The Deliberation Model of Organismic Agency

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 though the phenomena of mate choice and bacterial chemotaxis.

Keywords: Agency – Organism – Mechanism – Deliberation – Goal-directedness

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).

This development faces a number of challenges. One is that the use of the term "agency" in the biological context inevitably leads to the impression that that organismic agency is *in some sense analogous* to human agency. In areas such as philosophy of action, jurisprudence, ethics, or politics, an agent is generally understood to be 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. Not all these are necessarily appropriate for the biological context.

If the first worry concerns what agency "is", the second and perhaps deeper set of worries concerns the explanatory status of agency. Do we need a concept of agency? Certain areas of scientific practice do suggest a real explanatory function, ranging from attributing a "sense of justice" to apes (Brosnan, 2023) to a "sense of beauty" to birds (Prum, 2017). Even so, it remains controversial how widely applicable such agential

language is, and whether it cannot ultimately be dispensed with. As flawed as the machine metaphor may be (Nicholson, 2019), it has seen off many contenders over the centuries, from monads to élan vital, and the fate of agency is uncertain.

These questions and challenges tend to be answered simultaneously, and accounts of agency have been clustering in what I will for convenience term the "broad" and "narrow" senses of agency. Agency, in the narrow sense, refers to some mental capacity that humans possess, and perhaps some organisms too. Thus, agency has been viewed as the capacity for intentional representation (e.g., Allen & Bekoff, 1997; Sterelny, 2000), or as the capacity for utilitarian calculation (e.g., Okasha, 2018). In the broad sense of the term, agency refers to the capacity an entity to cause its own behavior (Moreno & Mossio, 2015; Walsh, 2015; Desmond & Huneman, 2020). On this understanding, agency is causal-explanatory concept – not a mental capacity – applicable in principle to any entity, whether bacteria, humans, swarms, or artificial intelligences.

Both senses of agency face their own difficulties. If one adopts narrow-sense agency, either one ends up reserving agency for a small subset of biological species – but then "agency" is no longer interesting as a fundamental biological concept that can rival the machine metaphor – or else one broadens the concept of mentality to include many species that may not even possess nervous systems. This raises the old danger of committing the category mistake of attributing mental terms to causal-physical processes (Ryle, 1949/2009).

For these reasons, this paper chooses to adopt broad-sense agency. Here the major difficulty lies in clarifying what exactly it means for an organism to "cause its own behavior". If A causes B, this seems to imply that A and B are nonidentical. The notion of "self-causation" thus appears to be self-contradictory when taken literally. Even so, it makes considerable sense at least in the human context to judge a person to be a cause of their own behavior (and hence responsible or culpable). How can self-causation be made sense of in causal terms – *without* relying on some mental function?

The target of this paper is arguably the dominant approach to broad-sense agency: the teleological approach. On this approach, an organism is an agent if it is capable of directing its behavior towards beneficial goals. I will argue we need another approach altogether, mainly because the teleological approach cannot adequately distinguish agency from function. If natural selection has shaped the goal-directed

behavior of an organism, the organism is not "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.

Instead, this paper proposes an alternative causal analysis of self-causation, and to that end, somewhat ironically, will also invoke an alternative narrow understanding of agency: agency as the capacity for *deliberation*. 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. Or in other words, if an organism exhibits a

capacity for goal-directed behavior, it can be considered an agent. This is the dominant way of thinking about agency today.

Note that the teleological approach, as such, does not define what counts as a "goal" or even a "capacity". Hence as such, not denote any specific dividing line between agents and non-agents. Thermostats, for instance, may or may not be agents on the teleological approach: it depends on how the underlying concepts of goal and capacity are defined. Instead, the teleological approach denotes a *logic* or style of reasoning about agency, and points to how self-causation can be analyzed in clear causal terms.

One major category within the teleological approach are cybernetic approaches, where goal-directedness is specifies as "control". This idea traces back to (Rosenblueth et al., 1943), where control refers to negative feedback between organism and goal, where signals about organism's proximity to the goal are used to modulate behavior. Hence the term "cybernetics", as an anglicization of χυβερνήτης (kubernḗtēs) or "steersman" (Wiener, 1948/2019, p. 18). On this approach, organisms as agents insofar they actively "steer" their activities towards certain goals. It is worth considering cybernetics in slightly more detail, as its influence continues to this day.

To yet further break down control into constituent causal processes, some of the form tools of attractor dynamics are used.[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. Attractor dynamics shows just how the negative feedback between organism and goal works.

Attractor dynamics still shapes the contours of today's teleological approaches. 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

[1] Cybernetics did not come from nowhere, and originated around the same time of many other efforts to further extend the explanatory reach of statistical physics to explain biological phenomena (Prigogine, 1947; Schrödinger, 1944/1992; Shannon, 1948).

> decisions about what will work best at various points in a dynamically unfolding situation. (Tomasello 2022, p. 11)

Alternatively, in Lee and McShea's analysis of teleology (Lee & McShea, 2020), a distinction is made between the two metrics *persistence* (resistance to perturbation) and *plasticity* (insensitivity to initial conditions) – both of which describe different basic properties of attractor dynamics.

