Naturalistic Foundations of Thompson's Theory of Agency

Wing Yi So

UCL

A thesis presented for the degree of
Master of Philosophy

I, Wing Yi So, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The present work is concerned with Thompson's Aristotelian approach to uncovering the minimal conceptual structure underlying attributions of action (X is doing/did A) and explanations of action (X is doing/did A because...). Aristotelian approaches, which conceive of action attribution as involving a relation between the act/event and the form/essence of the agent X, and which explicitly holds on to a teleological conception of action explanation, have been resurgent in contemporary philosophy. Such attempts however face a schematic dilemma: contemporary philosophers with a broadly naturalistic orientation often find the metaphysics involving teleology and form problematic, where teleological explanations are in particular difficult to reconcile with current conceptions of naturalistic and causal explanation; yet extricating Aristotle's insights from the metaphysics often leads to watered down claims that cannot do the work originally envisioned.

Michael Thompson has proposed a theory that can potentially overcome this difficulty. I shall present Thompson's proposal, and argue that, as it stands, it still does not escape the fundamental dilemma. However, as will be shown, one can view Thompson's ideas from a naturalistic point of view. Borrowing from systems theory, I show that the Aristotelian idea of form can be understood as designating systems with a particular physical organisation (displaying 'closure of constraints'). Further, I claim one can study in this naturalistic framework the deep conceptual structure underlying action attributions and action explanations that Thompson proposed through developing a theory of affordances. As I hope to show, this gives one a precise idea of the relation between teleological explanations and naturalistic, ausal explanations.

Impact Statement

The present work attempts to examine a position concerning the nature of agents and of action proposed by Michael Thompson. The aim is to study this in the light of contemporary theories of biological individual. It is hoped that this provides a new perspective on issues in the philosophy of action, and can lead to new approaches.

Acknowledgements

I would like to thank Doug Lavin for initial guidance on formulating the project. I would also like to thank John Hyman for supervision of the project at a later stage.

Contents

Chapter 1

Introduction

Philosophers of action have long been interested in studying the nature of actions and agents. What does it mean when one attributes an action to an agent (agent X is doing/did A), or when one explains an action (X is doing/did A because...)? What is distinctive of actions or activities, in contrast to just any other event or process; and what is distinctive of agents capable of action? What is the meaning of 'because' in an action explanation? The aim of the present work is to explore an Aristotelian approach to these problems proposed by Michael Thompson; more precisely, the aim is to provide the naturalistic foundations of parts of Thompson's theory. This first chapter will set the stage by presenting the background ideas.

1.1 Aristotle's View

The best way to introduce Thompson's position is to start with Aristotle's reflections on agency. Aristotle's inquiry into the nature of agency takes place against the broader background of an inquiry into the nature and explanation of change. For Aristotle, there are privileged entities to consider when investigating change, which he lists as animals and their parts, plants, and the simple bodies earth, fire, air, water. These entities are privileged in that "each has in itself a source of change and staying unchanged, whether in respect of place, or growth and decay, or alteration. A bed, on the other hand, or a coat, or anything else of that sort, considered as satisfying such a description, and in so far as it is the outcome of art, has no innate tendency to change, though considered as concurrently made of stone or earth or a mixture of the two, and in so far as it is such, it has." (192b14-20) For Aristotle, these internal sources include both principles which explain the change, as well as that which actually makes the change happen. Thus, such internal sources include the famous 'four causes': the matter, the form (which can be understood as the essence), the primary source of the change or rest (later known as efficient cause), and the end or that for the sake of which (194b24-195a2).

The significance of this general physical background for the inquiry into the nature of agency becomes clear starting with Book III of the *Physics*, where the pair action/passion is discussed. The pair action/passion is subsumed by the pair productive of change/changeable ("the active and the passive" is to be understood through "that which is productive of change and that which is changeable" (200b30-32)). Thus, the agent is identified with that which produces change, and moreover "[t]hat which produces change will always carry some form, either 'this' or 'of such a kind' or 'so much', which will be the principle of, and responsible for, the change, when it produces change." (202a9-12) Particularly noteworthy from the present perspective is the central explanatory connection between the nature of an entity, its essence or form, and change: change is partly explained through the nature of an entity, the kind of thing the entity is (thus moving upwards belongs to and is according to the nature of fire 192b35-193a1).

Let us consider the important case of animate bodies—plants, animals and humans. Here, the form they possess is their soul. In fact, as Shields noted in his commentary on *De Anima*, the form here is simultaneously the formal, final and efficient cause of the changes of the animal (Aristotle, 2020):

> For all ensouled bodies are organs of the soul—just as it is for the bodies of animals, so is it for the bodies of plants—since they are for the sake of the soul. 'That for the sake of which' is spoken of in two ways: that on account of which and that for which. Moreover, the soul is also that from which motion in respect of place first arises, though this capacity does not belong to all living things. (415b15-20)

In the same way, the soul is the cause of alteration, growth and decay.

Let us summarise the key Aristotelian ideas concerning actions of an agent and the explanation of action.

1. An agent is that which produces change and has in itself an internal source of change. Among such internal sources is the essence or form of the agent, the kind of thing that the agent is. For animate beings in particular, this form is identified with the soul. Actions are the changes produced by an agent on another entity (which may be itself) on the basis of its form. This connection between action and internal source of change accounts for the production of action as well as its explanation:

2. First, since the end could form part of the 'internal sources of change', the Aristotelian conception licenses teleological explanation of action: the action is done for the sake of an end, and can be explained by it

3. Second, explanation through the form of the agent: the agent produces such a change because of its nature

4. Finally, it supports an efficient causal explanation. Having such a soul should, somehow, bring about the change in question. Unfortunately, little is said about how this could be the case.

As remarked before, these are not always distinct for Aristotle, since the soul for instance is both the final, formal and efficient cause. In this case, the explanations all proceed via an appeal to the nature and structure of the soul.

From this brief presentation it is clear that the concepts of action and agency belong to the study of nature. But at the same time, the concept of action can support an ethical theory. This is because Aristotle's view of nature allows him to introduce the concept of an intrinsic function. Let us focus on the crucial case of the soul. By its form or essence, the ensouled agent has to act for the sake of an end, an end that thus belongs to the form of the agent. The existence of such ends belonging to the form of an agent allows one to define intrinsic functions for the agent X: X (or some trait of X) has an intrinsic function to F if exercising the function F belongs to the form of the agent as an end for X. The concept of an intrinsic function, in turn, is the basis for the idea of the good: if X has an intrinsic function to do F, X can be evaluated as good or bad relative to whether and how it performs F. This is even the case for the Supreme Good:

> To say however that the Supreme Good is happiness will probably appear a truism; we still require a more explicit account of

> what constitutes happiness. Perhaps then we may arrive at this by ascertaining what is man's function. For the goodness or efficiency of a flute-player or sculptor or craftsman of any sort, and in general of anybody who has some function or business to perform, is thought to reside in that function; and similarly it may be held that the good of man resides in the function of man, if he has a function.

The foundation of the Supreme Good for man in particular is placed in the function of man; this function, in turn, as Aristotle argues, is to be sought in the rational part of the soul or the form of man. Thus one might add a final point to the Aristotelian picture:

5. A concept of good that can be associated to each being which has a form with essential ends

Given the systematic coherence of these claims and the philosophical work they could do, it is unsurprising that these ideas came to be extremely influential (one can consult, for instance, Suarez's *Disputatio Metaphysica* 48 as well as his treatment of efficient causality for an overview of different views on action/passion and causality in many prominent authors in history; see also the textbooks of Eustache de Saint Paul (1609) and Dupleix (1626), in use in Descartes' time). Yet, with transformations in the conception of nature wrought by the early moderns, the central concept of 'form' came under attack (the rejection of forms in the crucial case of Descartes is treated in great detail in e.g., (Hattab, 2009; Ruler, 1995); transformations concerning the concept of animacy is detailed in (Riskin, 2016)). The result of this overturning of forms is as decisive as a philosophical attack can be; any attempt to revive the Aristotelian position, and thus the concept of form, would have to do so under very different premises.

1.2 Contemporary Revivals of Aristotle and a Fundamental Difficulty

One way to revive Aristotelian ideas, explored by a number of contemporary philosophers, is achieved through curtailing the scientific ambitions of the Aristotelian ideas substantially. Under this strategy, one first largely accepts that Aristotelian forms cannot do the scientific work required of them. Thus, one extracts the project from the general inquiry into natural change. In particular, point 4 above, which links forms to efficient causality now understood from a scientific point of view (e.g., through laws of nature, or mechanisms), is given up. This, however, does not mean that one has to give up the idea that some entities possess a form or essence. Although this form does not really help explain natural change, it can license teleological explanations or explanations from the form. But if such explanations are valid, the function argument can still go through, and the Aristotelian idea of the good can be retained. As such, one can, with minimal sacrifice from the point of view of ethical theory, retain much of what is attractive in the Aristotelian framework.

One can see this move at work in the pioneering work of Anscombe. The rejection of the ambition of viewing all agency under the general framework of 'change' of natural objects is clear. In discussing intentional action, Anscombe identified the applicability of the question 'why' as a distinctive mark, where this question is asked in a very specific sense. This sense is that the answer, if positive, gives the reason for acting (Anscombe, 2000, 9). Now Anscombe does not find this formulation by itself explanatory, and towards the end of her inquiry she writes "intentional has reference to a *form* of description of events.... Events are typically described in this form when 'in order to' or 'because' (in one sense) is attached to their description." (Anscombe, 2000, 85) Such a form of description is explicitly said to "go beyond physics", and are "vital descriptions" (Anscombe, 2000, 86). Bringing an event under such a form of description explains the event by rationalising it through its telos, an explicitly teleological mode of explanation somehow related to 'vitality' or animation.

Anscombe does not really further explain the concept of 'vital description' in *Intention*. But one may search for hints elsewhere, where it becomes apparent that a certain Aristotelian notion of form has been retained. For Anscombe defends

> the idea of a primary principle of the life of a living creature of a particular kind. I don't mean a general principle of life, but a particular primary principle of horse life, say, or strawberry life or human life. The primary principle of life, or the form, I call the soul of the living plant or animal. It may sound odd, because unusual, to speak of the soul of a daffodil. But a daffodil

> plant is certainly not inanimate, and so must have such a primary principle, its determinative form, as the principle of its being a living daffodil. (Anscombe, 2005)

One might think this is not much of an explanation. Later philosophers, such as Foot and Korsgaard, have attempted to formulate this key idea with greater clarity. They seek this determinative form, a universal, in the species of an individual which the individual participates in as a reproducing unit. Each individual can then be understood as engaging in activities to maintain its own form. This dynamic of the maintenance of form then allows one to understand the purposive (sometimes also described as imbued with intentionality or goal-directed) activities of an animal: in Korsgaard's words,

> it is because an animal has a self-maintaining form that we can assign intentional content to her movements. It is because an animal's function is the maintenance of her form that we describe even a very primitive animal "looking for something to eat" or "trying to escape the danger," in a way that implies criteria of success or failure.

The connection between the essential kind of thing an individual is and its activities licenses an attribution of goodness. Foot (2001, 33-34) articulates the chain of reasoning as follows:

> (a) There was the life cycle, which ...consisted roughly of self-maintenance and reproduction.
>
> (b) There was the set of propositions saying how for a certain species this was achieved: how nourishment was obtained, how development took place, what defences were available, and how reproduction was secured.
>
> (c) From all this, norms were derived, requiring, for instance, a certain degree of swiftness in the deer, night vision in the owl, and cooperative hunting in the wolf.
>
> (d) By the application of these norms to an individual member of the relevant species it (this individual) was judged to be as it should be or, by contrast, to a lesser or greater degree defective in a certain respect.

Very similar ideas are expressed in Korsgaard (2009, chapter 2) or Korsgaard (2018, 2). Of course, such philosophers do not only want to establish some kind of biological normativity. They wish to establish, through this type of argumentation, a substantive notion of goodness for humans (as Aristotle did): thus Foot (2001) claims that there exists 'natural goodness' of humans; in the same way, Korsgaard argues that the normative standards which follow from the form of a human being as a rational agent should rule out 'evil' and other defective forms of agency (Korsgaard, 2009, chapter 8).

Despite the appeal of these ideas, the suppression of the connection to efficient causation in the Aristotelian framework, and more generally the attempt to extricate Aristotle's ideas on ethics and action from the metaphysical and physical background, cause deep problems. These problems are manifestations of a basic dilemma. Contemporary philosophers with a broadly naturalistic orientation generally find the metaphysics and physics of teleology and forms objectionable; yet attempts to extract Aristotle's ideas on ethics and action from the metaphysics and physics threaten to bring only watered down versions of Aristotelian concepts that cannot do the work originally envisioned. This dilemma appears sharply in the severing of the connection of the Aristotelian framework to efficient causality. This can be seen in three ways.

First, these Aristotelian philosophers appeal to concepts with efficient causal implications (such as reproduction and self-maintenance) but do not engage in any detailed investigation of such concepts. As such, it is not clear that these concepts can do the required work. This lack of clarity concerning the naturalistic basis of the concepts can undercut the project in various ways. Here I only mention two possible cases. First, the appeal to self-maintenance and reproduction seems insufficiently restrictive. Different forms of organisation display self-maintenance, reproduction, or both. Hurricanes and vortices in a fluid may be said to be self-maintaining; forms of artificial life can reproduce, or both reproduce and self-maintain. A notion of 'goodness' or functional normativity might be definable for such systems, but it seems highly doubtful a substantive notion of good can be built on such a basis without a clarification of the difference between such systems and the systems the philosophers above have in mind, which usually are human beings. Second, although the philosophers appeal to a concept of species for introducing norms, it is not clear the modern concept of species can do this work. For instance, according to an influential view, a species is an individual (see for instance Ghiselin (1997); for a survey see (Zachos, 2016)), not a universal. Further, there are many more candidates for collectives or organisations that seem to be capable of playing the role of giving norms. Reflections on these collectives, including ‘species’, quickly lead to doubts concerning the naturalistic legitimacy of constructs such as the ‘form’ of the human being or of human nature (for skepticism concerning ‘human nature’, see e.g., contributions in Hannon and Lewens (2018)).

Second, these particular worries can be generalised to lead to a broader skeptical worry: what is the basis for our postulation that the an agent possesses such and such a form (e.g., motivation through reason and self-conscious reflection,...), and that its actions can be explained teleologically or through the form? Can it not be that our teleological descriptions and explanations are only misguided posits effacing the real causes which do not follow the grammar of ‘vital descriptions’?

