while a lion approaches. Once the lion has come within a certain distance, the gazelle may enter a state of alert hesitation. Does it continue feeding, or flee the predator? Two goals – nutrition and predator avoidance – compete, and hang in the balance. Neither goal is prioritized, but both goals cannot be simultaneously prioritized (or maximized): an organism cannot both feed and flee at the same time. The state of alert hesitation reflects a *symmetry* between these two goals. It is in such a state – where evolutionary goals underdetermine organismic action, or where there is a *symmetry* between the inherited evolutionary goals – that agency arises.

Informally, the symmetry means that the external designing principle "cannot tell" the organism what to do. We can assume that the cognition gazelle has a host of specialized modules (Carruthers, 2006): a fight-or-flight mechanism that is sensitive to certain types of input (such as large approaching animals), or a hunger-response mechanism that is sensitive to interoceptive signals of hunger or external signals of nutritional opportunity. We can assume these modules are inherited from previous generations where such a mechanism conferred clear fitness advantage, and thus that they are functional and in this sense goal-directed. However, in this particular environmental state, where a lion is approaching, the sensory input produces conflicting output (stay and eat, or run away). The exact environmental state in which the gazelle finds itself is "*novel":* the state of the environment surrounding the token organism cannot be immediately "recognized by natural selection". The selected functions are competing, and in this state of competition where neither has gained the upper hand, the designed attractor states are symmetrical to the organism.

To use the language of deliberation, the organism is "weighing" the different goals, and is also "unbiased" towards the goals. The goals of predator evasion and of nutrition-seeking are *both* "equally" important: one cannot say that one is "more important" than the other in general. In some circumstances, predator evasion must be prioritized; in other circumstances, nutrition-seeking. Agency-as-deliberation thus does not involve an intellectualist notion of agency: there need not be an intention, or mental representation. Deliberation may be present in organisms without a nervous system and without any "cognition" in the narrow sense of the term. Deliberation involves competing selected functions (or in general, competing externally designed goals). Each of the competing mechanisms is activated to different degrees by sensory input; only when sensory input underdetermines the response does agency arise.

However, agency-as-deliberation must involve a decision: a symmetry must arise, but the symmetry must also be broken by the organism. The most straightforward cases of agency involve clear activity by the organism itself to break the symmetry. The gazelle could seek out additional sensory information, for instance by modifying its angle of vision in order to have a better view of the exact mode of approach, or by moving a short distance away and observing whether the predator reacts by approaching further or not. This active searching to break the functional-selectionist symmetry is closely related to what Kim Sterelny once termed "epistemic action" (Sterelny, 2003, chap. 2): an animal may realize that a single cue implies unacceptable risk of a false positive (e.g., the apparent danger from the predator may turn out not to be a danger), and seek additional cues to determine a course of action. However, it would be a mistake to simply conclude agency to be an adaptive strategy to minimize false positives. Agency arises because of the *risk* or *uncertainty*, and false positives are not necessarily risks in themselves (e.g., when they are "cheap"). And one can only speak of risk arises when there are *opportunity costs* associated with competing goals: in this case, mistaken fleeing means that nutrition opportunities are missed. The risk of a false positive must become sufficiently large to cause indecision: it is in this state of symmetry that organism itself must act.

5.2 The Skeptical Case against Deliberation

One could grant that there are straightforward cases of agency. However, symmetry breaking is such a widely applicable notion that it raises the worry that some clearly non-agential systems will be counted as agential. This brings us to the leading class of counterexample (the counterexample that is for deliberation what the thermostat is for the teleological approach). Perhaps there is no simpler counterexample than a rock initially balancing on the precipice, but eventually tipping over towards one side rather than another. This is a form of symmetry breaking, where the symmetry might be broken by a slight wind. What precisely is different between the gazelle's and the rock's symmetry breaking? Like the wind blowing from one side rather than the other, if the lion approaches further, the gazelle will flee, but if the lion heads off in another direction, the gazelle will keep feeding.

This example serves to illustrate the difference between agential and non-agential symmetry breaking: only the former can be termed "deliberation". The rock is obviously not a deliberating agent, because the breaking of the symmetry is not occasioned by the rock itself, but by external forces. The behavior of the rock is similar to the middle of the rope that was being tugged in diametrically opposed directions by two tug-of-war teams. The rock cannot be attributed “self-causation”, since its behavior has been “pre-decided” by the laws of Newton, which determine the acceleration of the rock in response to external force.