Denis Walsh's influential work on organismic agency represents 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. However, when it comes to analyzing just how organisms causally interact with their goals, a cybernetic analysis is charted: agency is defined 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" (Walsh, 2015, p. 194) -- i.e., attractors.

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 consist of a single general and persistent goal.

All these accounts of agency adopt some core cybernetical ideas. However, one need not view goal-directedness in this way. Control is just one specific way of causally analyzing what is going on when an organism is deemed to "cause its own behavior". For instance, one could view organismic agency largely as a useful heuristic to represent fitness maximizing behaviors (the dominant approach in behavioral ecology: Grafen, 2002, 2014). The work of Samir Okasha can nonetheless be read as an instantiation of the teleological approach:

> "In [agential thinking about organisms], the telos belongs to an evolved organism (…) 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)

While this approach does not particularly draw on cybernetics, this still can be considered a teleological approach to agency, where agency refers to the capacity to pursue one general and persistent goal, namely fitness maximization.

Today's literature on agency is somewhat scattered and very transdisciplinary: the claim here is not that *all* approaches to agency today are teleological. Nonetheless, it does seem fair to assume going forward that the teleological approach is the *default* one, even to the point that it is often assumed that agency just *is* the capacity for goal-directedness – whether that is goal-directedness towards affordances, self-maintenance, maximal fitness, or other measures not discussed here such as minimal surprise (Friston, 2012). The teleological way of thinking of agency is very widespread, and whole debates (e.g., about whether agency is "real" or a mere heuristic) can take place within a teleological framework. The goal of this paper is to make the case is that it should be abandoned for an alternative (i.e., deliberation). The next section outlines the main weakness in the teleological approach.

3. Agency versus Natural Selection

Does goal-directedness in organisms require a concept of agency? The etiological account of *function* (Wright, 1973), building on (Mayr, 1961; Pittendrigh, 1958), shows how it does not. Consider the following passage by Mayr:

> "An individual who-to use the language of the computer-has been "programmed" can act purposefully. (…) A bird that starts its migration, an insect that selects its host plant, an animal that avoids a predator, a male that displays to a female-they all act purposefully because they have been programmed to do so. (Mayr, 1961, pp. 1503–1504).

Mayr's analogy between natural selection and a programmer should be a real concern for accounts of agency. It shows how organisms could plausibly be viewed as complex machines, and seems to preclude the necessity of any concept of agency. Of course, Mayr's analogy is *promissory*. Currently, we cannot explain organisms in this way. However, perhaps future generations of scientist might. This may seem weak and unfair to those who are unsympathetic to the machine metaphor. However, in the biological sciences as a whole (including genetics and the medical sciences), the machine metaphor is still very much the null approach to organisms. So a scientist could acknowledge that organisms are not "really" machines (Nicholson, 2019), and still be unwilling to buy into "agency".

Here is an analogy. A contractor has a set of trusted tools that gets all the jobs done, but is shown new, powerful, but awfully expensive tool. Should the contractor buy the tool? Pointing to the imperfection of the contractor's current tools would miss the point. Instead, one should point to an important job that can only be undertaken with the new tool, thereby justifying the expense. In a way, agency is an expensive, new concept. Is it really needed? To be of importance to science, the concept of agency must be explanatorily indispensable for certain types of problems (see also Desmond & Huneman, 2020). There is a host of other concepts – evolution by natural selection, genetic change, reaction norms – that can powerfully explain behavior. If agency is to become established as a scientific concept, one needs to actively make the case what explanatory jobs it is tailored for. The critique the teleological approach offered in this section is that it cannot do this.

3.1 The Organism as Designed Program

Cybernetics shows how Mayr's intuition can be made more precise – after all, it is predicated on the idea that organisms and computers are not dissimilar. 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.

Consider some of the major teleological approaches discussed in the previous section. Behaviors that appear to be affordance seeking – e.g., behavior directed

towards the goals of capturing this particular prey – could be analyzed as outputs a functional program (or reaction norm) that takes in environmental cues as input. Similarly, the behaviors that constitute autopoiesis can be analyzed as various functional capabilities, each of which have been shaped by natural selection. Autopoiesis becomes the the type of autonomy of a sophisticated *automaton*.

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 **Γ**:

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

Γ contains all possible input (sensory) variables. The reaction norm is an important part of this behavioral function, since it takes in environmental states and maps these onto phenotypic states (Pigliucci, 2001). The reaction norm commands the organism on how to react in different possible environments.

As such, this simply describes how organisms behave in response to inputs. The important question is: what explains *f*? In Mayr's intuition, it is natural selection – not the organism – that designs *f*. 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"). Sometimes natural selection gives many different commands, according to what exact environment the organism finds itself in: in "heterogeneous" environments, traits with a flexible reaction norm profile will be selected for (see models in Godfrey-Smith, 1996; Moran, 1992). In informal (and entirely metaphorical) terms, what natural selection does is *pre-decides* what organisms need to do. On this view, 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 mechanisms are universal. Evolution by natural selection is a long process of pre-decision.