Finally, as the causalists like Davidson have pressed in the context of intentional action, it seems that the concept of action should have an efficient causal dimension. An intentional action is not only an action open to some teleological explanation (explainable through reasons), but it is an action that is done for a particular reason. Clearly, there are many reasons that explain or rationalise an action but which are not the reason for which an act is done (an idea first introduced in (Davidson, 1963); for further discussions see e.g., (Mele, 2010), (Sehon, 2016)). As Davidson goes on to argue, that which distinguishes the reason for which an act is done from all other reasons is that the former reason is the cause of the intentional action. The point generalises to actions in general. The explanation of an action aims to explain why this particular action happened on this particular occasion. Introducing ends that an action may serve does not suffice unless the connection between the existence of the end and the way in which the action is brought about on this occasion is clarified.[1]

1.3 Thompson’s Contribution

None of these difficulties are fatal. They simply indicate that the Aristotelian revivals remain incomplete. Michael Thompson has over the years devel-

[1] Indeed, the efficient causal element has been seen as fundamental to action in general in the Aristotelian tradition. This is the view of Suarez and a number of scholastics (Suarez Disputatio Metaphysica 48: “actio [est] causalitas causae efficientis”), a point he argued for in greater detail in his treatment of efficient causality (DM 18)).

oped a version of the Aristotelian theory which has the potential to overcome the aforementioned difficulties. He does so by following what might be termed a conceptualist strategy: understanding the nature of agents and agency through understanding the circle of interconnected concepts we employ in using the concepts of action and agent, as well as the judgements through which such concepts are connected. Thompson's conceptualist strategy makes two further commitments. First, the relevant concepts and judgements are claimed to form part of the 'manifest image of the world'—the everyday mode of conceptualising the world which crucially exludes the introduction of new fundamental theoretical posits to explain the appearances (Sellars, 1963). Second, these concepts and judgements (such as the agent as life-form, or judgements concerning their life-cycle) are said to have a fundamentally different 'logical form' than concepts and judgements in the scientific image (presumably expressible in first or higher order predicate logic).

Nevertheless, these concepts are not supposed to be entirely disjoint from concepts describing nature as physical processes and events. Employing an old Aristotelian imagery, Thompson conceives of the relation between physical concepts involving processes and events (and perhaps certain conceptions of causality, although Thompson does not mention it) and concepts involving action and subjects as a relation between two distinct conceptual strata. In this sense, concepts of the higher strata should be seen as enrichments of the concepts of a lower strata, and the study of concepts of a higher strata can be conducted relatively independently of the concepts of a lower strata.

More concretely, Thompson inflects the Aristotelian claims above as follows. Following his conceptualist strategy, his fundamental problem is to understand the deep conceptual structure ('logical form') underlying action attributions (X is doing/did A) and action explanations (X is doing/did A because...) in everyday life. The crucial concept of form in Aristotelian theory is approached through unearthing the nature of the subject implicitly assumed in our everyday concepts and judgements about agents and their actions. This implicit conception of the subject Thompson claims to discern in a special class of concepts—life-form concepts. The specific features of these concepts are in turn determined through examining the judgements they enter into, which Thompson call natural-historical judgements. These judgements have two features—generality and temporality. Roughly, generality designates the fact that the activities of life-forms can be described through generics (e.g., bobcats breed in spring. . .) which are true in virtue of the nature or ‘form’ of the life-form (e.g., the species), while temporality designates the fact that the events in the life of a life-form display a distinctive temporal coherence that cannot be observed in inanimate physical process (e.g., creation of chemical elements in stars).

The language for the description of nature which conceives of nature as consisting of events and processes is enriched by our description and explanation of actions of life-forms in the following sense. Attributions of actions places an event, the action, in a new context—that of life-form concepts and thus events within the life and the life-form. Thus located, the event gains a new significance. It is on this basis that Thompson claims that all teleological action explanations have a minimal form: they make apparent the way in which events are located in the temporally coherent life of a life-form possessing a form; the force of the because in such an explanation then derives from the temporal coherence of the life of a life-form.

If Thompson is correct, he could claim to have sidestepped the dilemma above. Teleological explanation and the concept of form can now be introduced as irreducible concepts that belong to a higher strata but which remain enrichments of concepts of physical events and processes. This should account for how efficient causality at the lower strata meshes with teleological explanations on the higher strata and avoid worries cited previously. The way in which Thompson envisions his theory to be extended to social practices and dispositions of agents that arise from such practices would lead to a sufficiently sophisticated framework adequate to treating notions of goodness that are not too simplistic and limited to biological levels (e.g., treatment of virtues such as justice). All this depends on whether Thompson succeeds in providing an adequate analysis of action and agents through his conceptualist strategy, and whether his notion of ‘enrichment’ can do the required work. The examination of Thompson’s proposal constitutes the starting point of the present inquiry.

1.4 Summary of the Dissertation

In the next chapter, Thompson’s position will be presented in greater detail. I shall argue that as it stands, Thompson’s position does not allow one to sidestep the fundamental dilemma. Indeed, the three problems mentioned above will return in a different form. However, I shall argue that one need not curtail the scientific ambitions of the Aristotelian theory. In particular, I claim that temporality and generality can be understood equivalently as requirements on the forms of physical organisation that a life-form should have (chapter 2).

Now one might wonder what I mean here by a naturalistic point of view. By this phrase I do not assume at the outset any particular metaphysical doctrine about what there is, or about the reducibility of all concepts/laws to a certain set of (physical) concepts/laws; a very rough understanding will suffice for present purposes. Naturalism here designates something close to what Maddy has called second philosophy. In second philosophy, the philosopher starts out as a native to the modern laboratory, familiar with the going concern of the scientific community. The way in which one arrives at any conclusion has to be through the scientific method. There is a lot of controversy concerning what the scientific method is, but for present purposes, a rough and ready account suffices:

> an empirical study, beginning with careful observation of phenomena ('natural history'), moving on to deliberate experimentation, theory formation, and testing ('natural philosophy' in the stricter sense), always assessing and re-assessing its methods as it goes—a mode of inquiry that makes no principled distinction between questions we tend to think of as 'scientific' and questions we tend to think of as 'philosophical'. (Maddy, 2022, 19)

By a naturalistic perspective, then, I mean nothing other than an attempt to make intelligible certain phenomena (here, the agency of individuals and features of such agency such as purposiveness) through concepts and laws that form part of current scientific theories in a methodical manner.

Adopting such a naturalistic perspective means one cannot proceed according to a conceptualist strategy which takes the concepts and habits of thought in the 'manifest image' at face value. Rather, one uses the manifest image as a guide for the physical basis of purposive agency. What actually transpires in this inquiry has priority over the syntactic forms of ordinary life. As I hope to show in chapter 3, by borrowing from contemporary research into complex systems, one can largely provide an adequate account of Thompson's notion of generality and temporality, and thus of the concept of life-form, as Thompson envisions it (there are some caveats). I shall next argue that this leads to a more precise view about the deep conceptual structure of action attribution and explanation. First, attribution of actions can now be understood on the lines of attribution of functions and abilities; we shall develop an account of action through a theory of affordances. Second, action explanations do explain by locating actions within the life of a life-form, but one can attain a much more precise characterisation through the framework of affordances to be developed; in particular, we clarify the nature of teleological explanations as a form of scientific explanation, and elucidate the connection of such explanations to causal explanations (chapter 4).

The resulting work will not be a full examination of all aspects of Thompson's theory. In particular, arguably the most important concern—ethical life of human beings—will have to be neglected. However, it is hoped that the work does provide the naturalistic foundations for such a discussion so that more complex forms of life and action can ultimately be understood on this basis.

Chapter 2

The Agent as a Life-Form

In this chapter, we examine Thompson's theory of agency. His theory may be divided into three parts: first, a theory of the agent to which actions are attributed; second, an account of actions and action explanations which builds on this theory of agents; finally, an account of practices and habitual dispositions that support a theory of virtue. It is this final level which is for Thompson the proper context in which human actions can be located.

The aim of the present chapter is to explain and examine the first two parts of Thompson's theory. This will suffice for understanding the essential structure of actions and agents on which more complex human actions build on—the purposiveness of actions and the fact that they can be explained teleologically. I shall first study Thompson's theory of the agent (section 2.1), and then move on to his theory of action explanations (section 2.2). After presenting Thompson's analyses, I will argue that they lead to difficulties (section 2.3), which can be seen as concrete versions of the difficulties mentioned in 1.2. However, once one gives up on some of the background anti-naturalistic philosophical assumptions of Thompson's work, his observations are in many ways consonant with natural philosophical observations. In particular, they can be used as a guide for understanding the nature of agents and the source of purposiveness (2.4).

2.1 Life-Form Concepts and Natural-Historical Judgements

As noted in the previous chapter, Thompson pursues a conceptualist approach that aims to clarify the structure of actions through examining the relevant concepts and judgements we employ in everyday life (the ‘manifest image’ (Thompson, 2008, 10, 200)), instead of attempting to find a ‘real definition’ of agents in biological or chemical terms; indeed, the first chapters of Thompson (2008) are devoted to arguing against this possibility. In this strategy, the agent is conceived through the ‘life-form concept’; this concept in turn is clarified through understanding the judgements they enter into, the natural-historical judgements. Thus, the key lies in clarifying the structure of such natural-historical judgements.

Natural-historical judgements can be seen as a reconceptualisation of Anscombe’s vital descriptions. It is best illustrated through examples. The paradigmatic example of a natural-historical judgement comes from nature documentaries concerning animals, say the polypedilium vanderplanki (Wharton, 2002). The narrator might begin by narrating its life cycle: “the polypedilium vanderplanki lives in the rain-filled rock pools in Africa. It is born as an egg, develops into a larva, next into a pupa which then transforms into the adult.” The narrator can go on to describe its characteristic features: “its habitat experiences wet and dry seasons, as a result of which the larva is sometimes deprived of water for months. In these times the larva is capable of surviving anhydrobiotically, when all of the metabolic functions are paused reversibly.” Finally, its characteristic activities (foraging, mating, etc.) will be narrated. A judgement such as “the polypedilium vanderplanki survives anhydrobiotically when deprived of water for long periods”, and a narrative concerning the life-cycle of such a creature, are typical natural-historical judgements.

According to Thompson, these judgements possess a special logical form, which is exhibited in two key features: generality and temporality. Here, generality designates the generic character of these judgements. A judgement such as “the polypedilium vanderplanki survives anhydrobiotically when deprived of water for long periods” can at first sight be formalised through a universal quantifier. Yet this would be a mistake, for not all such larvae can survive anhydrobiotically, and the judgement could still be valid even if most cannot. In other words, it belongs to the class of sentences linguists call generics (e.g., 'water is liquid'; for more on the linguistic perspective on generics, see e.g., (Carlson and Pelletier, 1995)). In the semantics of generic sentences, it is generally accepted that one cannot model the semantic literally through a universal quantifier, but has to take into account the background or context in which the sentence is used.

However, as Thompson rightly points out, the natural-historical judgement is a special type of generic whose structure is not yet fully understood just on the basis of this linguistic classification. In particular, it seems different from generics like 'water is liquid', which is true generically only given assumptions about what constitutes 'normal' background conditions in locations containing water. Yet such conditions are not normal for water itself; they are normal for humans. The same cannot be said of the target of the natural-historical judgement: the fact that the larva can survive anhydrobiotically is 'normal' for the life-form itself; otherwise it would have gone extinct. Larva that fails to survive in this manner is unhealthy, or that its capacity for anhydrobiotic survival is malfunctioning. In other words, each individual life-form can be evaluated according to normative standards, standards that are associated with and intrinsic to each kind of life-form. It is this distinctive feature of judgements where the activities of the subject are evaluable with respect to something intrinsic to the subject of the judgement that Thompson calls generality.

Temporality arises when one attends to another feature of such narratives: in the description of the phases of development of the animal, no reference is made to the external timeline; one only describes the internal order of the phases. As Thompson puts it, one is viewing the timeline of the life-form as a B-series, not as an A-series in McTaggart's terms (Thompson, 2008, 65). Interestingly, Thompson emphasises that the propositions expressing the temporal order of the life of a life-form "has nothing to do with natural selection" in that "these propositions are in no sense hypotheses about the past." How then should one understand such temporality? This is brought out through a contrast between life-forms and technical processes and artifacts. The idea seems to be this: the production of technical objects depends on some external actor to maintain the process in a way that the life-form does not (Thompson, 2008, 79-80). However, Thompson does not provide much further illumination beyond these cursory remarks.

Let us elaborate on this contrast. Technical objects and processes display a prima facie similarity with life-forms. Just as in a nature documentary narrative, one might say the following of a technical process: "to synthesise chemical compound C, one first takes some available chemical A; then one increases the temperature by...; after that, one has to add X to produce an intermediate product Y so that we get Z...." These steps support means-end relations to each other; the success of the synthesis of the compound depends very much on the temporal order between the steps being respected. The difference between this case and that of life-forms however is that the coherence of the processes making up the life-form seems to derive from the nature of the life-form itself, whereas the coherence of the industrial process derives from an external purpose (someone has to make A available, increase temperature, add X...). There is some sense of 'internal purposiveness', to borrow Kant's phrase. By calling it 'internal', one is assuming that the life-form in question possesses some nature, or form, so that the purposiveness can be said to be intrinsic to such a nature or form.

As far as I know, Thompson does not provide any further elucidation of this obscure concept of form and what we have called internal purposiveness. A tempting move is to simply borrow Kant's characterisation of internal purposiveness, which designates a special form of organisation of the parts of a system whereby firstly the parts exist only in virtue of the whole, and secondly, each part is produced by some other part (AA5: 373-374). Thompson however rejects this. Indeed, he rejects any attempt to use the concept of organisation or complexity to capture purposiveness. Rather, generality and temporality should be taken as primitives, through which the life-form is comprehended. Perhaps Thompson's approach is closer to that of Spinoza, whose notion of 'conatus' introduced in Part III, proposition 6 and 7 of the *Ethics*, designates the effort of each thing to preserve itself in existence; this conatus simply is the essence of the thing in question. This accords fairly well with the intuition in Thompson's work, where the coherence of the life-history of a life-form should derive from the nature of the life-form, and the activities of the life-form can be understood as manifestations of this striving to preserve the life-form in existence. In this case, the 'form' of the agent would consist in the fact that the organisation of the agent can be described as self-preserving or self-maintaining. Yet Thompson also rejects such talk of 'self'-organisation. In any case, without further clarification, one may understandably object that temporality, generality, and conatus themselves are rather obscure concepts. We will return to Thompson's rejection of the concept of organisation in 2.3 where we criticise his arguments. For now, we will treat 'form' and 'internal purposiveness' as placeholders, relying on an intuitive understanding as Thompson does.

2.2 Action and Life-Forms

Having completed the preliminary survey of the concept of a life-form, we pass to actions and the explanation of actions. Recall that for Thompson judgements attributing actions and explanations of actions unveil a special form of unity between events and the subject of attribution. The simplest way in which this special form of unity comes into play is in what Thompson calls naive action explanation: "agent A is doing/did X because he is doing/did Y". Naive action explanations make no appeal to reasons, intentions, desires or other psychological and social factors; these are all further enrichments of the fundamental schema involving only a certain coherence between actions. Naive action explanation exhibits the structure underlying the unity of actions without further distractions. Let us follow Thompson's reasoning here.