Note also that agency does not mean some causal solipsism. There will always be some sensory input from the environment has tipped the organism in one direction rather than the other. If not -- if an organism would make a “decision” without decisive sensory input -- this would not be an example of agency-as-deliberation, but rather an example of noise or of random choice. The “self-causation” of agency thus does not lie in being cut off by external causes, but rather in the way that external input is processed, and whether the way in which that is processed must be accounted for as caused by the organism itself, or caused by some external design principle.

Agency-as-deliberation identifies, in effect, a selective process going on at the level of the organism. The organism is selecting – through deliberation – what behavior is best. So if a behavior is to be considered agential, one needs to primarily compare this with selectionist explanations, and not evaluate the question whether the behavior is “uncaused” by *any* process external to the organism. Why did the gazelle run away as the lion approached? A functionalist explanation will say “because the flight-or-fight mechanism was activated”. An agential explanation will give an answer such as “because the gazelle prioritized the goal of safety over the goal of nutrition”.

The deliberative capacity – the capacity to weigh and resolve the competition between selected functions – can *itself* be the object of natural selection. But this is not where the tension lies. Agency itself can evolve by natural selection – though it goes beyond the purposes of this paper to discuss under what types of circumstance a more powerful deliberative capacity (i.e., a greater “degree” of agency) would give a fitness advantage. But even if agency is a capacity that itself can evolve by natural selection, it is still the organism itself that is responsible for the action when agency is activated. Even if agency is a capacity that itself may have been evolved by natural selection, this does not mean that natural selection has pre-decided the deliberation by the agent. Instead, agency could metaphorically be viewed as a way for natural selection to outsource decision-making to the organism itself, and thus “decentralize” the production of adaptive behavior.

The significance of the counterexample of the rock lies in questioning how we know why the deliberation of the gazelle not like the pseudo-deliberation of the rock? The example pushes us to ask the same question of the deliberation model that was asked of the teleological approach. What grounds do we have for ascribing genuine deliberation to the gazelle, and not some complex form of calculation, where the "rules" have been pre-decided by external designing principles such as natural selection?

5.3 Grounds for Agency Attribution

To answer this question, it is important to set the expectations for the type of answer we can hope for. What is it precisely that an account of agency can hope to achieve?

First, one should not expect an account of agency to generate clear dividing lines that run through the biosphere, separating agents from the non-agents. Agency is a concept that is first and foremost used to explain a behavior of a token organism – it is not the organism as a whole, nor an organism type. Organisms are agential in some respects, and non-agential in other respects – so the question "is this particular organism an agent" is simply not well-formed. Even humans, the paradigmatic agents, are non-agential in many respects. The fact that I fall down in a gravitational field is a decidedly non-agential behavior of mine (that I share with a rock); the fact that I feel hunger after not having eaten for a long time also seems not to be the result of a deliberative process, but one of straightforward functional causation, with input leading to output in a way that is shaped by the evolutionary history of my ancestors. Am I an agent? With regards to many important behaviors, I am (or hope so!) – but with regards to many other behaviors, I am not. The primary explanandum of an agential explanation is a *token behavior* by a *token organism* – not types of behavior, nor types of organism, or even an organism as a whole. This does not mean that we cannot say anything about organisms as a whole, or that we cannot compare the agencies of different organisms. However, when we do so, it is a considered judgment on the basis of token behaviors that may or may not be agential.

Second with regards to token behaviors, one should not expect absolutely definite answers on whether they are agential or not. For instance, one *could* insist that my feeling of hunger is not only determined by my state of nutrition, and would not occur if, for instance, my life would be in danger and adrenaline would be coursing through my body. From this perspective, my hunger is the result of some deliberative process at the level of the "whole organism", and that thus *I* am responsible for the feeling.