In sympathetically portraying the analogy between natural selection and a computer programmer, I only wish to establish its plausibility and to justify the assumption, going forward, that natural selection operates as a designing principle.

How precisely this view relates to other views of natural selection, for instance as a causal force consisting of fitness differences (following Sober, 1984) or as a statistical description of births and deaths in a population (following Walsh, 2000), is a different question not considered here. The lesson is that one cannot simply assume that the observation of goal-directedness in and of itself justifies considering organisms as agents. If the behavioral function *f* can be reduced to a set of elementary commands ("if the individual sees/feels this, then they must do that"), then there does not seem to be any *a priori* reason why even goal-directed behavior could not be determined by natural selection. In such a case, it does not seem that "agency" fulfills any indispensable role in explaining the behavior (including development) of the organism.

3.2 Refining Teleology at the Cost of Clarity and Simplicity

Can the behavioral function always be reduced to a set of commands? This is the key assumption: *computability*. Computability is a concept that is native to logic and computer science (for a good introduction, see Hamkins, 2020 chapter 6), and as such does not usually feature much in biological contexts. However, because we have presented an organism's behavior as a functional mapping (i.e., a many-to-one mapping, of sensory inputs to behavioral/developmental outputs), it is instructive to enquire in rough and qualitative terms what it means biologically for computability to fail to hold.

Consider decision theory frameworks as a way of predicting organismal behavior (see Kochenderfer, 2015). For instance, Markov decision processes model an organism's behavior as determined by current sensory inputs (past inputs are ignored). These are computable, or decidable. However, *partially observable* Markov decision processes allow the sensory inputs only yield incomplete and imperfectly reliable information about the environment. These decision processes seem to better describe biological reality, since cues about the environment are incomplete and imperfectly reliable. Such processes have been shown to be undecidable (Madani et al., 1999). In intuitive terms, this means that the behavior of the organism may still be *causally determined* by its sensory input (i.e., there is still a behavioral function), but that there is no set of commands (an "algorithm") that exhaustively decides how behavioral output is to be generated from sensory input.

The trio of environmental *uncertainty, heterogeneity* and *novelty* describes the types of environment where computability fails. For instance, organisms can react adaptively in "novel" environments, defined as environments that have not occurred before in their selectionist history. Further developing line of reasoning will be the topic of the next section, but provisionally it can be precisified as follows: an organism can behave adaptively even if the immediate environment cannot be categorized as one particular state of the selective environment (see Brandon, 1990; Desmond, 2022). This means, per definition, that natural selection could not have shaped the adaptive response.

The main lesson here is that goal-directedness *per se* is not definitive of agency: it is goal-directedness in the face of certain types of environmental uncertainty, or in the face of certain types of environmental novelty. Thus, the teleological approach is not necessarily *wrong:* it can potentially be refined. My critique is intended to persuade why it would be *preferable* to have an alternative approach that does not require such refinements. After all, the teleological approach is attractive as a *logic:* it helps to structure thinking about agency in terms of a clear and simple notion, namely, "goal-directedness". However, once we define agential behavior in terms of a "non-computable goal-directed behavioral functions", this clarity and simplicity is compromised. To return to the tool analogy, the teleological approach to agency yields at best an expensive and difficult-to-wield tool. Ideally, a technical concept of agency in biological context would be easily communicable to non-specialists, and not require scientists to expend precious mental energy in following philosophical-conceptual sophistications. These are the motivations for a clean, alternative approach to agency.

4. Deliberation and Human Agency

To outline this thinking, I suggest that we first revisit our representation of that paradigmatic form of agency, *human agency*. Reflecting on narrow-sense agency (i.e., agency as mental function) can generate lessons on how broad-sense agency (i.e., agency as self-causation) can be intuitively construed.

4.1 Intentionality and Autonomy

In fact, the teleological approach to broad-sense agency is implicitly allied with a particular construal of narrow-sense agency, which we can for purposes here title the

intentionality model of human agency. Both the intentionality model and the teleological approach to organismic agency 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 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). While not always so explicitly, intentionality is also in the background of 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, given 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

[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, and only inferred from behavior.

[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)

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 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". 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 close connection between narrow-sense and broad-sense agency (mental function and self-causation), it may not seem so strange to introduce a different concept of organismic agency by means of a different angle on human agency: deliberation. This third way of characterizing human agency is perhaps most commonly found in applied ethics and virtue ethics.

As a preliminary illustration, 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 so-called "attenuating" or "aggravating" circumstances. Deliberation refers to the

nature of the judge's mental process, where all factors relevant for a range of possible outcomes are weighed, and synthesized into a single, final decision. One easy and accurate way of characterizing deliberation is by means of the three symbolic features of lady *Justitia*: the blindfold (no pre-determined outcome), the scales (weighing of factors), and the sword (a decision).