2.2.1 Naive Action Explanation

The central claim Thompson makes concerning naive action explanations is that the distinctive unity which is at the basis of the teleological connection between activities A and B lies in the postulated temporal coherence or temporal unity of the activity A and the activity B. The clue for understanding this temporal coherence is the fact that such activities are described through verbs which have an aspect. Aspect in general concerns "different ways of viewing the internal temporal constitution of a situation"(Comrie, 1976, 3-5). For Thompson's purposes, the key internal temporal constitution is expressed in grammatical aspect: the contrast between the perfective and the imperfective for verbs. Very roughly, the perfective conveys the fact that a certain activity has been completed, whereas the imperfective conveys that the activity is progressing, is habitual or is iterative. More generally, the perfective and imperfective "correspond roughly to the contrast between whether the event is perceived from an external viewpoint which is located outside the running time of the event, or from an internal viewpoint located within the running time of the event." (Rothstein, 2016) The progressive sense of a verb implies that an act A has the following structure: in doing A, one is embarked upon completing A, yet one might never complete A. For Thompson, this 'being embarked upon' marks the fundamental logical structure of actions in general: first, it presumes that there is a way in which the action can be completed, that is, one views the sequence of events in light of this end. Second, the parts of an action have an internal temporal coherence—the events are not just temporally juxtaposed with each other; the action might logically be divided into segments such that each segment is done in order that a later segment can occur.

One can illustrate this through examples. Consider making a sandwich. The action of making a sandwich can be divided into segments (steps of making a sandwich such as breaking eggs, cutting cheese into slices, etc.) which follow each other according to a definite order in order to complete the act of making a sandwich. All these segments are unified into one action in view of the end of making a sandwich. Or consider building the Milan Cathedral, which is done over centuries. Here, although there is no particular point of time at which such an activity is envisioned to end when started, there is still a way in which the action can be said to be completed—when the plan of the cathedral has been fulfilled. The coherence of the activity throughout the centuries then come from the fact that all the events of building a cathedral is oriented towards realising the blueprint laid down for building a cathedral; phases of building the cathedral relate to each other, and to this end, teleologically. Finally, there are actions that are not supposed to end, for instance, maintaining a dynasty. The temporal horizon here is indefinitely extended. The meaning of 'complete' would then not be the end of the activity, but its continual success, and all steps taken to ensure the maintenance of the dynasty cohere in that they serve this end.

How do such unities arise? Thompson (2008, 91) employed 'synthesis' as a description of such unities. This has Kantian overtones, where the various syntheses are ultimately moments of the transcendental unity of apperception, thus linked to a transcendental subject. Synthesis thus seems to suggest the existence of a subject who synthesises. What synthesises the temporal parts of an action in Thompson's view? One would have expected that the unity of the phases of an action is fundamentally connected to unity of a life-form. However, Thompson does not make this explicit; indeed, he leaves the connection between the discussion of life-forms he made and the discussion of action in the dark. But it seems there is a natural way to complete Thompson's theory on these lines.

Let us start with simple biological individuals and their actions. Consider bacterial chemotaxis which might happen after a bacterium has sensed some gradient in resources in its environment. The bacterium moves towards the region where there is greater concentration of useful resources. The movement of the bacterium does have the temporal unity described by Thompson:

phases of its movement are oriented towards an end—arriving at an environment with sufficient resources. One might describe the bacterium as sensing or discriminating its environment and actiong accordingly to fulfill its preferences. Now this might seem too anthropomorphic. What is the difference between the behaviour of the bacterium and that of a point particle whose trajectory in a potential can be described through the principle of stationary action (that it takes the 'path of least action')? Do we also say that the particle 'prefers' the path of least action? Perhaps it is useful to extend the talk of 'preference' to this case as well. But regardless of whether that is true, it is here that Thompson's intuitive concepts of form and internal purposiveness has a role to play to distinguish the bacterium from the particle. The preference of the bacterium arises because of the form of the bacterium: its activities attempt to maintain the form. One can then explain the action of reaching a resource rich region of the environment in light of the self-preserving nature of the bacterium. It is only because one can describe the life-form as possessing some internal purposiveness following from its own nature that it makes sense to say that the change effected or the event that happened is an action or activity of the life-form.

The attribution of an action to a life-form inserts the action within the temporally coherent life of the life-form, a coherence derived from the posited form that the life-form possesses, and expressed in the temporality of natural-historical judgements. The unity in a naive action explanation has to be read against the background of the temporal coherence of the life of a life-form. Consider, for instance, a beaver building a dam. The end may be 'protection against predators', an end which follows directly from the self-preserving imperative, which one may understand to be part of the internal purposiveness of the beaver. A naive explanation of the form, 'B is pushing logs into the mud, because it is diverting the stream of the river'; 'it is diverting the stream, because it is building a dam' connects various actions to the end which follows from the self-preserving imperative. The unity or teleological connection between different actions and activities comes from the fact that this is an expression of the internal purposiveness of the beaver itself.

Thus, strictly speaking, Thompson's use of synthesis does not assume a subject who actively synthesises. The unity of an action comes from its place within the temporally coherent life of a life-form. The various phases of an action are decompositions of the action relative to some end; but such ends are themselves located within the life of the life-form, and can be indirectly linked to the activities the life-form engages in to continue to exist.

2.2.2 Sophisticated Action Explanation

The challenge now is for Thompson to explain how the more sophisticated forms of explanations, including reason explanations, can be attained as enrichments of naive action explanation. This unfortunately remains rather under-developed in Thompson's work. However, there is one important thought. In the above, we have linked the explanation of the activity of the bacterium to the fact that it discriminates gradient differences and acts accordingly, thus displaying a primitive preference, a preference which arises due to its form. The action is unified when seen in the light of this preference. In the same way, Thompson writes "it is the fact that the agent *wants* to do B—or even, as I think, the fact that the agent *is doing* B—that joins *the thought that ...to do A... to do B...* and the *agent's doing A* together as cause and effect of the right kind."(Thompson, 2008, 94) Crucially, 'to want' here does not mean just any sort of inclination towards doing A (what Aquinas called appetite in Summa Theol., I-II, Q. viii, a. 1). It is closer to what Aquinas would call will, understood as the rational appetite (Summa Theol., II-I, Q. viii, a.1). It is, very roughly, a capacity attributed to the agent whereby the agent exercises preferential control (through reason) over its appetites, and is to be distinguished from just feeling the prick of desire. One thus gets the following picture: for the agent acting for reasons, besides preferences there exists such 'rational appetites'. As such, when the agent actually acts on on the basis of such rational appetites, it can be described as acting for a reason.

This simple way of scaling up the discussion from the bacterium's preference to the human being's rational appetite is, however, too hasty. Indeed, introducing the will is simply introducing a term for the difficulty—characterising the exact control it has over all the other appetites. To further understand Thompson's theory, one has to look elsewhere. Now Lavin and Boyle (2010) has extended Thompson's theory to provide a more refined theory of desire and will. They make two moves. First, they explicitly identify the 'form' of an agent with the organisation of the agent as a system of self-maintaining (causal) powers, where each power is directed at some end. This system is self-maintaining in that the realisation of an end by a power enables or facilitates (perhaps when cooperating with the realisation of other ends by other powers) the realisation of an end by at least one other power in the system, such that the whole organisation of powers is reproduced or renewed. Second, they claim that the specificity of the rational agent lies in the fact that it is not only capable of goal-directed behaviour in view of its own self-preservation (self-generated goal-directed activities, in their terms). It further has a special 'power' as part of its form, that is, the self-maintaining system of powers: the power of self-conscious reflection and deliberation. In particular, the agent can deliberate on means-end relations (the basic form of which is 'S is doing B because S is doing A') and allow this deliberation to result in the execution of action. Given this extra structure, a naive action explanation of the form 'S is doing B because S is doing A' is enriched. Previously, the 'because' in this explanation links the actions, and the internal structure of the action, to the overarching internal purposiveness of the life-form as a self-maintaining system of powers. Now, if the agent has this new power of reflection and deliberation, the 'because' acquires a new significance. The rational agent has the capacity to know what it is doing and provide reasons for what it is doing, and thus to employ sentences of the form above. A sentence of the form 'S is doing B because S is doing A' holds for S only if S endorses it. It is not just a third person description of S, unless the agent is deluded or self-deceiving concerning his own reasons. It is not because this sentence is already true that S endorses it; rather, the endorsement of S makes it true. To endorse it however is to acknowledge the force of the 'because' involved: as noted above, this is equivalent to endorsing the teleological connection between actions or activities that follow from the form of the agent of these actions. The endorsement by S of this sentence, which is about S itself, thus means S judges the teleological connection on the basis of its own form, which includes the fact that S is a rational agent. Here, it is not necessarily the case that S explicitly thematises its preferences. But it is in principle possible. For the action to be more than habit or automatism, S has to constitute the explanatory relation by endorsing the reason through the reflective knowledge of its own preferences, the system of ends that corresponds to the powers constituting the self-maintaining system. This is the underlying structure of acting for reasons.

This certainly represents a further development of Thompson's theory. However, as we shall see, it inherits problems from Thompson's general framework. We now turn to a critique of the theory.

2.3 Criticisms of the theory

2.3.1 The Concept of Organisation

At several points above, we have seen the centrality of the concept of the form of the agent, which implies that the agent has some kind of 'internal purposiveness'. Yet, this notion has remained obscure in Thompson's work. The obscurity arises from the fact that Thompson rejected any further explication of this concept, in some sense taking it as a primitive through which the circle of concepts including life, agency, reason, and explanation can be illuminated. One might find it odd that he rejected the most natural way of explicating this concept of 'form'—through self-organisation or self-maintenance. He justifies this as follows. Any attempt to clarify the distinctive unity of events that is witnessed in natural-historical judgements "by an employment of prefixes as 'self-' or 'auto-'—as in, say, 'self-reproduction', 'self-organization' or 'auto-regulation'—is ...completely empty. The phrase to which the prefix is attached is always a distraction, and the whole problem is already contained in the reflexive" (Thompson, 2008, 45) This rejection of a natural philosophical explanation of the phenomena led him to his peculiar strategy of clarifying concepts through studying forms of judgement.

The content of Thompson's objection is rather hard to make out. It seems that the problem is not really about 'emptiness', but about the inadequacy of the concept of organisation to capture the nature of life-forms. The objection is rather the following: either the concept of organisation is too general to be able to divide what is a life-form from what is not; or it is too particular (e.g., defining life as a particular organisation of DNA), restricting life-forms to contingent features of the natural world as observed and conceivable now.

The history of attempts to define life might lend Thompson's objection some support. To date, there is no consensus on what is life, and all proposed definitions admit counterexamples or overgeneration. But perhaps this is because the attempt to define life in accordance with our intuitions is a doomed enterprise in the first place. A parallel can be drawn with the study of states of matter. Just as classifications of things into liquids and solids gave way to a sophisticated theory of matter in condensed matter physics, where states of matter are classified based on different principles that are not part of everyday intuition (order parameters, correlation length, symmetry breaking, renormalization...), the same can be true of organisation (indeed, the study of matter is one part of the general study of possible organisations).

From this perspective, what one should conclude from Thompson's argument is not that one should give up on the attempt to clarify the reflexivity through the concept of 'organisation', but that one should not cling to intuitions. The concept of agency is to be determined through possible forms of organisation that we may discover, not defined through prior assumptions on what agents are.

It seems to me that Thompson is prevented from taking this step because of his commitment to taking the manifest image at face value. But, as we shall now see, accepting the manifest image at face value is questionable—the intuitions we inherited from the manifest image responsible for the intuitive conception of temporality and generality seem to be merely inherited from parochial features of our cognitive system. It is to this point that we now turn.

2.3.2 Critique of Thompson's Manifest Image

One can put the challenge in more skeptical terms. How can Thompson guarantee that the concept of a life-form has any objective purchase at all? It seems, on the contrary, that temporality and generality are products of our cognitive system which evolved to make rough distinctions fit for purposes of our survival and propagation, but not as the foundations for a deep understanding of action.

The case with generality is straightforward. As psychologists have pointed out, there is an essentialist cognitive bias built into our cognitive system (Gelman, 2003). For instance, children have a tendency to attribute to certain categories, certainly including 'animals', a true nature which accounts for the features displayed by instances of the categories in question. Such essences precisely support generics of the kind we encountered above. Yet, of course, such 'kinds' need not correspond to kinds one discovers in science. There is thus no justification in simply accepting the forms and kinds that we, given our cognitive makeup, see in the world.

Temporality derives from a different set of cognitive capacities. It might roughly be understood as comprising our ability to recognise goal-oriented agency and phenomena like natural growth. As has been noted by psychologists and cognitive scientists, humans (including infants) and other animals like chimpanzees have a primitive capacity to detect and respond to a minimal form of agency, a capacity which can be traced back to the visual system and is thus pre-linguistic and pre-cognitive (see (Burge, 2022, 466-468) for a survey and references from a philosophical perspective; see Spelke (2022); Spelke and Kinzler (2007) for a psychological view). This minimal form of agency has the following features. First, it is based on our ability to recognise shapes and bodies separate from their environment, and to detect their motion. Infants preferentially respond to certain shapes, for instance shapes of faces or snakes; later, they also show an ability to recognise 'biological shapes' which are asymmetrical in some way. Experiments have also shown that movements of abstract shapes are sufficient to cause subjects to attribute types of goal-directed activities like pursuit or flight (Michotte (1963), Heider and Simmel (1944), Simondon (2013)); further, infants preferentially respond to biological motion that are not uniform.

Second, this minimal form of agency consists in goal-directed activity of a body. Observations of chimpanzees have shown that they do track the goal-directedness of movements of their prey: in particular, they can anticipate where to wait for their prey based on what targets their prey are tracking. It is not implausible to think this holds true for many sophisticated animals, and it definitely holds true for humans. Human infants already reliably distinguish between random autonomous movements of a body and seemingly goal-directed activity of a body (Opfer, 2002). Furthermore, infants not only track the presence of goal-directed activity. They track the value of the goal, and whether the goal is in the line of sight of the agent seeking the goal (Spelke and Kinzler, 2007). Similar studies have pointed to how infants preferentially respond to movements like natural growth. The notion of temporality seems to be connected to this primitive ability to recognise goals and plans specific to individuated bodies.

If features of the manifest image can be directly traced to features of our cognitive system which evolved most likely for practical purposes, there is no reason to think that they provide a valid explanation of how actions are brought about. One can certainly claim that it is still valid within our practice of using such explanations; they will therefore be relegated to playing a pragmatic role in our everyday life. This of course does not answer the skeptical challenge raised in chapter 1. Moreover, Thompson rejects this. He envisions the structures discerned on the basis of the manifest image to have logical and metaphysical significance: the study of life-forms is modelled on Frege's logic which attempts to make explicit the logical forms of thought and judgement. It is claimed that life-forms concepts and judgements about life-forms have a distinct logical form, corresponding to a distinct 'strata of being' (Thompson, 2008, 2).