To return to the example of the gazelle: whether or not its deliberation is "genuine" or an ersatz form that has been predicted by natural selection, is a question that simply cannot be answered without further empirical details. How hungry is the gazelle? How valuable is the patch of grass or water hole it is feeding or drinking from? A gazelle might allow the lion to approach much closer if the gazelle is close to starvation, or if there is no other watering hole around for tens of miles. How powerful is the evasion capability? A fit, athletic gazelle may be more relaxed than an older or sickly gazelle. Does the gazelle have offspring in the area it needs to protect? Does it have other conspecifics that can protect it? There is no context-free triggering point at which the gazelle will prioritize the goal of predator evasion. It will be weighed against other goals of nutrition, offspring protection, or the goal of staying in a herd. Where the precise tipping point lies depends on empirical details that may differ from environment to environment. As we add further details, our judgment of whether the behavior is agential or not may change.

To draw a general lesson from such considerations: it would be misguided to search for overarching, universally applicable rules to dictate agency attributions. This is not a realistic expectation of what a conceptual-philosophical account of agency can achieve.[4] Ascribing a capacity for deliberation to an organism must *itself* be the outcome of deliberation – a deliberation on the part of the observer who is weighing the available evidence, and then selecting the best explanation for that evidence. Agency attributions thus always occur within epistemic contexts – but that does not mean that agency attributions are epistemic constructs, because the phenomena constrain what can be considered the best explanation.

Instead, what a conceptual-philosophical account can achieve is to identify the contours of the deliberation underlying agency attribution: what are the competing explanation types, and what types of evidence *tend* to support each type? As mentioned in the introduction to this paper, agency competes with the machine metaphor. Agential explanations compete with mechanistic explanations, where behavior is explained as the output of a complex system of interacting functional parts, each designed by some external principle. The important question to ask is, not how

[4] One could surmise that it goes in against the nature of the agency concept, because it would seek to categorize certain stimulus-response patterns as agential regardless of the token organism or further details regarding its exact environment. If such a rule were available, the behavioral function would be computable (see section 3), and that would undermine the rationale for ascribing agency in the first place.

to draw the dividing line *in general,* in some *a priori* way – but rather, how should an observer *reason* about different types of evidence and come to a considered conclusion on whether to consider the target behavior in an agential or mechanistic fashion?

6. Investigating Agency Attributions

The purpose of this paper is to introduce the concept of deliberation (and symmetry breaking) and make a general case why it is a promising way to understand organismic agency. The purpose is not to provide a detailed discussion on how to adjudicate empirical cases. Nonetheless, the fruitfulness of conceptual innovation must lie in more powerful or accurate such adjudication, and to make the case why the deliberation model promises such fruitfulness, I will schematically discuss two empirical case studies. These cases involve a set of observations of how animals behave in response to cues from the environment. The question is then whether the patterns of behavior must be viewed as agential.

6.1 Sexual Choice

Peahens seem to prefer mating with some peacocks over others. The question is what explains their behavior. An agential explanation would model the process as a deliberation: weighing the various traits of potential mates, and making an all-things-considered judgment. A selectionist explanation would model the process as an output of a functional cognitive mechanism that can be triggered by certain types of input.

There is no dearth of theoretical hypotheses of how peahen preferences may have evolved by natural selection. According to the handicap principle (Zahavi, 1975), evolution of peahen preference structure *P* can be explained if *P* allows the organism to better track traits (such as handicaps: a large, heavy train) that convey fitness advantages such as health or immune strength, compared to rival preference structures *P'*. According to the Fisherian runaway mechanism, *P also* allows organisms to track traits that convey fitness advantages, but just a different type of fitness advantage: conformity to pre-existing mating preferences.

However, the question is: what do the empirical data say about the strength of these hypotheses? I review this liteature in other work (Author forthcoming), and only summarize it here. There are at least 6 visual variables that peahens are sensitive to: number of eyespots, eyespot density, train length, train symmetry, eyespot coloration, and eyespot iridescence. Further, there is at least one relevant audiovisual variable: the frequency of vibration of the feathers.

How do peahens process all this information from seven independent sensory variable? What combinations lead to acceptance (copulation), and what combinations lead to rejection? Our understanding of this issue is surprisingly patchy, but researchers have hit on some generalizations. First, a minimum number of eyespots is necessary but not sufficient for acceptance. In other words, a peacock with a lot of eyespots is not guaranteed acceptance by peahens, but a peacock with a low number of eyespots is pretty much guaranteed to be universally rejected. Second, and similarly, the presence of the blue-green eye-color is necessary but not sufficient for acceptance. The third generalization is a significant non-result: researchers have preliminarily concluded that does not seem to be any single variable which, once it assumes a particular value, can guarantee acceptance. Either peahens are picking up on variables that are unknown to observers, or else peahens are evaluating *particular combinations* of sensory inputs.