When does a person act, as opposed to being acted upon? On the deliberation model, the difference maker lies in the details of the decision-making process, and the requirement that this process has a broadly deliberative structure. In particular, the presence or absence of intentionality is not necessarily what matters. Contrast deliberative judging with types of *goal-directed judging*, which would more commonly be called biased or ideological judging. Here, the full range of outcomes is not considered. A set of values or ideology has pre-decided how the judge will form the sentences – it is not really the judge *themselves* who is doing the deciding. Thus a judge may intentionally aim at a particular state of affairs (e.g., a particular outcome), but in the process not deliberate on the course of action, and be acting as an instrument for a broader social ideology. In the extreme case, such a case becomes a "show trial": where there is only a mere a semblance of a genuine deliberation, and where the defendants are sentenced 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".

Some type of intentionality does seem to be present in genuine judging, insofar it is 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. Intending to realize "justice" is not like intending to drink a cup of coffee: it may be unclear beforehand, before all the minute details of a case have been taken into consideration, what exactly the "just sanction" is among the range of possible sentences. So if deliberation involves intentionality, it is intentionality towards some abstract value that does not directly correspond with particular empirical states of affairs. Insofar organisms can be assumed not to possess such intentionality, we will

later assume that goal-directedness is neither sufficient, nor necessary for deliberation.

A second important contrast is with *mechanistic judging*. Mechanistic judging has a set of rules on how to act in a given situation, and a finite set of criteria for how to categorize particular situations according to the rules (i.e., more rules that govern the application of rules). For instance, a rule could tell the judge: "when three attenuating factors are present, the sanction may be reduced by 25%; when four factors are present, 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 not so much deliberation rather than *calculation*. 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, narrow-sense agency is conceptualized as *utility maximization,* with organismic agency viewed as a heuristic where organisms are treated *as if* they are maximizing an analogue of utility, namely, fitness (Okasha, 2018). This reflects the default approach to human agency in economics, but according to the presentation in this paper, utility maximization is not necessarily deliberation. Once utilities have been assigned to the possible courses of action, and once a decision rule has been given (i.e., maximize utility), the true "deliberation" is over. There may be deliberation involved in assigning utilities to courses of action, but this is not what utility calculus covers. Making decisions by maximizing utility presupposes that behavior has been "pre-decided" by some designing principle. In the next section, we construct a broad-sense notion of agency by adding biological and causal detail to the concept of deliberation.

5. Deliberation as Symmetry-Breaking

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.

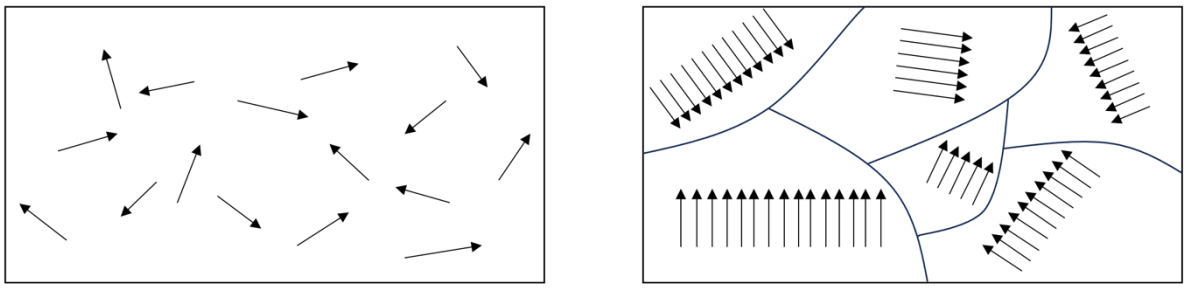

Figure 1: The Breaking of Spatial Symmetries

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 comes to be "preferred" over another, even though the initial state and laws were indifferent to the two outcomes. 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., superconductivity, ferromagnetism) or novel particles, or even novel forces.

There is a dedicated debate on the nature of symmetry breaking (Brading et al., 2021), but we do not need to engage with it in detail since we will only be using symmetry breaking in a rather qualitative sense. As justification, it is important to note that symmetry breaking does not 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. And with rare exceptions (Longo & Montévil, 2011), it has been largely overlooked in the biological context, and certainly has not featured to the extent that the concepts of equilibrium and approach to equilibrium have.

5.1 Organismic Agency as Breaking Symmetries of Natural Selection

How should symmetry breaking be applied to organisms? Ideally, we would first specify the laws or dynamics that govern the behavior and development a token organism. However, there are of course no such laws available, at least if we understand a "law" to be an exceptionless generalization. We can exactly predict the motion of planets many centuries in the future, but we cannot predict where exactly a swimming duck even as we stand by a pond. That is not to say that there are pockets of individual-level predictability. For instance, the exact sequence of behaviors by which a wasp burrows its eggs may be explainable by reference to the inheritance of fitness-contributing behaviors. However, in general, the behavior of token organisms is relatively noisy.