Given this, it seems that Thompson's refusal of an organisational elucidation of life-forms is an error. After all, the talk of ontological strata is simply a roundabout, and historically loaded way (in particular by conceptions one can trace back to the medieval worldview, a view certainly expressed by Dupleix, and refashioned by Hegel) to express a concept of the emergence of complex organisations. There is thus no reason for endorsing Thompson's metaphysical background and commitment to the manifest image. This opens up the space for an elucidation of life-forms as a physical system, with temporality and generality as specifications of the structure of physical systems.

2.3.3 Critique of Reflection

Finally, the concept of sophisticated action explanation as defended by Lavin and Boyle is unsatisfactory. Firstly, they have assumed a teleological conception of nature through the concept of powers directed at an end. However, they have not yet explained what this means concretely from a physical perspective. Secondly, it is not clear what the capacity for reflection appealed to is. Indeed, philosophers who appeal to self-conscious reflection usually take the authority of such reflection for granted. Yet, without an account of how such powers function concretely, there is no reason to think that such reflective capacities allow one to identify one's reasons as opposed to merely constructing a fictional etiology for one's actions. Finally, the connection between the rational faculty and the life-form remains unclear. How does such a capacity for reflection and deliberation fit in with the life-form concept and its self-maintaining structure? It seems that one has simply retrofitted the special capacity or power of reflective self-consciousness onto the concept of a life-form, thereby losing the progress gained by grounding agents in life-forms. Without this further clarification, it is hard to find Thompson's invocation of 'desire' or the will as explanatory.

The fundamental problem is this. Thompson's theory sketched a progression: from life-forms to rational life-forms possessing the power of reflection on definite contents, making possible the use of sophisticated explanations and the classification of actions as intentional actions. However, Thompson prematurely imports a conception of deliberation and conscious reflection into his model of life-forms, leaving in the dark how these two components cohere. A crucial missing step in this account is how knowledge of the world and self-knowledge of one's own preferences, which enter into the argument of Lavin and Boyle (2010), can be understood as a power that a life-form possesses just like the other powers or capacities of a life-form (such as chemotaxis, cell signalling, metabolic activities...).

We shall follow a different path in providing a naturalistic foundation for Thompson's ideas. In particular, although we will not have the space to develop a theory of desire and the will of rational agents, it is hoped that the naturalistic foundation serve as a foundation also for a future discussion of these issues. We thus turn to developing the content of temporality and generality from the perspective of physical organisation.

2.4 Thompson's Theory as a Naturalistic Theory

In this concluding section of this chapter, we attempt to unearth an organisational understanding of temporality and generality. In the next chapter, we will see if such organisations can be realised in reality.

2.4.1 Generality, Form and Function

Recall a typical generic expressing generality: 'the polypedilium vanderplanki survives anhydrobiotically when deprived of water for long periods'. In such a generic, we are attributing a behaviour or capacity of the life-form based on the 'kind' of thing it is. Thus, it presupposes that the subject possesses an essence or form defining its kind. This kind entails a sort of biological normativity governing the traits of the life-form, where trait is understood in a most general sense comprising features, structures or activities of the life-form: the polypedilium vanderplanki normally can survive anhydrobiotically; one that does not do so is abnormal. This characterisation of the generic seems to suggest that the behaviour or capacity thus attributed has a certain function for the subject. Here, the concept of function is not merely instrumental. Rather, as noted by Garson (2016, 3-7), the concept of function plays three theoretical roles: first, in attributing a function to a trait, one wishes to distinguish certain accidental properties or features of the trait from its function (e.g., the function of a nose is for breathing, not holding up glasses; the function of the heart is to pump blood, not make thumping noises). Second, the attribution of functions have an explanatory value: it is supposed to explain why the trait exists. Finally, a theory of function provides the basis for an account of biological normativity: we evaluate a trait token relative to its function. Thus a trait token may be unable to perform a function, yet can still be evaluated relative to the standard provided by its function. The generic concerning the polypedilium vanderplanki is valid because anhydrobiotic survival serves a function for individuals belonging to the kind 'polypedilium vanderplanki'. The observation that life-form concepts can be judged with judgements displaying generality thus implies that the organisation of the life-form should be able to support attribution of functions, indeed, possession of traits which have a function in virtue of the nature of the life-form in question.

It is in a way unsurprising that Aristotelian philosophers typically look to members of a species to play the role of life-forms. For it seems that species provide a natural environment for articulating functions. Indeed, a prominent theory of functions, the "selected effects theory" pioneered by Millikan (1984), seem to offer a direct way in which the above conception can be realised. In such a theory, the basis for attributing functions to a trait of a life-form lies in the evolutionary history of the life-form. Very roughly, a trait acquires a function for a kind of life-form because it has been selected for because it performs that function. The textbook account of natural selection states that such selection takes place whenever there exists a population of individuals displaying variation, differential fitness, and heritability of fitness (Lewontin, 1970). The precise characterisation of these features, and the study of how they apply to actually existing populations, is a subtle affair. Godfrey-Smith proposed a Darwinian space to conceptualise different sorts of Darwinian populations (Godfrey-Smith, 2009, chapter 3), as well as three parameters for capturing the concept of reproduction (Godfrey-Smith, 2009, 2013, chapter 5). Since the details do not matter for present purposes, I refer the reader to those works. Members of such populations constitute what Godfrey-Smith calls Darwinian individuals.

What is important however is that the concept of a population of Darwinian individuals is more general than that of a population of individuals of a species.[1] Examples of units that can undergo selection but which do not form a species abound: assemblages from different species (symbiotic systems) can co-operate in ways that increase the reproductive advantage of

[1] From this point of view, the formulation of many Aristotelian philosophers which appeal to the species as providing the form of the individual is insufficiently general.

each, or which are even indispensable to each other (e.g. termites, the bioluminescent squid or leaf-cutting ants and their fungus); super-organisms which are collections of individual organisms possessing functional integration and shared fate Wilson and Sober (1989) (e.g., ant colonies or other insect societies (Hölldobler and Wilson, 2009)), but the superorganism itself is not an instance of a species. But the same is true beneath the level of the organism: many experiments in prebiotic chemistry attempts to trace the Darwinian evolutionary dynamic back to the dynamics of chemical interactions. Famous experiments involving RNA molecules shows RNA molecules in an appropriate environment also evolve (Joyce, 2007; Furubayashi et al., 2020; Mizuuchi et al., 2022; Bansho et al., 2016).

The evolutionary framework on functions can accommodate complex functions. This can be seen more directly through the theory introduced by Szathmáry and Smith (1995), who conceptualised the various scales at which Darwinian individuals can emerge through 'major transitions of evolution':

> There are common features that recur in many of the transitions: (1) Entities that were capable of independent replication before the transition can only replicate as parts of a larger unit after it.... (2) The division of labour: as Smith, pointed out, increased efficiency can result from task specialization.... (3) There have been changes in language, information storage and transmission. (Szathmáry and Smith, 1995, 227)

These transitions can be understood as a kind of emergence: if one inquires as to why or how such transitions occur, one invariably has to appeal to the structure of the individuals and their interactions at an earlier stage before the transition, where the existence of the individual at a higher level cannot be taken for granted yet. Thus, for instance, one has to account for how the usual relation of competition between individuals is suppressed in some circumstances, leading to the formation of a higher level of organisation (for one account for how this might be done, see Szathmáry and Smith (1995, 229), drawing on the seminal work of Eigen and Schuster (1977, 1978a,b)). The transition to be driven by the interactions of the replicating individuals, which in turn depend on the structure ("language, information, storage and transmission" in the quote) of the individuals in question.

One might wish to conclude that the organisational basis of the generality of life-forms lie in the fact that such life-forms are Darwinian individuals at some stage of Szathmary and Smith's evolutionary transitions. However, here I think one should not be too hasty. The objection of Thompson that one risks tying the concept of life-form too closely to contingent physical facts holds. For it seems that organisations displaying purposive agency (and thus generality and temporality) is more general than the concept of Darwinian individual. Consider, for instance, the following cases. Hanczyc and Ikegami (2010) have manufactured chemical systems capable of self-movement—an oil droplet which, due to a chemical reaction within the droplet, breaks symmetry and begins moving through the surrounding aqueous solution. The authors claim to have observed feedback cycles that exhibits self-regulation and a form of homeostasis. In a way, one might say that the oil droplet is moving in order to maintain itself. The feats of engineering performed regularly in modern laboratories raise further questions. Should self-assembled micro-nano-robots which are capable of achieving a precise task be said to possess functions or agency (Wei et al., 2021)? These cases do not fall within a Darwinian paradigm; although some authors would simply rule them out (BURGE (2009) even rules out bacteria as possessing agency), as far as I can see no principled distinction has been given.

If we think back to the role of evolutionary theory, one sees that its role here is to secure ascriptions of functions to life-forms which possess a form or essence. The difficulty of accounting for such functions, historically, lies in the fact that the laws of nature known at the time do not seem to be able to account for the precise way in which the life-form as a physical system can evolve such that its traits have specific functions. This is because the laws of nature describing the evolution of the physical system and its parts seems to only make possible 'random' outcomes; the probability of the physical behaviour of all parts of the system conspiring in such a way as to form parts which have and fulfill functions is vanishingly small. Evolution provides a theoretical framework for reconciling these observations, since the long period of evolutionary selection now accounts for why the physical behaviour of a system can have such precision as to its end state. But this also allows us to see that we need not restrict our attention to evolution per se. More fundamentally, we are interested in the possibility of physical systems displaying such precise and seemingly designed behaviour through the laws of nature alone, where such laws are not limited to those of evolutionary theory.

Thus, generality is best understood through the postulation of agents with a certain 'nature', and whose activities have functions, where such functions are manifested in the contrast between ‘random’ behaviour of a physical system and the precision of certain physical processes which play a role in maintaining the agent in existence. The often mentioned need for heredity and the frequent invocation of species by Aristotelians is, I think, a confusion and in this respect Thompson’s refusal to employ an evolutionary model for generality is correct. What is important is rather the existence of individuals of a certain organisation capable of supporting attribution of functions. It is to this concept of an individual that we now turn.

2.4.2 Temporality and Organisms

Let us now turn to temporality. As discussed above, by temporality Thompson designates a sort of temporal coherence of the life-form, the fact that its episodes of activities on the whole seem to display something like an A-series—an internal order. This internal order explains why one episode of activity can appear to be teleologically directed at another future episode, and thus is at the basis of the ‘internal purposiveness’ observed in life-forms. We have also mentioned that this internal order seems to be connected with the individuation of the life-form: the order between the episodes is determined by the sequence of activities required for the life-form to maintain itself, that is, to continue to individuate itself from the environment, although Thompson himself does not talk in terms of self-maintenance or self-preservation.

This description of temporality points to some unity which underlies the temporal coherence of the life-form, something which ‘binds together’ the events in the life of a life-form and which makes these events attributable to this particular life-form. The relation between different phases also display some directedness: an activity begun at an earlier point can be completed at a later point. The existence of temporal order and such directedness is elegantly described by Merleau-Ponty:

> II faut mettre dans l’organisme un principe qui soit négatif ou absence. On peut dire de l’animal que chaque moment de son histoire est vide de ce qui va suivre, vide qui sera comble plus tard. Chaque moment présent est appuyé sur l’avenir, plus que gros de l’avenir. A considérer l’organisme dans un moment donné, on constate qu’il y a de l’avenir dans son présent, car son présent est dans un état de deséquilibre. (Merleau-Ponty, 1995, 207)

Moreover, transitioning out of its present state of disequilibrium does not just lead to the dissolution of the system. The system is, in some sense, reproduced. These features are present in what Godfrey-Smith has called organisms: they are "systems comprised of diverse parts which work together to maintain the system's structure, despite turnover of material, by making use of sources of energy and other resources from the environment." (Godfrey-Smith, 2013, 25) Such an organism is in a constant state of activity; this activity is driven precisely by disequilibrium of its present state, which nevertheless leads to the persistence of the system.

Thompson has objected that phrases such as self-maintenance are not helpful, since the entire difficulty for clarifying the temporal coherence and individuality of the life-form lies in the reflexive. Merleau-Ponty's work and Godfrey-Smith's definition also makes use of the reflexive—the system's parts work together to maintain its own structure. How should we understand the organisation that underlies this reflexive? Godfrey-Smith (2016) points to concept of autopoiesis first introduced by Maturana and Varela (1980). An autopoietic system has the following features.

- An autopoietic system consists of a network of processes of production (transformation and destruction) of components which:
  - through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produced them
  - constitute the system as a concrete unity in space in which they (the components) exist by specifying the topological domain of its realization as such a network.

There are three kinds of relations (Maturana and Varela, 1980, 88):

- Relations of Production and Transformation: these consist of
  - Relations of constitution: determining the components produced constitute the topology in which the autopoiesis is realised
  - Relations of specificity: determine that the components produced be the specific ones defined by their participation in the autopoiesis
  - Relations of order: determine that the concatenation of the components in the relations of specification, constitution and order be the ones specified by the autopoiesis

The system of relations define an autopoietic space:

- Autopoietic space: this is the space whose dimensions are the relations of production of the components that realise it

The notion of an autopoietic space is rather obscure. To conceive of such a space one must abstract from the actual physical realisation of the autopoietic system. In such a physical realisation in a macroscopic system, one can define a temporal order to the events of production of components and describe the components of the system as residing in a three dimensional Cartesian space. However, in the autopoietic space, what matters is only the specification of order, not an absolute order. In this sense, the absolute temporal order does not hold within autopoietic space; similarly, since the autopoietic space is defined by contiguity relations, the three dimensional Cartesian space is in general not an appropriate description. Maturana and Varela thus suggestively say that such an autopoietic space is 'closed and curved' and 'entirely specified by itself'.

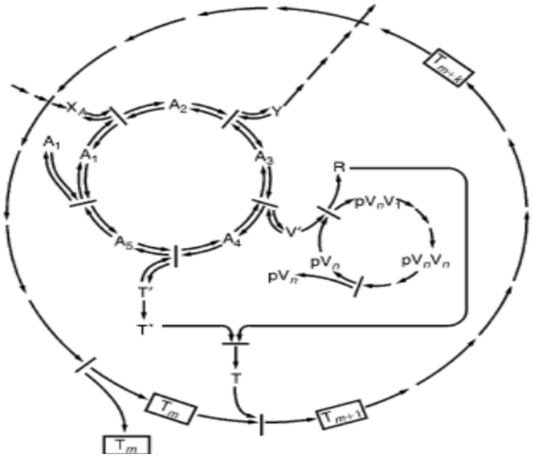
The concept of an autopoietic space can be roughly illustrated by a model these authors introduced with Uribe (Varela et al., 1974). The model in question is essentially a cellular automata. In this case, the autopoietic space consists of the rule space—the space of rules specifying the evolution of the automaton.