The question here is whether the preference structure of peahens evolved because it (at least at some time) successfully tracked differences in fitness components (strength, susceptibility to disease, attractiveness to other mates) between peacocks. If so, then peahens with those preference structures would have themselves had higher fitness (because their offspring would have higher fitness, yielding a greater number of grand-offspring). However, the state of the empirical data is such that peahens seem to be more picky than what would be expected from natural selection. They prefer some sensory inputs over other sensory inputs, even though there does not seem to be any implied fitness difference. (In principle, peahens could also be *less* picky than natural selection, preferring some sensory inputs that imply a lower fitness peacock.) This is grounds for attributing agency: peahens are conducting *their own deliberation,* and their choice has not be "pre-decided" by natural selection.

Such an agency attribution is fallible, and could be rejected in the future if additional evidence coming to light (e.g., some unaccounted-for sensory variable that predicts peahen choice, suggesting the presence of a cognitive mechanism that takes that variable as input). Moreover, one could wish for more and more detailed observations of the minutiae of peahen behavior to allow for higher-confidence inferences. Nonetheless, given the available evidence, it is not unreasonable to infer the presence of some agential deliberation done by the peahen when evaluating potential mates. It is still an open question what "values" guide the peahen's deliberation – is it the peahen's own judgment of fitness of the peacock, or do aesthetic values play a role in the peahen deliberation? – but that question concerns the further structure of agency. The question at stake here is whether the evidence allows for an inference of agency, and current evidence does suggest it may make more sense to explain the exhibited preference structures as resulting from a genuine deliberation by the organism itself, rather than the activation of a cognitive mechanism where the preferences have been "pre-decided" by natural selection.

6.2 Bacterial Locomotion

Locomotion – the changing of an organism's spatial position from one time to the next – is one of the most basic ecological interactions with the external environment. Is it an example of agency? The selectionist-mechanistic explanation is rather straightforward: chemotaxis is a mechanism that takes certain sensory cues as inputs, and produces motor outputs. It is adaptive in heterogeneous environments (Keegstra et al., 2022), has evolved multiple times, and therefore can be hypothesized to be the result of evolution by natural selection.

However, the empirical details of real chemotaxis behavior are much more complicated than would seem given such a selectionist-mechanist hypothesis. For instance, it turns out that bacteria, surprisingly often, are attracted by compounds that are not particularly nutritious – and conversely, are not strongly attracted by highly nutritious compounds (Keegstra et al., 2022). The relation between the strength by which bacteria are attracted by a compound, and the positive effect that compound has on the bacteria's growth rate, is – in general – very noisy. Hence, even if one assume that a particular type of chemotaxis evolved through natural selection, one cannot simply assume that the evolutionary function of chemotaxis lies in maximizing exposure to compounds that benefit the growth rate. And in fact, Keegstra et al. review how chemotaxis has other ecological functions, including expansion into novel environments (Keegstra et al., 2022).

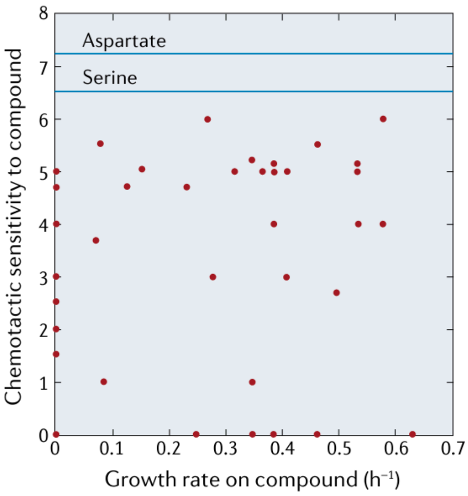

Figure 2: The relation between the attractiveness and nutritiousness of a compound is surprisingly noisy. (Keegstra et al., 2022, p. 493)

So let us return to the question: is it plausible to believe that natural selection pre-decided how an individual *E. coli* behaves? Table 1 is a non-exhaustive summary of the "preference structure" of *E. coli*, which characterizes the explanandum (*E coli* behavior) with more empirical precision. The table can be read as a mapping from sensory input (amino acids, sugars, etc.) to behavioral put (attraction/repulsion). Unlike in the peahen case, the mapping is highly modular, in the sense that each sensory input is processed by a dedicated mechanism (i.e., a receptor). Of these mechanisms, Tar and Tsr are the most abundant, and are sensitive to two of the most important sensory inputs – aspartate and serine level. The other chemoreceptors (Tap, Trg) that are much less abundant in the periplasm, but that can modulate responses in collaborative networks of chemoreceptors.