Instead, biological science tends to be interested in type-level behavior. Now, even at this level, biology textbooks will never use the term "law" as a physics textbook might. Even so, the description of any species – the brown bear, the purple emperor butterfly, the parasitic hookworm – will consist of various generalizations about their habitat, their reproductive lifecycle and mating behaviors, their typical appearance, their dietary habits, their interaction with predators or competitors, and so on. Some (not all) of these generalizations may describe *functional* behaviors. Insofar functional behaviors are goal-directed, attractor dynamics could be repurposed to reformulate type-level behaviors. These may involve attraction to ecological goals (competition and cooperation, nutrition and mating, etc.) or to goals entailed by inherited developmental plans (Jaeger & Monk, 2014). While we will especially consider natural selection as the source of these goals, in general it need not be 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 or natural selection may explain certain aspects of an organism's development or behavior.

So let us assume these evolutionary goals as given. The types of situation of interest for this paper are environmental states where the goals *underdetermine* a token organism's behavior. The different goals "tug" the organism towards different, mutually incompatible behaviors. 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. Both goals cannot be simultaneously prioritized (or maximized): an organism cannot both feed and flee at the same time. Once the lion approaches closely enough, the goals will enter into competition with each other, and each goal underdetermines the organism's behavior.

The state involves a *symmetry* between evolutionary goals. 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,* in the following sense*:* the state of the environment surrounding the token organism cannot be subsumed as one of the states of the selective environment. 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.

The subsequent action of the organism, whatever that may be (the gazelle fleeing, or continuing to graze), involves a *breaking* of this symmetry. A particular goal gains the upper hand, and this is translated into action. How precisely the symmetry is broken will depend on the particular situation, but natural selection may not have shaped a cognitive mechanism with step-by-step instructions to resolve the symmetry. The *organism itself* is causally responsible for the breaking of the symmetry. This is what self-causation means.

To illustrate why symmetry breaking has the structure of deliberation, let us go through the three characteristics of deliberation. First, the organism is blind or "unbiased" towards its different goals. This means that one cannot judge in general, averaging across all the types of environment the organism may be exposed to, that feeding or fleeing predators is "more important". They are both important. Only in specific circumstances, does one goal become more pressing than another.

Second, the organism is "weighing" the various goals. This need not be a mental operation: the "weighing" simply refers to a causal process of *competing* selected functions (or in general, competing externally designed goals). Each of the competing mechanisms is activated to different degrees by sensory input; the mechanism that is most activated will be the one that weighs most heavily on the response. There is nothing causally mysterious here, and such "weighing" may be present in organisms without a nervous system and without any cognition in the usual senses of the term.

Third, symmetry breaking involves a decision: as the organism acts, the symmetry is broken. One goal gains the upper hand, or else a course of action is chosen that attains various goals to varying extents.

5.2 The Skeptical Case against Deliberation

Symmetry breaking is such a widely applicable notion that it raises the obvious worry that some clearly non-agential systems will be counted as agential. This in fact a type of challenge that *any* broad-sense concept of agency must face: once self-causation is specified in terms of constituent causal processes, many systems will seem to be possess such "self-causation". The teleological (cybernetic) approach faces this challenge in the counterexample of thermostats, since they too interact with their goal through negative feedback.

The balancing rock. The deliberation model's equivalent counterexample is the rock balancing on a precipice, but eventually tipping over towards one side rather than another. The rock has two goals (falling to the left, and falling to the right) that are in competition, and once the wind picks up, a "decision" is made. Calling this "agential" is clearly undermotivated. There is no self-causation: 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. Thus, the skeptical case against the deliberative model states, quite simply, that if deliberating organisms are agential, so are balancing rocks, and this is a reductio ad absurdum.

The reason why the deliberation model can deal with this counterexample is because its epistemic component. An organism is deemed an agent when an agential explanation of a behavior is deemed better *in contrast to* a selectionist explanation. Agency (and self-causation) is a design principle, and must be evaluated as such, in

contrast to other design principles. The reason why the rock cannot be attributed "self-causation" is not because its behavior is causally set in motion by the wind, but because its behavior has been "pre-decided" by the laws of Newton, which determine the acceleration of the rock in response to external force. Similarly, whether or not the gazelle can be attributed self-causation does not depend on the question whether its behavior was causally set in motion by the lion or not. Rather, the question is whether its behavior has been pre-decided by natural selection.

Indirectly touching on the free will debate (O'Connor & Franklin, 2022), agency-as-deliberation can not only be categorized as compatibilist with causal determinism, but even as *in*compatible with causal *in*determinism. 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".

This is why the deliberative capacity can *itself* be the object of natural selection, but without natural selection being able to account for just how the deliberative capacity maps inputs onto behavioral outputs. In such a case, it is justified to say that the organism itself – and not some external design principle -- is responsible for the action. Metaphorically, agency can be viewed as a way for natural selection to *decentralize* decision-making, and to *outsource* decision-making to the organism itself: it is pre-decided that the organism will decide for itself what to do.