More realistic examples come from other theories. The influential concept of an autocatalytic set, first introduced by Kauffman, is one illustration (the formalisation introduced here is based on that of Hordijk (2013)). The definition of an autocatalytic set is based on the following data. First, there exists a collection of processes which transforms material elements into other material elements: a set of process types $P = \{p_1, ..., p_n\}$ and a distinct set of element types $E = \{e_1, ..., e_m\}$. Each process takes a subset of E into a subset of E (inputs and outputs for the process), say $p_i : I_i \to O_i$, where $I_i, O_i \subset E$. One process can support another because its outputs are at least part of another's inputs. However, they may be mutually supporting in another way. Now a central operation in chemistry has to do with the catalysis of one process by some element (that is, the process of a molecule increases the rate of reaction of a chemical reaction). As Kauffman (1993) noted, this is central to the origin of life. Thus, we will need a further catalytic set $C = \{(e, p) | e \in E, p \in P\}$ to designate which elements catalyse which processes.

Within the set of elements, there exists a distinguished set of abundantly available elements in the environment, sometimes called the food source $F \subset E$. The set of all elements that can be generated from the food source through a set of processes P is denoted $cl_P(F)$. An autocatalytic set is then defined as the following data: $P' \subset P$ such that for each $p \in P'$ there exists $x \in cl_{R'}(F)$ for each such that $(x, p) \in C$ and for each $p : I \to O$ in P', $I \subseteq cl_{R'}(F)$.

In this case, the autopoietic space would be the space of pairs of a function (p_i, e) designating the processes and their catalysts.

A more complete illustration of the idea, which is architecturally much more concrete and grounded in biological and chemical principles, comes from the chemoton theory of Ganti (2003a,b). Consider the following figure, taken from Ganti (2003c, 4):



Here, we have three different autocatalytic sets coupled together. The T process designate membrane formation; the A process designates the metabolic system with intermediates A_i, consuming X as food from the environment and producing Y as waste to be ejected from the chemoton; pV_n undergo template replication; R is a byproduct of replication needed to turn T' into T, the molecule needed for the membrane. The three autocatalytic sets together reproduce each other by drawing on the environmental resources.

Let us return to the nature of an autopoietic system. Although the physical realisation is abstracted from in the above specification of the organisation of the autopoietic system, it is essential for the actual existence of such a system. The specification of the autopoietic space as a system of relations is an abstract object—a network or graph. But this abstract system of relations can be concretely actualised in a physical system in multiple ways. One can thus distinguish structure and organisation:

- Organisation: this is the set of relations specifying the autopoietic space
- Structure: this is the set of all physical relations between components in a physical realisation of an autopoietic system.

Finally, the closure of the autopoietic system is not in contradiction with the fact that it can be coupled to the environment in various ways. Autopoietic systems are adaptive, in that they can preserve their own autopoietic organisation through 'deformations' of its structure. Maturana and Varela define this as cognition. That this should be said to be cognitive is not obvious. However, the rough idea is this: the processes taking place in the environment impacts the autopoietic system, leading to deformations. Insofar as such deformations do not lead to the destruction of the autopoietic system, these impacts are in some sense internalised by the autopoietic system. But these impacts come from the environment and simply are part of the structure of the environment. Thus, the structure of the environment is in some sense internalised by the autopoietic system in a self-relating way. But cognition is just such ability to 'internalise' the environment in some sense relative to one's 'self':

- Cognitive domain: this is the domain of all the interactions in which an autopoietic system can enter without loss of identity.

This definition of a cognitive domain might seem to be an unwarranted extension of the usual notion of 'cognition'. It might remind one of the functioning of neural networks which change their weights in response to 'stimulus' provided by input; the weight changes are in these sense modifications which preserve organisation (the connection between neurons). Without evaluating the suitability of this definition of cognition, I will simply note that the aim here is to establish a minimal way of articulating cognition as a form of information processing by a physical structure. We will comment on this later.

2.5 Towards a Unified Theory of Individuation

To recapitulate, from a naturalistic perspective, the generality and temporality of life-forms designate features of systems with a particular physical organisation. The special form of unity in natural-historical judgements reflects the unity of a life-form, a unity which we understood through the concept of autopoiesis. Generality and temporality are aspects of an individual displaying autopoiesis. Generality requires the existence of an individual capable of supporting ascriptions of functions; temporality describes the constant, structured process of individuation which constitutes the individual. In the next chapter, I shall first give a more precise description of autopoietic systems; and in the final chapter, I study what functions are on this basis, and how action and action explanation may be understood in this context.

Chapter 3

The Constitution of Individuals

Introduction

The previous chapter argued that Thompson's generality and temporality should be understood as descriptions of physical organisations, instead of as features of judgements and concepts. Here, we will attempt to illuminate the nature of the individual displaying autopoiesis. This will finally allow us to identify the exact meaning of form and internal purposiveness. We will do so in several steps.

1. In the previous chapter, we have used rather vague terms (self-maintaining system of powers, autopoietic space). Here, we explain how these concepts may be articulated rigorously. For this purpose we introduce the concept of a state space, a central tool for describing and explaining the evolution of complex systems. (3.1)

2. Next, we briefly introduce a way in which such self-maintaining individuals can form spontaneously in nature—the idea of 'contextual emergence' (Bishop et al., 2022). (3.2.1)

3. Next, we give a precise characterisation of purposiveness as 'error correction' (Smith and Morowitz, 2016), the agent with internal purposiveness as a self-maintaining system. We further interpret a self-maintaining system as a system displaying 'closure of constraints', a concept introduced by Kauffman (1993, 2000); Montévil and Mossio (2015). (3.2.2)

4. This characterisation of individuality directly entails structural properties of the individual (Ashby, 1957; Conant and Ashby, 2024): as we shall see, the system has to display a degree of internal complexity such that its interactions with the environment has a temporal horizon—it utilises stored infromation to plan for the future. (3.2.3)

It has to be emphasised that the theory proposed is very much preliminary, and only sketches a number of general ideas and principles. A fully mature, formal theory of individuality has to be left for a future work. Nevertheless, this is sufficient for illuminating key conceptual aspects of temporality and generality (3.3), as well as the status of action explanations (as we shall see in the next chapter).

3.1 Basic Concepts of Systems Theory

Thompson claims that the subject of the attribution of action should be made explicit as a life-form. In our naturalistic theory, the subject of attribution is to be understood as a system. By a system, I mean an organisation specified by the following data: a collection of all physically possible states of the system, the 'state space' of the system; a set of inputs to and outputs from the system (if these sets are empty, the system is a closed system); a specification of how the system changes states based on its inputs and outputs.[1] This is an extremely general notion. First, the notion of a system is scale-free: it can be used to describe organisations from the smallest level (quantum systems) to the largest scales (societies, galaxies,...). Second, it does not respond to our 'intuitions' concerning what constitutes an object or a substance; any organisation possessing such data forms a system.

One might wonder what the utility of this extremely general notion is. As I hope to show, this notion has a central role in scientific explanations; furthermore, it gives one a precise way for articulating 'dispositions' which are useful in providing causal explanations.

3.1.1 The Explanatory Role of Systems

Systems play a central explanatory role because of the way the state space of the system is selected. The concept of the state space has a long history

[1] In its full generality, the concept of a system can be formalised using 2-category theory (Myers, 2022), but we will not need this formalism.

in physics (for a historical view, see Nolte (2018)). The idea of applying the concept of a state space to systems in general, however, seems to have developed through the studies of Kalman and Stratonovich in what is now known as control theory (Kalman, 1960; Kuznetsov et al., 1965). As Kalman (1960) writes, a state is, "intuitively, some quantitative information (a set of numbers, a function, etc.) which is the least amount of data one has to know about the past behavior of the system in order to predict its future behavior." This entails two important features of state spaces.

First, state spaces provide a determination of a certain organisation through the determinable-determinate relation (Wilson, 2023). Each state of a system consists of particular determinates of certain determinables—it is, say, red, rectangular, and so on. Moreover, as Kalman wrote, the determinables and determinates are to be captured quantitatively. Typically, a determinable can be modelled as a geometric space—for instance, colour resides in colour space, extension in our immediate environment as an Euclidean space, velocity in a vector space, 'meanings' in semantic spaces (Gardenfors, 2014) (this last example might be more controversial).

Now the determinable-determinate relation is not just a relation between the more specific and the less specific (typical counterexamples introduced to illustrate this point are being disjuncts: 'yellow' is not a determinate of 'yellow or angry', although it is more specific (Searle, 1959)). This is connected to the idea that determinables could be seen as 'natural', or 'carving nature at its joints', in some sense. While the exact meaning of naturalness is controversial, this fits well with the concept that the quantitative information required to specify a state should be minimal, which intuitively means one should find the most efficient way of modelling the system without sacrificing the predictive power or adequacy of one's model. What this means in practice is less straightforward, and is case dependent. There has been various attempts to model this sense of minimality or naturalness formally (for a hint as to how one can find such efficient models of complex dynamical systems from data obtained through observing or simulating the dynamical system, cf. Koopman operator theory; minimisation of variational or expected free energy; Kolmogorov complexity; ϵ-machines...). It will take us too far afield to evaluate different notions of simplicity, but it is worth emphasising that the 'simplicity' or 'naturalness' involved is not a subjective intuition. It is best understood as an objective relational property of the system in relation to another system—its model.

Second, in order to model the future trajectory of the system, one needs to take into account the ‘tendencies’ of the system. This is best illustrated through a simple example.

Example 3.1.1. *State space of a point particle in 3 dimensional space: here, the state of a particle consists of its position in three dimensional space (taken to be Euclidean) and its velocity (which is also specifiable with three values, and thus reside in the Euclidean space. In our world these six parameters are all that is required for determining the future trajectory of the particle given initial conditions; any less would be insufficient, and any more would be redundant. Thus, the state space, which is usually called the phase space, of the particle is* $\mathbb{R}^6$.

Now three of the parameters required are the directions of the velocity—the tendencies of the particle to move in a certain direction. In general, the concept of the state space models tendencies to change by including the information contained in the derivatives—the rate of change of some parameter needed to specify the system.

3.1.2 Systems and Dispositions

The systems theoretic framework further allows one to rigorously formulate the talk of dispositions. The starting point for relating the talk of dispositions with systems is the following observation. The concept of a state space captures concisely three things: first, all the physical possibilities of a system; second, the ‘tendencies’ inherent in each state of the system; third, the interactions between a system and its environment through the specification of inputs and outputs. These features invite interpretations through the concept of a disposition as follows.

I will take the conceptualisation of dispositions by Martin (2008) as the reference point. Martin argues for the following characteristics of dispositional properties.

1. Dispositions are correlated to their manifestations: a disposition is for certain manifestations rather than another; manifestations come from certain dispositions.

2. A realist attitude towards dispositions: a disposition is actual even if its manifestations are not actual.

3. Dispositions are best conceptualised as forming an interwoven network. They have reciprocal manifestation partners (e.g., soluble salt and solvent water) (Martin, 2008, 2-3)

4. Dispositional properties are inseparably linked to 'categorical', or non-dispositional properties (e.g., spatiotemporal location): as Martin articulates it, no real property of an object, event, process or even space-time segment or field can be thought of as existing at either the potency-free purely qualitative limit, or the purely dispositional non-qualitative limit (Armstrong et al., 1996, 74). Indeed, for Martin each property is simultaneously categorical and dispositional (Martin, 2008, 65).

The applicability of this conceptualisation to the systems theoretic view is evident:

1. Dispositions describe tendencies for change of a system. A particular trajectory traced out by the system in its state space following these tendencies can be seen as a particular manifestation of the disposition.

2. One may take a realist stance towards state spaces: the space of possibilities of a system is something we have discovered constituting the nature of the system; the system occupies one state at each time, and this state includes the 'tendencies' required to determine its future trajectory. Thus, the tendencies are taken to exist even when they are not manifested.

3. The data constituting a system captures all the possible states the system could be in given all the inputs and outputs that the system can receive. The inputs and outputs originate from or end in other systems, which might be understood as dispositional partners. Of course, as a scientific model, one can only be selective in modelling such partners.

4. The systems theoretic formalism makes no distinction between dispositional and categorical properties. It starts with the idea of the minimal data required to account for the future evolution of a system. As such, such data might involve both what we call dispositional and categorical properties. In that sense, both dispositional and categorical properties are understood on the same basis. This differs slightly from Martin's claim that every property is simultaneously both categorical and dispositional. However, it captures the idea that the same objective structure

> characterising a system is simultaneously categorical and dispositional. In a way, this suggests that perhaps we need not be too attached to our everyday notions of 'categorical' versus 'dispositional'.

It is thus appropriate to think of systems as possessing a number of dispositions; indeed, one may view the state space formalism as providing a rigorous language to talk about dispositions, although here one has to be careful not to articulate dispositions in terms of natural language but in terms of the state space.

The above comparison however does not entail any particular philosophical view of the nature of dispositions. The dispositional view merely suggests a relational conception of systems: systems are only defined relative to other systems with which it can interact; a system is in a certain state with certain observable determinates only when viewed relative to other systems. The interactions may be specified through laws of nature, in which case one would also have to take the concept of a law of nature as primitive; it could also be specified as mere regularity, in which case the collection of interacting systems comes to resemble a 'Humean mosaic'; one can equally consider global constraints on collections of systems (e.g., symmetry considerations) which define how they interact; finally, one can of course attempt to supply a powers ontology to ground the talk of dispositions, where powers are modally fixed ontic properties (properties whose identity is given by their causal/dispositional/nomic role (Bird, 2016, 345), and which cannot have other dispositions in other possible worlds). We need not enter into this discussion here.

Having introduced the concepts of systems theory, one may proceed to a discussion of the autopoietic individual.

3.2 The Constitution of Internally Purposive Individuals

3.2.1 Contextual Emergence

In the previous chapter, we noted that a traditional difficulty with understanding the kind of internal purposiveness of life-forms lies in the fact that the laws of nature seems to give one little resources for explaining how physical processes that appear to have a function or purpose can arise. A complex system consists of a large number of components. The probability that all these components conspire to produce a precise outcome, an outcome which can be described by us as serving a particular purpose or function, is vanishingly small. Yet this is precisely what can be observed. For this to be possible, there has to be some way in which the space of physical possibilities of a system actually differs from the total space of possibilities. This may be conceptualised through contextual emergence (Bishop et al., 2022).