Input		Receptor	Output
Amino Acid	Aspartate	Tar	+
	Dipeptides	Tap	+
	Serine	Tsr	+
	Leucine	Tsr	-
Mineral	metal ions	Tar	-
Sugar	Ribose	Trg	+
	Galactose	Trg	+

	Maltose	Tar	+
Oxygen		Aer/Tsr	+/-

Table 1: Simplified preference structure of E. Coli (Ortega et al., 2017; Taylor et al., 1999; Yamamoto et al., 1990)

A crucial empirical fact is that thus different receptors form *arrays*. This means that receptors do not simply compete, but also “cooperate”: within an array, the stimulation state of one receptor can influence the output of the whole (Parkinson et al., 2015). This raises the further question: to what extent is the array – the precise frequency and location of receptor types – itself designed by natural selection? If the precise structure of the array is idiosyncratic, then this is grounds for viewing the array as an individual property, of the token bacteria, rather than as a trait that has been inherited over generations.

Unlike in the previous section, I do not wish to make an evaluative judgment on the agential nature of the behavior in question. Rather I will outline the *types of evidence* that would allow one to judge an agential explanation to be the best one. Bacteria can evidently be in a "symmetric state” with regards to the externally designed goals: if the sensory input underdetermines the behavioral output, by triggering the different chemoreceptor arrays in opposing ways, bacteria exhibit “pausing behavior” (Eisenbach et al., 1990). However, the difficult question is: if a bacteria exits such pausing behavior, opting to move in one direction rather than another, is the token bacteria responsible for this decision, or has it been “pre-decided” by natural selection?

Natural selection may have designed how to associate a *type of response* to a *type of input,* and to associate that *type of response* with a higher fitness outcome in the environment that tends to correspond to that *type of input*. Sensitivity to serine and aspartate seem to be explainable that way: while these amino acids do not strongly stimulate growth rates, they are taken as signals of nutrition-rich environments. The behavioral trait of attraction to aspartate thus can be explained as evolution by natural selection. However, can a *token response* to a *token input* be explained as a combination of selected *types*? How good is the explanation of a particular instantiation of behavior as a combination of various *functional behaviors*? Once functional pathways start to compete, is the mode of the resolution of the competition also pre-decided by natural selection? Or is it left underdetermined, and must we explain the symmetry breaking by reference to the organism itself? The answer to this question would determine whether bacterial locomotion is at least in part an agential process.

7. Conclusion

Agency-as-deliberation identifies, in effect, a selective process going on at the level of the organism. The organism is selecting what behavior is best, by weighing and selecting a course of action in light of its various competing goals. Agency-as-deliberation integrates some core insights from the teleological approach, but differs from the teleological approach in that the essential element of agency lies in the competition and selection between goals. Moreover, deliberation clarifies in an elegant way how agency is *a counterpart to natural selection*. Natural selection describes a selective process "carried out by" the environment (though this is an entirely metaphorical way of speaking). Agency-as-deliberation describes a selective process carried out by the organism. This parallel between natural selection and agency, I submit, is the most fundamental attraction of the deliberation model proposed here. It suggests a conceptual framework with which agency can be further developed into a major principle of biological science, on par with natural selection. Agency-as-deliberation clarifies just how the whole organism – and not its selective environment, or its ancestors – is the cause of its own behavior, and this is the main reason for speaking of agency in the biological context.

REFERENCES

Abrams, M. (2014). Environmental Grain, Organism Fitness, and Type Fitness. In G. Barker, E. Desjardins, & T. Pearce (Eds.), *Entangled Life* (Vol. 4). Springer Netherlands. https://doi.org/10.1007/978-94-007-7067-6

Allen, C., & Bekoff, M. (1997). *Species of mind: The philosophy and biology of cognitive ethology*. MIT Press.