Thus, balancing rock counterexample identifies a form of symmetry breaking that simply consists of competing causal forces, and not "deliberation" in any plausible

sense of the term. The response lies in pointing to the epistemic context. Symmetry breaking can only be defined relative to the laws governing the behavior of the system. Since organismic agency, in the broad sense, means the capacity for self-causation, symmetry breaking in organisms is defined relative to *external* design principles that pre-decide an organism's behavior. This is why a rock cannot be judged to be an agent, and an organism exhibiting symmetry breaking can be judged to be an agent.

The automaton. The second class of counterexample is the automaton (or: machine, computer, robot, artificial intelligence, and so on). The automaton, at least as defined here, processes information about the environment according to "rules" that were pre-decided by external designing principles such as natural selection. It can mimic the deliberation of an agent, but it is not itself causally responsible for the action, having been pre-decided.

For instance, 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 symmetry is closely related to what Kim Sterelny has termed "epistemic action" (Sterelny, 2003, chap. 2): an animal may realize that a single cue implies unacceptable risk of a false positive, and seek additional cues to determine a course of action. The gazelle may possess a cognitive program, shaped by natural selection, that is activated in certain states of indecision, and that generates behavioral outputs that minimize false positives. In this way, the gazelle can act while following pre-decided rules.

This counterexample identifies *pseudo-deliberation,* and cannot be dismissed as resting on a confusion. Moreover, it *should* not be dismissed, because taking the possibility into consideration allows for more disciplined reflection about organismic agency. The response is thus that agents and automata can only be distinguished on a case-by-case basis. In fact, hoping for universally applicable rules (or criteria) on how to distinguish agents from automata implies a misunderstanding of what agency is about.

Agency is not a concept that picks out objects in the world, like the concept of "cat" or "mat". It does not generate judgment-free dividing lines that run through the biosphere, separating agents from the non-agents. Agency is an explanatory concept used to explain a *token behavior* of a *token organism*. Only derivatively can we use it to characterise types of behavior, or types of organism, or organisms as a whole. Thus,

organisms are agential in some respects, and non-agential in other respects. 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. However, even here, one could redescribe the explanandum to make it appear agential. For instance, one could ask why I am feeling hungry *instead of* fleeing to preserve my life. 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 action results from genuine or pseudo deliberation 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 brings the deliberation model to the meta-level: 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 specific epistemic contexts.

What a conceptual-philosophical account of agency can hope to achieve is to identify certain lines of reasoning that shape the deliberation underlying agency attribution. What types of evidence *tend* to support agential explanations? How should conflicting sources of evidence be weighed? The important question to ask is, not how 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

These questions cannot be answered in full here. 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. Nonetheless, as a way to illustrate how the concept of deliberation can help reasoning about organism behavior, 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, or whether natural selection can potentially be sufficient to explain them.

6.1 Mate Choice

A few peacocks are widely preferred by peahens; most are universally ignored or rejected. This empirical literature on charting peahen preference is surprisingly complicated (this subsection draws on (Desmond, forthcoming)). There seem to be 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. There may be other sensory cues that peahens are picking up on. However, evidently some combinations of audiovisual cues lead to acceptance (copulation), and and others lead to rejection?

The "behavioral function" here is a preference structure, mapping sensory inputs onto just two behaviors (copulation and non-copulation) indicating a preference (acceptance and rejection). The question is what best accounts for the observed preferences. An agential explanation would model the preferences as resulting from a deliberation: the peahen weighs the various traits of potential mates,

and makes an all-things-considered judgment. A selectionist explanation would model preferences as an output of a functional cognitive mechanism triggered by certain types of input. In the latter explanation, peahen choice is a mere "as if" choice, since it has been pre-decided by natural selection under what conditions the peahen will copulate.

There is no dearth of theoretical hypotheses of how peahen preferences may have evolved by natural selection. The most prominent one has perhaps been the handicap principle (Zahavi, 1975), where the 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'*.

However, do the empirical data support selectionist hypotheses? 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 (see Figure 2). Second, the presence of the blue-green eye-color is necessary but not sufficient for acceptance. Third, eyespot number was inversely correlated with heterophil level (a type of white blood cell), indicating immune effectiveness or lower rate of infection (Loyau, Saint Jalme, et al., 2005). These observations support selectionist hypothesis such as the handicap principle: peahens may possess some cognitive program that is sensitive to particular cues (minimal eyespot number, blue-green eye coloration) because these are correlated with immune system health.

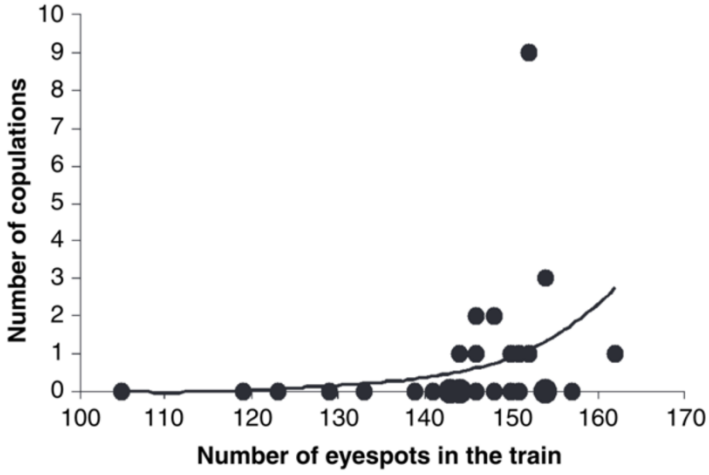

Figure 2: Peahens seem to universally reject peacocks under a minimal eyespot number. However, it is not known what drives acceptance above that minimal eyespot number. One peacock in this study copulated nine times,

but it is not known why the peacock seemed to be widely preferred. Reproduced with permission from (Loyau, Jalme, et al., 2005).