According to contextual emergence, a certain system emerges from a background environment due to an alteration in the space of possibilities in this environment. This alteration takes place due to the existence of constraints in this environment (these constraints form the context for the system, explaining the name). These constraints can be treated as unchanging in the timeframe where the emergent system exists and are thus called stability conditions. For instance, the formation of emergent patterns in fluid motion in Rayleigh-Benard convection requires that there be a sustained temperature gradient as well as stable dynamics at larger length and time scales which damp out microscopic fluctuations, stability conditions which when combined make possible the emergent patterns in the system of fluid flow.[2]

The stability conditions alter the space of physical possibilities—the state space introduced in the previous section—in the following way. Sets of stability conditions "define the accessibility of modal space of the system to specific regions of physical possibility through the constraints at relevant scales defining access to those subspaces." (Bishop et al., 2022, 31) This restricts the possible states a system enters into, but also enables the system to reach states that, were it possible to explore all states, it would not have been able to reach. The simultaneously restrictive and enabling role of stability conditions transforms the original system's state space, leading to the formation of systems with different properties, a complex process which has been called 'emergence'. I shall not dwell on whether this is an adequate concept of emergence (for this, see (Bishop et al., 2022)); what is important for our purpose is that the underlying mechanism leading to a modification of the space of possibilities makes possible the existence of systems that appear to be ordered (ultimately, possess purposiveness), as we shall see.

The existence of systems that possess some form of order or pattern (in contrast to randomness) can precisely be captured through the formalism of

[2] For a more careful definition of stability conditions, see (Bishop, 2019) or (Bishop et al., 2022, chapter 2).

state spaces. For most systems (in particular for macroscopic systems), the number of variables required to provide a state space description is immense. Now the reason why some discernible pattern can arise in observing such an immense space is that the system does not occupy all of these possible states; moreover, even in the states that the system occupies, it does not transition from one state to any other state randomly. In other words, the system is confined to a small volume of its state space.

How is this confinement possible? An interesting concrete mechanism for contextual emergence has been given by Smith and Morowitz (2016), inspired by condensed matter physics (Nishimori and Ortiz, 2011), (Smith and Morowitz, 2016, 431-432). First, the confinement to a number of states of order arises from the collective effects of the interaction of many subsystems (in general, small-scale degrees of freedom) when appropriately constrained within a certain background. Collective effects can be modelled by global or macroscopic parameters which abstract from the details of the microscopic degrees of freedom, leading to states of order where individual fluctuations are damped out due to these collective effects. Now each microscopic component may fluctuate, that is, be in different states. Yet, this macroscopic parameter is robust in that it still expresses the property of the system despite these fluctuations. This is due to the fact that fluctuations of one part of the microscopic structure is corrected by the general tendency of other parts to remain in a certain average value.[3] Although in principle the system can still explore all states through a collective large fluctuation, such events are exceedingly rare. This accounts for the stability of the system as a whole, and the observed confinement of the system to a small number of states.

This mechanism, when specified in technical detail in particular situations, can be used to explain the stability of various sorts of systems, not exclusively systems associated with internal purposiveness. In the following subsections, we will consider the underlying structure of autopoiesis.

[3] Technically, this is justified through an appeal to the law of large numbers and its generalisations through large deviation theory. Very roughly, this theorem states that the probability that the average of n identically distributed random variables X_i is equal to the average of each random variable tends towards one as n increases to infinity. If one considers the fluctuating microscopic states as identically distributed random variables, as may often be done as an approximation or idealisation, the large number of such microscopic states entail that their fluctuation tends towards an average value.

3.2.2 Error Correction and the Closure of Constraints

In the previous chapter, we saw that autopoietic systems are dependent on the environment, yet constantly distinguishing itself from it through its activities, a dynamic captured by Merleau-Ponty's description that the system is constantly in a state of disequilibrium and directed towards the future in a self-producing manner. The dynamic underlying the simultaneous dependence and distinction has been stated with great clarity by Schrödinger. Dependence on the environment leads to wear and tear on the system:

> When a system that is not alive is isolated or placed in a uniform environment, all motion usually comes to a standstill very soon as a result of various kinds of friction; differences of electric or chemical potential are equalized, substances which tend to form a chemical compound do so, temperature becomes uniform by heat conduction. After that the whole system fades away into a dead, inert lump of matter. A permanent state is reached, in which no observable events occur. The physicist calls this the state of thermodynamical equilibrium, or of 'maximum entropy'. (Schrödinger, 1944, 69)

Yet, the autopoietic system can fortify itself against this decay. Its activities driven somehow allow it to produce itself, that is, to evade equilibration by having some way to reverse the effects of wear and tear. The autopoietic system precisely possesses a form of order; thus, from the point of view of state spaces, one can say that the system, instead of 'exploring' more of the state space, is instead confined to a part of the state space. What is distinctive of the autopoietic order is however that such confinement takes place through the interactions between the system and the environment. The system is not confined because it is insulated from the environmental perturbations. Rather, it is through a constant flux of matter and energy with the environment that the system maintains itself.

This is the background against which particular activities that contribute to the autopoietic dynamic appear purposive, for they appear to serve self-preservation through attaining some sort of end with great precision. Typical examples of such purposive processes can be found not only in behaviour of animals, but in many molecular processes, such as the processes in the 'central dogma'—DNA replication, transcription and translation (the energy for such operations ultimately come from metabolism). It has long been noted that the error rates of such processes are very low. Tania Baker has proposed the following analogy for DNA replication to bring the point home. Think of DNA has the size of a sewer pipe of diameter 1m; imagine DNA replication to be a delivery process by a truck. The accuracy of DNA replication, when scaled up, is the same as such a truck "travelling at the speed of 500km/h making a delivery of one of four colored boxes on both sides of the street once every 10cm, completing its daily journey (for the case of bacterial replication) in 40 minutes. In this highly efficient delivery process, the truck would deliver a wrong package only once every 3 years." (Phillips and Orme, 2020, 370) It is unsurprising then to find statements such as the replicosome seems directed at a goal—the replication of DNA. The purposiveness of systems thus designate the possibility of utilising the flux of energy and matter through the system to complete (usually extremely precise) tasks. How can this be done?

An illuminating way for understanding this sense of purposiveness has been provided by Smith and Morowitz (2016), borrowing from information theory. In this framework, the capacity of a system to perform precise tasks is conceptualised along the same lines as decoding a message transmitted across a noisy channel while minimising the error rate, where a task completed is likened to a message decoded with small error. The capacity to successfully carry out a task (e.g., the processes of the central dogma mentioned earlier) despite environmental fluctuations is then likened to the possibility of decoding a message with a certain degree of accuracy despite noise.

Let us first describe the theory of error-correction in transmission of messages before passing to a justification of this analogy. In information theory, one is interested in reliably transmitting a message across a noisy channel. Each channel has a certain channel capacity, intuitively the maximal rate at which information can be transmitted across a channel. A fundamental result of Shannon shows that, when transmitting at rates less than the channel capacity, there is a way of coding the message such that asymptotically the probability of error of decoding is zero. The key insight of Smith and Morowitz (2016) is the following: the same mechanism is at work in the suppression of fluctuation in the mechanisms responsible for contextual emergence of stable systems as in the suppression of error in asymptotically reliable transmission of information. The structural parallel is justified because we conceptualised that emergence of order through probabilistic principles, and the principles governing contextual emergence in the particular case of purposive individuals can precisely be linked to the principles of asymptotic error-correction (for details of this link, see chapter 7 of (Smith and Morowitz, 2016)).[4]

The above explains the concept of purposiveness, but not yet internal purposiveness. To explain this, one has to see how such error-correction mechanisms can be put to use by the system for the system's own self-preservation. This requires a more precise determination of the meaning of self-maintenance. To do so, we draw from the work of Kauffman (1993, 2000). Kauffman started from the observation that each process within the purposive system utilises energy to fulfill a certain task (doing work, that is, a constrained release of energy) which, in some sense, is useful for the system; the resources however have to be used in a very precise manner—that is, the way work is done is highly constrained. The constraint here plays a crucial enabling role. Without the constraint, the release of energy cannot be put to useful work. Now the constraints that operated on a certain physical

[4] Technically, Morowitz and Smith posits a parallel between the large deviation principle underlying asymptotic error correction and the large deviation principle for stability of individual systems. In asymptotic error correction, the large deviation principle has the following form:

$$P_{error} \sim e^{-\mathcal{D}(\mathcal{C}-\mathcal{R})}$$

Here, $\mathcal{D}$ is the block length of the code, $\mathcal{R}$ is the rate of the code, and $\mathcal{C}$ the channel capacity, and $\sim$ means $\frac{lnP_{fluc}}{N} + s \to 0$ (the rigorous formulation of large deviation principles is more technical, and will not be used here). That is, the probability of that the decoder will make a mistake about a particular codeword received is bounded at leading exponential order (see Cover and Thomas (2006) for a rigorous treatment of the large deviation approach to information theory). Parallel to this, Smith and Morowitz (2016) also understands the stability or self-maintaining nature of the system through a large deviation principle. Let N be a parameter that characterises the scale of the system (e.g., the size of a system), and s a certain rate function. Then one has the following principle expressing the stability of a system: the probability of a fluctuation from the ordered state, in the large N limit, scales as a negative exponential according to a certain rate function:

$$P_{fluct} \sim e^{-Ns}$$

One reinterprets the informational quantities as follows: the probability of error means the probability that the system is perturbed out of its target state; $\mathcal{D}$ can be understood as a measure of the scale of the system at which it can be robust against perturbations; $\mathcal{R}$ is the growth in the number of possible ordered states with system size; and $\mathcal{C}$ is the maximal number of ordered states that can be maintained with reliability in the asymptotic limit. The large deviation principle gives a formal statement to the idea of robustness and its relation to error-correction and scale of the system (see (Smith and Morowitz, 2016, section 8.5) for further elaboration).

process doing work may disappear after the process runs to completion, in which case one cannot perform the same task without re-introducing constraints. Kauffman proposes that the concept of self-maintenance should be understood as a 'work-constraint cycle': the set of constraints facilitating the work done by the set of processes of the system should be maintained or reproduced by the operation of one or more of these processes that are facilitated by these constraints. Montévil and Mossio (2015) made this idea more precise, naming it the closure of constraints. To understand this idea, let us define the concept of a constraint more precisely.

The nature of such constraints is best understood by considering an example. For instance, in research into the origin of life, one hypothesis posits that before the existence of cell membranes the relevant antecedents of the chemical reactions now taking place in cells took place within rock formations, where the crevices of the rocks form a natural 'boundary' for such reactions before the existence of cell boundaries. Such a boundary constrains the movement of chemicals and thus their interaction. But it is itself made of physical processes, albeit changing in a negligible way at the timescale at which chemical reactions take place.

There are three crucial aspects concerning constraints that we can abstract from the example. First, one should not understand a constraint as something fundamentally different from the physical process at work. A constraint is realised in certain physical processes and physical elements. Being a constraint is a role that a physical process might play for other processes. However, and this is the second point, the processes in which the constraint is realised and the processes constrained usually run at different timescales. For instance, in the above example, the timescale of chemical reactions and timescale of the change of rock formations are separated by orders of magnitude. Thus, associated to each constraint is a timescale. Third, the changes wrought by the constrained process (the chemical reactions) on the constraining process (the rock formation) can be treated as negligible in the relevant timescale; in other words, the constraint remains invariant under the operation of the constrained processes. Montévil and Mossio (2015) formulates such invariance and the difference that a constraint makes via the concept of symmetry—each invariance is some kind of invariance under a symmetry: a process can be modelled as a geometric object in a state space. Its invariants are thus symmetries under certain transformations of this state space, for instance the transformation brought about by the dynamical evolution of a constrained process. In these terms, they define the concept of a constraint on a process as follows: given a process $A \to B$, C is a constraint on $A \to B$ at a timescale τ if and only if:

- $A \to B$ and $A_C \to B_C$ (the process as constrained by C) are not symmetric at τ (e.g., enzymes as catalysts)
- A temporal symmetry is associated with all aspects of $C_{A\to B}$ (the constraint while the process $A \to B$ runs) at timescale τ

Constraints may depend on each other. Constraint C_1 operating at timescale τ_1 depends on C_2 operating at timescale τ_2 if C_2 is a constraint on a process that produces the elements in which C_1 is instantiated. One says there is a direct dependence between constraints C_1 and C_2 if the relevant aspect of C_1 does not depend on C_2 through another process.

Closure can then be defined as follows: a set of constraints $\mathcal{C}$ realises overall closure if for each constraint $C_i \in \mathcal{C}$

- C_i depends directly on at least one other constraint belonging to $\mathcal{C}$
- There is at least one other constraint $C_j \in \mathcal{C}$ which depends on C_i

It is important to note that not all constraints that contribute to the maintenance of the system are included within the closure of constraints, since some of these constraints might not depend on constraints within the closed system of constraints. Thus, the framework accommodates a dependence on the environment.

The work constraint cycle and the conceptualisation of the system as satisfying closure of constraints provide a concrete way of understanding autopoiesis. In fact, it subsumes and improves the concept of autopoiesis. Autopoiesis is subsumed since the mutually supporting network of processes posited in the autopoiesis can be understood in terms of processes and their constraints. It is improved on in two ways. First, in conceptualising autopoiesis, Maturana and Varela did not provide much details concerning how such a system of mutually supporting processes are possible. Here, the physical basis—constraint release of energy—becomes explicit. Second, the present framework makes explicit the role of constraints, and thus the possibility of contextually emergent properties of the system.

3.2.3 The Structure of Internally Purposive Individuals

It is perhaps surprising that from the principles discussed in the previous subsections, one can derive conclusions about the capacities of these internally purposive individuals, in particular allowing us to attain a more precise understanding of cognition (the cognitive domain in autopoiesis). This is pointed out by Conant and Ashby (Ashby, 1957; Conant and Ashby, 2024) who formulated the principle of requisite variety and the good regulator theorem. To explain these ideas, we first set down some terminology.

Variety is defined for a set. It may mean two things: the number of elements in the set, or the logarithm of the number of elements in base two. On the basis of this definition, the law of requisite variety states the following. Consider a system, called the regulator R, which is supposed to regulate a number of variables within a range; the variables to be regulated are subjected to influence by another system, E in the environment. Then the law of requisite variety states: only an increase in the variety of R, which roughly designates the internal complexity of R (the number of responses it is capable of giving to environmental perturbations) can reduce uncertainty of the effects of E, that is, to regulate well. In other words, the number of distinct states a controller can sense, and the number of distinct actions it can take in response to samples, must at least equal the number of distinct forms of disturbance that can drive the controlled system out of its target domain. This is a statement about the complexity of the regulator.