Arnellos, A., & Moreno, A. (2015). Multicellular agency: An organizational view. *Biology & Philosophy*, *30*(3), 333–357. https://doi.org/10.1007/s10539-015-9484-0

Borrelli, A. (2021). Between symmetry and asymmetry: Spontaneous symmetry breaking as narrative knowing. *Synthese*, *198*(4), 3919–3948. https://doi.org/10.1007/s11229-019-02320-8

Brading, K., Castellani, E., & Teh, N. (2021). Symmetry and Symmetry Breaking. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/symmetry-breaking/

Brosnan, S. F. (2023). A comparative perspective on the human sense of justice. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2022.12.002

Burge, T. (2009). Primitive Agency and Natural Norms. *Philosophy and Phenomenological Research*, *79*(2), 251–278. https://doi.org/10.1111/j.1933-1592.2009.00278.x

Carruthers, P. (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Clarendon Press.

Dennett, D. C. (1989). *The Intentional Stance*. MIT Press.

Desmond, H., & Huneman, P. (2020). The Ontology of Organismic Agency: A Kantian Approach. In A. Altobrando & P. Biasetti (Eds.), *Natural Born Monads: On the Metaphysics of Organisms and Human Individuals*. (pp. 33–64). De Gruyter.

DiFrisco, J., & Wagner, G. P. (2022). Body Plan Identity: A Mechanistic Model. *Evolutionary Biology*, *49*(2), 123–141. https://doi.org/10.1007/s11692-022-09567-z

Dubber, M. D. (2015). *An Introduction to the Model Penal Code* (2nd ed.). Oxford University Press.

Eisenbach, M., Wolf, A., Welch, M., Caplan, S. R., Lapidus, I. R., Macnab, R. M., Aloni, H., & Asher, O. (1990). Pausing, switching and speed fluctuation of the bacterial flagellar motor and their relation to motility and chemotaxis. *Journal of Molecular Biology*, *211*(3), 551–563. https://doi.org/10.1016/0022-2836(90)90265-N

Fulda, F. C. (2017). Natural Agency: The Case of Bacterial Cognition. *Journal of the American Philosophical Association*, *3*(01), 69–90. https://doi.org/10.1017/apa.2017.5

Gambarotto, A., & Nahas, A. (2023). Nature and Agency: Towards a Post-Kantian Naturalism. *Topoi*. https://doi.org/10.1007/s11245-023-09882-w

Gibson, J. J. (2014). *The Ecological Approach to Visual Perception*. Psychology Press. (Original work published 1979)

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge University Press.

Grafen, A. (2002). A First Formal Link between the Price Equation and an Optimisation Program. *Journal of Theoretical Biology*, *238*, 541–563.

Grafen, A. (2014). The formal darwinism project in outline. *Biology & Philosophy*, *29*(2), 155–174. https://doi.org/10.1007/s10539-013-9414-y

Hofstadter, D. R. (1982). Metamagical Themas. *Scientific American*, *247*(3), 18-M18. https://www.jstor.org/stable/24966674

Husserl, E. (2014). *Ideas: General Introduction to Pure Phenomenology*. Routledge. (Original work published 1913)

Jaeger, J., & Monk, N. (2014). Bioattractors: Dynamical systems theory and the evolution of regulatory processes. *The Journal of Physiology*, *592*(11), 2267–2281. https://doi.org/10.1113/jphysiol.2014.272385

Keegstra, J. M., Carrara, F., & Stocker, R. (2022). The ecological roles of bacterial chemotaxis. *Nature Reviews Microbiology*, *20*(8), 491–504. https://doi.org/10.1038/s41579-022-00709-w

Kochenderfer, M. J. (2015). *Decision Making Under Uncertainty: Theory and Application*. MIT Press.