However, the literature also yields 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 predict peahen acceptance (Loyau, Jalme, et al., 2005). For instance, normality qua eyespot number is a necessary, but not sufficient cause for acceptance (see also Figure 2). Multiple signals are being evaluated simultaneously.

In and of itself, this does not imply that the peahens are genuinely deliberating on these signals. There are potential fitness advantages of integrating multiple signals, since it may allows to more powerfully discriminate between fitness components of peacocks (Choi et al., 2022). This raises the hypothetical possibility that particular combinations of audiovisual cues are activating inherited cognitive mechanisms, shaped by natural selection, that trigger acceptance. In other words, the peahen could conceivably be an automaton rather than an agent. However, the studies that point to the fitness value of signaling multiple cues, do not identify such combinations. Rather, the fitness advantage seems in allowing the peahen to integrate *more* information about the displaying peacocks. In other words: multiple cues aids deliberation. If true, this is the type of evidence that would support the inference that peahen choice has not been "pre-decided" by natural selection, but are conducting their own deliberation.

6.2 Bacterial Locomotion

The second case study concerns locomotion, one of the most basic ecological interactions between organism and the external environment. Is it an example of agency? Let us consider chemotaxis as one of the most basic forms of locomotion. The teleological approach would here deem chemotaxis to be a manifestation of organismic agency because it directs the organism towards the goal of being close to the source of nutrition. However, there is a straightforward selectionist-mechanistic explanation of this goal-directedness: chemotaxis is a mechanism that takes certain sensory cues as inputs, produces motor outputs, and has evolved multiple times since it is adaptive in heterogeneous environments (Keegstra et al., 2022).

To apply the deliberation model to chemotaxis, we must look at the empirical details of real chemotaxis behavior. Can the behavior function of chemotaxis really be modelled as taking in one type of sensory input (a nutrient gradient) and generating one type of behavioral output (swimming up the gradients)? It turns out the mapping of sensory inputs on behavioral outputs is not simply one-to-one. Distilling three studies (Ortega et al., 2017; Taylor et al., 1999; Yamamoto et al., 1990), Table 1 gives a non-exhaustive summary of the "preference structure" of *E. coli* with some 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 its own dedicated mechanism (i.e., a receptor). (By contrast, there is no evidence that different audiovisual cues are processed in different spatially distinct "modules" in the peahen brain.) Of these mechanisms, Tar and Tsr are the most abundant in the *E. coli* membrane, and are sensitive to two of the most important sensory inputs – aspartate and serine level. The other chemoreceptors (Tap, Trg) are much less abundant in the periplasm, but 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,* summarizing (Ortega et al., 2017; Taylor et al., 1999; Yamamoto et al., 1990).

Sometimes these signals compete, and when they do, bacteria exhibit "pausing behavior" (Eisenbach et al., 1990). And when bacteria move in a particular direction, it is not necessarily nutritious compounds that trigger this attraction. 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 (see Figure 3). Hence, 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).

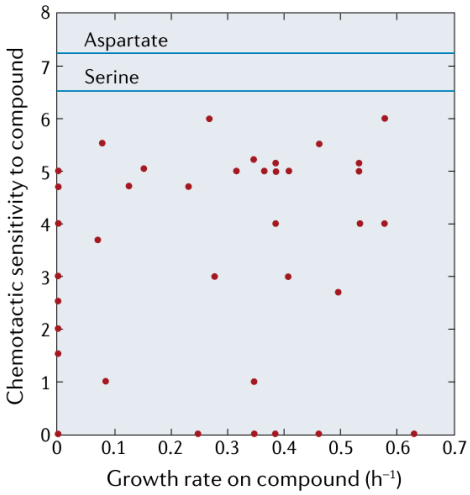

Figure 3: The relation between the attractiveness and nutritiousness of a compound is surprisingly noisy. Reproduced from (Keegstra et al., 2022, p. 493)

Another complicating consideration is that different receptors form *arrays*. This means that receptors do not simply "compete" to determine the whole organism's behavior , 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 are array structures – the precise frequency and location of receptor types – 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.

These empirical details are crucial for reflecting on whether bacterial chemotaxis (of the *E. coli* in this case) should be considered as an agential behavior. According to the deliberation model, the *ideal types* of evidence would be evidence indicating that bacteria can be in states of "symmetry" with regards to the (ecological) goals shaped by natural selection, and evidence that the way in which this symmetry

is broken does not seem to be pre-decided by natural selection. In other words, the uncertainty faced by the token bacteria must be subsumable under a *type* of uncertainty that recurs in the selective environment. Then it would be plausible to believe that bacteria behavior can be adequately predicted by a population-level selected function. However, if there is no evidence that this is possible, then this is grounds for attributing agency.