The law of requisite variety may be generalised as follows. Consider two systems, R, S, where R is trying to regulate S such that S remains within certain states. A regulator that is able to respond accurately to perturbations of the environment in fulfilling this task has to not only match the complexity of the environment. It has to, in some sense, be able to keep track of and exploit the structure in the environment so as to respond appropriately. This intuitive idea is stated more rigorously through the ‘good regulator theorem’, which states that under some general conditions, the uncertainty of the output of a controlled system is minimised if there is a deterministic map from states of the controlled system to states of the regulator. In this sense, R is able to accurately fulfill its task only through the information about S it stores as represented in this deterministic map (in Ashby and Conant’s terms, the regulator R has to have a model of S to be a good regulator).

The pioneering theory of Ashby and Conant is important in linking the task of regulation with the physical structure of the systems (in the good regulator theorem, the physical structure stores a model) involved. This is of great significance to the present problem since one can view self-maintenance through the lens of regulation: self-maintenance is possible only when many different variables and subsystems are regulated to within a certain range. The work of Ashby and Conant thus implies that one can reach conclusions concerning the structure of self-maintaining systems on the basis of the fact that they are self-maintaining.

This is precisely what is done in recent work extending the original ideas of Ashby and Conant. One line of work shows a connection between the efficient use of thermodynamic resources and the memory and predictive capacity of the system. Very roughly, the reasoning runs as follows. The environment in which systems are placed evolve temporally. Systems which are able to maintain themselves have to do so through storing information about some of the environment's past states and predicting its future states. Now just as Ashby and Conant showed that a system that is capable of self-regulation has to have sufficient internal complexity, these results show that systems capable of utilising thermodynamic resources efficiently has to be able to retain information and predict environmental structure (see, e.g., (Still et al., 2012; Still, 2020; Boyd et al., 2024)). Furthermore, it has been argued that systems capable of evading equilibration have to be capable of utilising such thermodynamic resources efficiently, for otherwise they could not have formed.

The existence of systems possessing some form of memory and predictive capacity is the basis for cognition; it also gives a concrete way for one to understand Merleau-Ponty's statement that the present is supported in the future. For the present states of the system can be seen as partly containing predictions about the future, predictions which are crucial in determining the tendencies of the present state of the system.

3.3 Thompson's Temporality Revisited

Having developed the above ideas, we can finally apply them to interpret Thompson's ideas. First, the crucial notion of 'form' receives a straightforward interpretation as the system of constraints displaying closure that defines the system. The connection between such a conception and temporality is straightforward. The temporal coherence in the phases of the life of a system displaying closure of constraints can be traced to the fact that, each constraint plays some role for the instantiation of processes constrained by other constraints, where the temporal order in which the constraints are brought into play are crucial for the closure of constraints, and where all processes and constraints serve to maintain the system. The fact that the processes constrained seem directed at some end, that is, display some form of 'internal purposiveness' comes from the fact that they are processes that are constrained to fulfill precise tasks within the overarching dynamic of self-maintenance. Internal purposiveness arises from two factors: first, the purposiveness comes from the reduction of physical possibilities that the presence of constraints lead to, creating the appearance that the result of the constrained process has been selected for; second, such 'selection' contributes to the persistence of an individual form.

Returning to Merleau-Ponty's observation, one can now see that the 'disequilibrium' associated to the life-form is simply the out of equilibrium state of the life-form: the system displaying closure of constraints is in a state of balance between the equilibrating flow of matter and energy, and the order maintaining flows instantiated in processes constrained to complete precise tasks. One could describe the system as attempting to minimise disequilibrium, but structurally prevented from actually returning to equilibrium. This tension is the foundation of the temporality of the system.

To conclude the present chapter, we will mark two differences of the present conception with Thompson's life-forms. First, the present view shifts the focus from *life*-forms to systems (not necessarily living) displaying a particular organisation. The question, whether this organisation also serves as a real definition of life, is derivative for our purposes. The focus is simply on the existence of purposive forms of organisations. Second, here one is not defining the life-form in terms of a system of powers as in Boyle and Lavin's approach, but as a system of constraints on material elements and processes. The fundamental dynamic is simply a disequilibrium in matter and energy. Describing the system in terms of certain powers or dispositions like desire, belief, or cognitive powers is a theoretical move whose justification has to be provided independently. To justify such conceptualisations, one has to connect Thompson's project with that of the naturalisation of content: how can the processes which are fundamentally driven by disequilibrium in terms of energy and matter be legitimately described in intentional or even psychological terms. This is a difficult question which we shall not treat here. We will restrict our attention to minimal forms of agents where such questions do not arise.

Having treated temporality, we may now treat generality and move on to answer our initial questions concerning the nature of the attribution and explanation of actions. We shall start with the idea that generality assumes a conception of individuality which supports attribution of functions. In the next chapter we show that the conception of individuals presented above can do explain functions and actions.

Chapter 4

Function, Affordance, Action

In this chapter, we will proceed as follows. First, we will define the concept of a function for individuals which possess closure of constraints. This theory of functions develops the 'organisational theory of functions' (Mossio et al., 2009), which we defend against a number of criticisms (4.1). Next, the relation between the concept of function and of action will be examined. Here, we shall draw upon the the concept of an affordance (4.2). In the final section, we will discuss the nature of actions and action explanations.

4.1 The Organisational Theory of Functions

4.1.1 The Theory

As noted in chapter 2, functions play three theoretical roles (Garson, 2016, 3-7): first, distinguishing accidental properties or features of a trait from its function (e.g., the function of a nose is for breathing, not holding up glasses; the function of the heart is to pump blood, not make thumping noises); second, explaining why the trait exists in the first place; finally, providing the basis for an account of biological normativity.

The organisational theory of functions is one candidate for such a theory of function. I take its most mature development to be in the work of Mossio et al. (2009); Moreno and Mossio (2015); Mossio and Bich (2017); Saborido et al. (2011). These authors argue that the most general concept of function is one defined in terms of systems possessing closure of constraints, an idea that meshes directly with the framework developed in the previous chapter.

We will first present and develop the theory, and then turn to some criticisms.

In the organisational theory, a trait which has a function has to belong to a system whose organisation displays constraint closure. Given this, a trait T has a function if the following three conditions are satisfied:

- T contributes to the maintenance of the organisation of the system
- T is produced and maintained under some constraints exerted by the organisation of the system
- The system is organisationally differentiated (Mossio et al., 2009, 828)

Here, organisational differentiation is introduced to distinguish the systems having functions from self-organising systems or dissipative systems such as the convection cells that form in Rayleigh-Bénard convection (that is, when a horizontal layer of fluid is heated from below, the fluid develops and maintains patterns in the form of convection cells as long as the heat is applied continually) which seem too simple to possess functions. Organisational differentiation imposes the following condition on systems: the system can generate distinct structures which contribute to self-maintenance in different ways. This rules out simple self-organising systems like Rayleigh-Bénard convection cells or other dissipative systems because these systems typically can only support one structure or pattern.

Functions thus defined can indeed play the three theoretical roles mentioned at the outset. First, accidents are distinguished from functions because the trait which has a function has to not only contribute to the maintenance of the organisation, but is itself produced and maintained under some constraints within the organisation of the system. Now a common objection draws upon example of the following sort: the definition of function here seems to include holding up glasses as a function of the nose, which intuitively is not a function of the nose. The glasses do belong to a closed system of constraints, and is maintained by and contributes to this system. After all, being able to hold up glasses means one can see more clearly, which contributes to maintenance; in turn, one learns how to keep glasses on, repairs the glasses and so on, and thus the pair of glasses is maintained and produced under some constraints. This is however not an objection. The nose, like any trait, can have multiple functions. Indeed, the acquisition of functions by a trait due to placement in new environments (e.g., in a society where glasses can be obtained) is related to the phenomenon of exaptation, where traits can be co-opted for new functions.

Second, attributing a function to T has explanatory value in two ways: first, the way in which T is produced and maintained is explained through the second condition; second, the first condition implies that T contributes to the self-maintenance of the system, which partly explains why the system itself continues to exist in the first place. Finally, biological normativity of a function is determined with respect to the self-maintaining organisation of the system to which the functional trait belongs: a trait within an organisation is not performing its function if it no longer contributes to the maintenance of the organisation. This remains true even if the trait fails to perform the function regularly.

Montévil and Mossio (2015) have added a further clarification to the theory of functions which points to the relation between functions and constraints. The way in which a trait serves its function is that it works as a constraint within the closed system of constraints defining the individual: it constrains some process while remaining unchanged at a certain timescale. For instance, the heart has as function to 'pump blood'; this satisfies the definition of a constraint because first the heart modifies blood flow in a way that would not be possible without the heart in the biological organisation of the human body, and second the structure of the heart remains invariant with respect to the flow of blood at a certain timescale.

To complete our presentation of the theory of functions, it is important to note a final feature of the theory of functions. This has to do with the fineness of grain of the description of the function of a trait. In the above we have described the function of the heart as 'to pump blood'; yet, one might ask whether this is an accurate characterisation. Should one not say that the function is rather to regulate the velocity of blood flow, or to do so according to a certain rhythm? A very similar question has arisen a propos sensory functions several decades ago when philosophers discussed the naturalisation of content. The controversy may be illustrated through a famous example. Consider a frog sensing and eating flies that fly around it. Among philosophers who agree that the frog has a certain representation of its environment, there are different intuitions concerning what this representation is about: the content of the frog's representation has been identified variously as 'fly', 'moving, nutritious black dot', 'something, small, dark, moving', 'food'. What is the fineness of grain of the content of the representation for this particular sensory function of the frog? The organisational theory, as stated, does not address this issue. It simply mentions that the trait which has a function contributes to the maintenance of the organisation, without specifying how one should describe this contribution.

With respect to this issue, I think Neander (2017) is exactly right to state that the answer to such a question should not be a matter of intuition. An identification of the right description of a function is, to a large part, an empirical matter. There is no point in discussing what the content of the representation of the frog is without an understanding of the information processing capacities of the central nervous system of the frog. A consideration of the problem of content given such biological facts is presented by Neander (2017, chapter 5).

This problem may be clarified using the concept of the state space. The state space provides the minimal description of a system that accounts for all the potential behaviours of the system (or a sufficiently large range of them). The correct description of the function of a certain trait of a system is also based on this state space description of the system. The function of the heart or of a part of the sensory system should be described through the role these systems have for the individual system in question. This point becomes more concrete when we consider some examples.

In a specific case, the investigation of the right description of a function may take place as follows. Take sensory systems first. The goal is to describe the space of distinctions the system can make as a space with a number of states—for instance, a space in which a black dot is located might consist of states defined by parameters including a small range of spatial extent, colour, velocity. Discovering such a space of distinctions has to be an empirical matter: one subjects the system to various stimuli and observes the reactions in order to classify such stimuli into equivalence classes. Now while the choice of stimuli may be straightforward for simple cases, this might not be so for more complex cases; one thus needs a principled approach. The principled approach should derive a space of distinctions that a system can make on the basis of the structure of the system. This is largely an open problem. However, I shall illustrate the potential of this approach to contribute to the issue of the fineness of grain with a simple example.

The example concerns the space of distinctions that a part of the visual system can make. Recall that a receptive field is, according to textbook definitions, an area in which stimulation leads to response of a particular sensory neuron. Thus, for instance, the stimulation of some receptive fields on the retina has been known to lead to response of particular neurons coding for a certain direction. The general problem of discovering a space of distinctions for a set of neurons has been formulated, with some simplifications, by Curto et al. (2013). Now it is generally accepted that external stimuli are encoded by the brain via neural codes—patterns of firing of neurons. Neurons downstream of the receptors generally only have the neural code as information to infer what the stimuli is. In other words, the distinctions encoded by this neural code is a good candidate for a preliminary understanding of the 'content' of a representation.

More precisely, let us first consider a space of stimuli X, for instance, a space of colours or colours at locations. The receptive fields form a way of covering the space of stimuli, in that each receptive field detects features of one patch of the space, such that the receptive fields collectively detect all regions of the space (the receptive fields thus form what is called a cover of X). Denote the elements of the cover by U_i. Now let the neurons which react to stimuli belonging to the space X be denoted by N_X. A neural code is modelled as a binary code, that is, a code where each word is a map $w : N_X \to \{0, 1\}$. For each word $w = a_1...a_n$, we may define its support—the set of i such that $a_i = 1$. The code we are interested in is the code associated with the covering U_i. This code, denoted $C(U)$, is defined as follows:

$$w \in C(U) \leftrightarrow \bigcap_{i \in suppw} U_i - \bigcup_{i \notin suppw} U_i \neq \emptyset$$

Here each word is a map $w : N_X \to \{0, 1\}$. The problem of identifying the content of the representation given by the neural codes is then the following: what properties of the space X can be recovered by knowing only the neural codes C_X? Curto et al. (2013) proved that in fact one can recover the homotopy type of X.

The ideal procedure for determining the fineness of grain of an ascription of function is thus a circular process between experimentation and mathematical modelling based on first principles. Models allow one to propose a space of distinctions the system can make, which can then be tested experimentally; experimental results might then lead one to revise the theoretical model concerning the structure of the system.

Such a procedure is not restricted to sensory functions. Consider the heart and its function. One first figures out the general structure of the heart, as well as its connections with other parts of the body as part of the cardiovascular system, which is a closed circuit including two other components (systemic and pulmonary circulations, and the microvasculature). Anatomical study of the heart shows more precisely that the electrical activity and the dynamics of the heart allows it is to pump blood through the aorta and the pulmonary artery, which it does through maintaining the cardiac cycle. The failure of self-maintenance of this system, which leads to death (through cardiopathologies), gives one a rough idea as to what the function of the heart is: for instance, in ventricullar fibrillation there is a chaotic excitation of cells, which leads to irregular pumping, leading to death. From such cases of failure, and from observing the normal functioning of the heart, one can hypothesise that the main function of the heart is to constrain blood flow in a regular manner. One then may build a mathematical model (a survey of such models are in (Quarteroni et al., 2017)), through which one identifies the parameters that allow one to best model the functioning of the heart.

Of course, although an accurate understanding of the contribution of a functional trait has to rely on such procedures, in everyday communication one can simply abbreviate and state ‘the function of the heart is to pump blood’.

Criticisms

Having presented the theory, we turn to consider a number of common objections. Such objections either point to traits that intuitively should have functions but which cannot under the definition of the organisational theory, or traits that intuitively should not have functions but are attributed functions by the theory. Let us consider the these objections in turn.

A number of counterexamples have been discussed in the literature to substantiate the first objection—for instance concerning functions in symbiotic systems or collectives, or intergenerational functions. These have been discussed elsewhere ((Mossio and Bich, 2017) for supra-organismic organisations, (Saborido et al., 2011) for intergenerational functions) and I shall not recapitulate them. Indeed, the fundamental idea could be illustrated by considering another example, that of programmed cell death (PCD), which has not been studied in this context. The phenomenon of PCD is roughly the ‘suicide’ of a cell through highly regulated and multi-step pathways in the cell. These pathways seem to have a clear function and yet do not contribute to the self-maintenance of the cell—they rather contribute to its dissolution. In what sense then can they be said to have functions on the present theory’s definition?