Lee, J. G., & McShea, D. W. (2020). Operationalizing Goal Directedness: An Empirical Route to Advancing a Philosophical Discussion. *Philosophy, Theory, and Practice in Biology*, *12*(20220112). https://doi.org/10.3998/ptpbio.16039257.0012.005

Liljeholm, M. (2021). Agency and goal-directed choice. *Current Opinion in Behavioral Sciences*, *41*, 78–84. https://doi.org/10.1016/j.cobeha.2021.04.004

Longo, G., & Montévil, M. (2011). From physics to biology by extending criticality and symmetry breakings. *Progress in Biophysics and Molecular Biology*, *106*(2), 340–347. https://doi.org/10.1016/j.pbiomolbio.2011.03.005

Moran, N. A. (1992). The Evolutionary Maintenance of Alternative Phenotypes. *The American Naturalist*, *139*(5), 971–989. https://doi.org/10.1086/285369

Moreno, A., & Mossio, M. (2015). *Biological Autonomy*. Springer. https://doi.org/10.1007/978-94-017-9837-2

Nadolski, E. M., & Moczek, A. P. (2023). Promises and limits of an agency perspective in evolutionary developmental biology. *Evolution & Development*, ede.12432. https://doi.org/10.1111/ede.12432

Nicholson, D. J. (2019). Is the cell really a machine? *Journal of Theoretical Biology*, *477*, 108–126. https://doi.org/10.1016/j.jtbi.2019.06.002

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

Ortega, Á., Zhulin, I. B., & Krell, T. (2017). Sensory Repertoire of Bacterial Chemoreceptors. *Microbiology and Molecular Biology Reviews*, *81*(4), e00033-17. https://doi.org/10.1128/MMBR.00033-17

Paolo, E. A. D. (2005). Autopoiesis, Adaptivity, Teleology, Agency. *Phenomenology and the Cognitive Sciences*, *4*(4), 429–452. https://doi.org/10.1007/s11097-005-9002-y

Parkinson, J. S., Hazelbauer, G. L., & Falke, J. J. (2015). Signaling and sensory adaptation in Escherichia coli chemoreceptors: 2015 update. *Trends in Microbiology*, *23*(5), 257–266. https://doi.org/10.1016/j.tim.2015.03.003

Pigliucci, M. (2001). *Phenotypic Plasticity: Beyond Nature and Nurture*. The John Hopkins University Press.

Prigogine, I. (1947). *Étude thermodynamique des phénomènes irréversibles:* Desoer.

Prum, R. O. (2017). *The Evolution of Beauty: How Darwin's Forgotten Theory of Mate Choice Shapes the Animal World - and Us*. Knopf Doubleday Publishing Group.

Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, Purpose and Teleology. *Philosophy of Science*, *10*(1), 18–24.

Rosslenbroich, B. (2014). *On the Origin of Autonomy: A New Look at the Major Transitions in Evolution*. Springer Science & Business Media.

Ryle, G. (2009). *The concept of mind*. Routledge. (Original work published 1949)

Schlosser, M. E. (2015). Agency. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2015). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2015/entries/agency/

Schrödinger, E. (1992). *What is Life?: With Mind and Matter and Autobiographical Sketches*. Cambridge University Press. (Original work published 1944)

Searle, J. R. (2000). *Mind, language & society: Philosophy in the real world* (1. paperback ed., [2. pr.]). Basic Books.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sterelny, K. (2000). *The Evolution of Agency and Other Essays*. Cambridge University Press.

Sterelny, K. (2003). *Thought in a Hostile World: The Evolution of Human Cognition*. Wiley.

Sultan, S. E., Moczek, A. P., & Walsh, D. (2022). Bridging the explanatory gaps: What can we learn from a biological agency perspective? *BioEssays*, *44*(1), 2100185. https://doi.org/10.1002/bies.202100185

Taylor, B. L., Zhulin, I. B., & Johnson, M. S. (1999). Aerotaxis and Other Energy-Sensing Behavior in Bacteria. *Annual Review of Microbiology*, *53*(1), 103–128. https://doi.org/10.1146/annurev.micro.53.1.103

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

Varela, F. J. (1979). *Principles of Biological Autonomy*. North Holland.

Walsh, D. (2015). *Organisms, Agency, and Evolution*. Cambridge University Press. https://doi.org/10.1017/CBO9781316402719

Wiener, N. (2019). *Cybernetics: Or, Control and communication in the animal and the machine* (Second edition, 2019 reissue). The MIT Press. (Original work published 1948)

Yamamoto, K., Macnab, R. M., & Imae, Y. (1990). Repellent response functions of the Trg and Tap chemoreceptors of Escherichia coli. *Journal of Bacteriology*, *172*(1), 383–388. https://doi.org/10.1128/jb.172.1.383-388.1990

Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214. https://doi.org/10.1016/0022-5193(75)90111-3