To what extent does the real, available evidence fit this ideal? The pausing behavior of bacteria indicates symmetry between goals. However, the difficult question is whether the mode of resolution is pre-decided by natural selection. For this one would need detailed empirical studies of the precise microconditions in the bacterial environment that trigger specific decisions by individual bacteria: in other words, the analog of studies on peahen behavior. Research on this is ongoing and driven by the use of microfluidic devices which can control environments on the scale of micrometers (Chait et al., 2017; Hochstetter et al., 2015). If it turns out that the *noisiness* of individual bacterial motion can only be explained by referring to very specific, idiosyncratic circumstances – the way a judge's exact sentencing can only be understood in reference to the specific, idiosyncratic circumstances of a case – then this would be grounds for viewing bacterial chemotaxis as agential.

7. Conclusion

Agency-as-deliberation defines agency, in effect, as a selective process going on at the level of the organism. The organism, and not natural selection, is selecting what behavior is best, by weighing and selecting a course of action in light of its various competing goals. The deliberation model mainly 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 implied parallel between natural selection and agency is a fundamental attraction of the deliberation model as proposed here. Looking forward, it suggests a conceptual framework where 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.

Word count: 9903

(including footnotes but excluding references, abstract, and captions)

Acknowledgments

Blinded.

REFERENCES

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/

Brandon, R. N. (1990). *Adaptation and Environment*. Princeton University Press.

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.

Chait, R., Ruess, J., Bergmiller, T., Tkačik, G., & Guet, C. C. (2017). Shaping bacterial population behavior through computer-interfaced control of individual cells. *Nature Communications*, *8*(1), 1535. https://doi.org/10.1038/s41467-017-01683-1

Choi, N., Adams, M., Fowler-Finn, K., Knowlton, E., Rosenthal, M., Rundus, A., Santer, R. D., Wilgers, D., & Hebets, E. A. (2022). Increased signal complexity is associated with increased mating success. *Biology Letters*, *18*(5), 20220052. https://doi.org/10.1098/rsbl.2022.0052

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

Desmond, H. (2022). Adapting to Environmental Heterogeneity: Selection and Radiation. *Biological Theory*, *17*, 80–93. https://doi.org/10.1007/s13752-021-00373-y

Desmond, H. (forthcoming). Sexual Selection, Aesthetic Choice, and Agency. In E. Gayon, P. Huneman, V. Petit, & M. Veuille (Eds.), *150 Years of the Descent of Man*. Routledge. https://philpapers.org/rec/DESSSA-3

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

Friston, K. (2012). Prediction, perception and agency. *International Journal of Psychophysiology*, *83*(2), 248–252. https://doi.org/10.1016/j.ijpsycho.2011.11.014

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

Hamkins, J. D. (2020). *Lectures on the philosophy of mathematics*. The MIT Press.

Hochstetter, A., Stellamanns, E., Deshpande, S., Uppaluri, S., Engstler, M., & Pfohl, T. (2015). Microfluidics-based single cell analysis reveals drug-dependent motility changes in trypanosomes. *Lab on a Chip*, *15*(8), 1961–1968. https://doi.org/10.1039/C5LC00124B

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

Loyau, A., Jalme, M. S., & Sorci, G. (2005). Intra- and Intersexual Selection for Multiple Traits in the Peacock (Pavo cristatus). *Ethology*, *111*(9), 810–820. https://doi.org/10.1111/j.1439-0310.2005.01091.x

Loyau, A., Saint Jalme, M., Cagniant, C., & Sorci, G. (2005). Multiple sexual advertisements honestly reflect health status in peacocks (Pavo cristatus). *Behavioral Ecology and Sociobiology*, *58*(6), 552–557. https://doi.org/10.1007/s00265-005-0958-y

Madani, O., Hanks, S., & Condon, A. (1999). On the Undecidability of Probabilistic Planning and Infinite-Horizon Partially Observable Markov Decision Problems. *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 541–548.

Mayr, E. (1961). Cause and Effect in Biology. *Science, New Series*, *134*(3489), 1501–1506. http://www.jstor.org/stable/1707986

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

O'Connor, T., & Franklin, C. (2022). Free Will. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2022/entries/freewill/

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.

Pittendrigh, C. S. (1958). Adaptation, natural selection, and behavior. In A. Roe & G. G. Simpson (Eds.), *Behavior and Evolution* (pp. 390–416). Yale 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

Sober, E. (1984). *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. University of Chicago Press.

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. (2000). Chasing shadows: Natural selection and adaptation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *31*(1), 135–153. https://doi.org/10.1016/S1369-8486(99)00041-2

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)

Wright, L. (1973). Functions. *The Philosophical Review*, *82*(2), 139–168.

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