To respond to this, one has to first understand PCD better. There are two distinctions to be made: first, PCD may happen in multicellular and unicellular organisms. Second, there are two broad classes of PCD (Durand, 2021, chapter 8): true PCD and ersatz PCD. In terms of mechanism, one can roughly understand PCD in general as "active, genetically controlled, cellular self-destruction driven by a series of complex biochemical events and specialized cellular machinery." The distinction between true and ersatz PCD arises when one looks at the phenomena in evolutionary terms: one may define true PCD as "an adaptation to abiotic or biotic environmental stresses resulting in the death of the cell", and ersatz PCD as a form of programmed death intrinsic to the cell, but which has not been selected for; it is rather a side-effect of a selection process which targets another trait (although later ersatz PCD might itself develop a function). Now how does PCD in either forms, for unicellular or multicellular organisms, have a function under the organisational theory?

The case of multicellular organisms is relatively uncontroversial. Some cells have to be killed off for the survival and development of other cells (indeed, it is an important process in, e.g., normal development of the embryo to an adult). This is hypothesised to take place through kin selection. As such, PCD of either form clearly has a function for the self-maintenance of the multicellular organism, although not for the cells that are killed off.

The unicellular case is more difficult. Yet, true PCD is something that has been selected for, and indeed it intuitively has a clear function—although not always for the cell itself (e.g., for the E coli, PCD has group-level fitness advantage (Durand, 2021, 112)). How can the organisational theory account for this aspect? First, the organisational theorist may claim that the traits causing PCD do not form part of the self-maintaining organisation of the cell, that is, the constraints making up the closed system of constraints constitutive of the cell; it is only triggered during extreme conditions, and thus strictly speaking does not constitute part of the organisation of the cell. Second, the trait nevertheless has a function, although not for the unicellular organism. Rather, it has a function for the system in which the unicellular organism is embedded. In particular, one may claim that PCD has a function for the species of which the unicellular organism is an instantiation, now itself viewed as a self-maintaining system extended in time through reproduction; indeed, as Durand (2021) notes, PCD in general has many benefits for the species. In fact, this gives us a general way of accounting for evolutionary effects which typically operate on a supra-organismic level: we need only consider the population displaying heredity, variation and differential rates of fitness as a system extended in time, self-maintaining when viewed at a longer timescale than the lifetime of an individual. Having included evolutionary effects into the organisational theory, one can now easily deal with ersatz PCD. For the unicellular organism, one may claim that ersatz PCD need not have any function for the organism, although it may develop a function through exaptation or further adaptations.

Let us now turn to the second class of objections—the overgeneration of functions. The example introduced by Garson (2016, 2017) for the organisational view is typical. There, Garson considers panic attacks. A variety of bodily dispositions together cause a panic attack (hypervigilance to bodily sensations, aversion to strenuous exercise...). If we view these dispositions as a system, then this system seems to be self-maintaining because they display organisational closure; further, it displays organisational differentiation, since the system is capable of sustaining several different patterns or behaviours. Thus, Garson claims that according to the theory, hypervigilance to bodily sensations would have as a function to cause a panic attack. Yet this is intuitively not the case. A similar situation is obesity (indeed, many pathologies such as depression that are caused by a variety of factors that produce and reinforce each other).

However, this seems to be a misunderstanding of the theory. Functions are attributed to systems as a whole. The function of hypervigilance to bodily sensations for the human being as a psychophysical whole, and the function of hypervigilance for the panic attack system, are different. Garson has transported intuitions about the function of hypervigilance for the human being as a whole illegitimately to the case he described, which only concerns the function of hypervigilance for the panic attack system. A similar response can be used for obesity.

4.2 Affordances

The previous discussions have completed our treatment of the issues raised with Thompson's notions of generality and temporality. We may proceed finally to apply these concepts to the clarification of the concept of action.

The central idea that Thompson wants to do justice to is the Aristotelian one that actions are somehow initiated or generated by the subject (there is some internal source of the action/change). Moreover, the relevant sense of initiation is both causal and explanatory. In other words, the action is caused by the agent, and it can also be explained teleologically or through the 'form' of the agent. It is relative to this teleological explanation or the form of the agent that one has a standard to measure whether an action is complete.

The starting point of our present elucidation of action is the observation that actions have functions: intuitively, for instance, an animal executes an action because doing so would help it, say, acquire food or escape from a predator. Let us first clarify the relation between actions and functions.

4.2.1 Actions Possess Functions

The link between actions and functions has been theorised by Moreno and Mossio (2015) as follows. One may reconstruct it as follows, appealing to an as yet unexplained and intuitive notion of 'capacity'. A capacity for a particular action may be seen as a trait of the system with a function: firstly, this capacity is supposed to contribute to the self-maintenance of the system: for instance, an animal hunts when hungry, bacteria move away from high temperature or towards regions with higher concentration of nutrients. Secondly, the capacity is produced and maintained not as an accidental effect of the processes making up the system, but is produced and maintained by constraints within the closed system of constraints (e.g., a capacity for movement of an animal is supported by the motor system). The capacity is time assymmetric with respect to the processes it constrains, since it can be redeployed multiple times. Finally, the system is by assumption organisationally differentiatied.

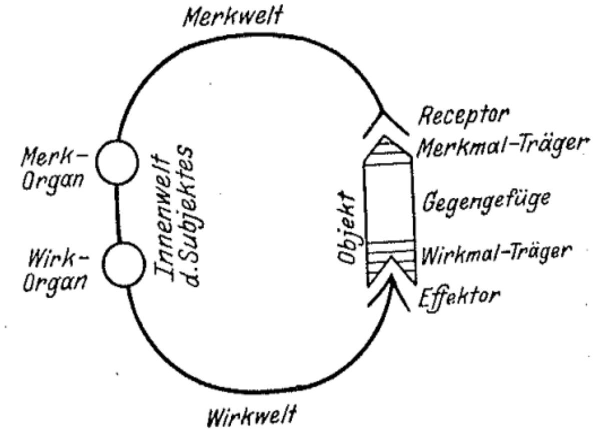
Now although the capacity for particular actions has a function, not all traits which have a function generate action. Thus, Moreno and Mossio (2015, 92) think that capacities for action possess a distinct group of functions—interactive functions, "a subset of biological functions that exert a causal effect on the environmental conditions of the system." But this simple definition does not suffice. This is because the theory overgenerates, including any sort of change produced by the system, for instance reflexes produced by stimulation of a reflex arc, or more generally any kind of reaction mechanism to external stimuli as an action. Now there is a sense in which such reactions are actions: they can be seen as actions of a subsystem of the system in question, which can itself be a self-maintaining system. Yet the characterisation of Mossio and Moreno would attribute this reaction as an action not only of the subsystem, but of the whole system, which seems too liberal.

This difficulty arises because Mossio and Moreno's characterisation of action failed to explain when a system can be said to possess a capacity for action. Possessing a capacity for action, or an ability, is usually understood as involving some sort of reliability in the exercise of the ability, as well as some the flexibility of the system in selecting alternative means for an end or altering the end in face of the feedback the system receives when undertaking or after undertaking an action. Failure in accounting for this leads to the overgeneration. As we shall see, an adequate theory can be sought in the concept of an affordance, which embeds a theory of ability within a more general theory of how systems interact with its environment. We will first introduce the concept of affordance; next, we will see how this helps.

4.2.2 The Theory of Affordances

The Definition of an Affordance

Our use of the concept of affordance draws from two distinct but overlapping traditions. The first is psychological. Indeed, the word 'affordance' is first introduced in a psychological context by Gibson, who was heavily influenced by Gestalt psychology; in fact, the Gestsalt psychologist Lewin already proposed the concept of 'Aufforderung' close to Gibson's affordance (Lewin, 1926). The second is ethological (and cybernetic). The pioneer of this tradition is von Uexküll, who proposed the concept of an Umwelt associated to each animal: the animal's world does not consist of all the physical properties of its environment, but only those that are complementary to the sensory functions and action possibilities of the animal. The behaviour of an animal in the world can thus be conceptualised as a feedback cycle involving the Umwelt:



The resulting view can also be seen as an early precursor of cybernetics, very roughly a theory which attempts to understand behaviour and function through feedback control and information theory (Wiener, 1961; Ashby, 1957; Sobolev et al., 1955); for a discussion of von Uexküll and cybernetics in particular, see Lagerspetz (2001).

These two traditions are convergent and complementary, as we shall see. We shall start with Gibson's concept of an affordance:

> The *affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*, either for good or ill. The verb to afford is found in the dictionary, but the noun *affordance* is not. I have made it up. I mean by it something that refers to both the environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment. (Gibson, 2015, 127)

There are a number of ambiguities in this formulation, which have led to controversies over the proper understanding of affordances (differing views are e.g., (Turvey et al., 1981), (Chemero, 2009, chapter 7), (Stoffregen, 2003), (Heft, 2003)). Michaels (2003) lists four major points of debate: first, whether affordances exist independently of the fact that it is perceived by the animal in question; second, whether affordances are action related; third, whether the existence of affordances depend on the animal's motor capacities (sometimes called effectivities); finally, the fineness of grain of the concept of affordance. To this we might add a further question about the range of applicability of the concept of affordance: Gibson applies the concept to animals, but it seems to be applicable to any system for which some concept of 'good and ill' applies. Thus, a number of researchers in robotics have employed the concept of affordance in the context of robots. In particular, we wish to apply the concept of affordance to any self-maintaining system.

To tackle these problems systematically, we will start with the influential definition of affordances given by Chemero (2009). Chemero (2009, 141) rightly points out that the basic logical structure of an affordance is relational: for a behaviour ϕ of an organism, an affordance is a relation between features of the environment and abilities of the organism, which he writes as affords-ϕ(feature, ability). Here, feature and ability have specific meanings. The concept of feature is introduced here in the context of a contrast between feature placing and recognising properties of objects in perception. In feature placing, one need not explicitly recognise a certain entity which has a property; one simply notes the presence of a feature. Although strictly speaking the concept of feature placing only makes sense in perception, it can easily be used to describe the sensory functions of a self-maintaining system, for instance in terms of the distinction between the task of object recognition/attributing properties to objects and the task of registering equivalence classes of stimuli (features) and responding appropriately. The concept of ability here has two features of note. First, there is a normative dimension. It sets up a standard to judge how X is performed as a function. Furthermore, as with biological functions discussed previously, the agent can systematically fail to do X, yet be performing the function. Second, the concept of an ability is conceived from an evolutionary perspective: "Abilities come to play the role they do in the behavioral economy of the animal because, at some point in the past, they helped the animal (or its ancestor) to survive, reproduce, or flourish in its environment." (Chemero, 2009, 145)

This definition of affordances has come under criticism. Rietveld and Kiverstein (2014) have objected that Chemero's conceptualisation of abilities fails to account fully for the basic normative dimension of abilities. For Rietveld and Kiverstein, the normative dimension of abilities comes not only from the fact that an ability is a function that has been selected for evolutionarily. Rather, the normativity often comes from the fact that abilities are embedded within "sociomaterial forms of life", relatively stable ways of living that impose sociocultural normative constraints. Thus they provide an alternative definition: "Affordances are relations between aspects of a material environment and abilities available in a form of life." (Rietveld and Kiverstein, 2014, 335) It seems that underlying the objections of Rietveld and Kiverstein is the fact that the functions as selected for through natural selection do not suffice to account for the normativity of abilities as we see in communities with a developed social life. But one can sidestep this objection if one employs the organisational concept of function: indeed, a self-maintaining system might include social constraints as part of its organisation or as part of the environment on which the organisation depends. Replacing Chemero's theory of functions by the organisational theory and the inclusion of such social constraints accommodates, at least in principle, the concerns of Rietveld and Kiverstein. However, a proper account of social practices as constraints will take us too far afield, and we will not further delve into this issue here.

What exactly does the concept of ability entail if one replaces the evolutionary theory of function with the organisational one? It is precisely through the concept of an ability that we explicate the notion of capacity. To gain a more precise understanding of an ability, we can return to the basis of the concept of affordance in the self-maintaining dynamic of the system in question. An ability is a trait with a function; it thus contributes to the self-maintenance of the system in question. To do so, as discussed in the previous chapter, the exercise of an ability amounts to the completion of tasks in a regulated manner. The perspective of error correction implies that under appropriate conditions, the perturbations leading to errors in the execution of an ability can be corrected for. This entails that the execution of an ability can be done reliably. Moreover, as discussed in the previous chapter, for regulation to be possible, the system also has to be equipped to react to a range of circumstances, that is, environmental perturbations in order to fulfill the same task (requisite variety and the good regulator theorem). In other words, the system can change the ways in which it fulfills the task when needed to overcome a range of challenges. Thus, the system has to display some degree of adaptivity; there is an increase in adaptivity if the system has greater capacities of memory, prediction, and locomotion, as discussed in the 3.2.3 of the previous chapter. Abilities are thus reliable and adaptive.

This conceptualisation of abilities have direct implications for the theory of affordances. Many affordances given as examples often focus on features in the immediate environment of the system—surfaces affording walking or climbing, features of objects affording grasping, and so on. Yet if an ability has a function, and a function is instantiated by a constraint operating at a certain timescale, associated with each ability is a natural timescale. Thus, theoretically, abilities are not restricted to immediate reactions to the environment. They could involve longer timescales, in particular involving other abilities of the system such as prediction and planning. Thus, affordances are not just features that offer opportunity for immediate reaction; they might involve action over a period of time or at a certain point in the future, involving plans with a certain temporal depth.

At this point, one can settle the controversies raised at the beginning.

- Whether affordances exist when there is no system to exploit them can be settled as follows. An affordance, as a relation, is instantiated when both the feature of the environment and the ability of a system are instantiated. Where there is no system to instantiate the relation, the affordance is not actual. It would exist just in case a system with the appropriate ability were to exist.

- Affordances relate to abilities in general, and not so much to action or effectivities; in fact, one can extend the concept of affordance to mental actions (e.g., finding a mathematical proof as responding to particular affordances provided by a certain mathematical object)

- the problem of the fineness of grain of affordances is resolved by utilising our previous discussion of the fineness of grain of functions. Let us consider how it plays out through an example (taken from Michaels (2003)). Suppose one is playing baseball. A baseball swing comprises of a step and swing phase; are these two distinct affordances or a single one? One might further ask about a swing: is, say, a high swing and a low swing both afforded by the bat and the environment? Is there a point at which the concept of affordance no longer applies? Within the present formalism, this is best understood as a question about the fineness of grain of abilities; but the fineness of grain of abilities follow from the fineness of grain of functions. There are two affordances stepping and swinging, or of high swing and low swing, if the physical constitution of the baseball player is best understood as supporting two abilities, each displaying reliability and adaptivity as supported by error correction mechanisms, that is, if one can identify in the player two functions, each with a contribution to the system with the appropriate fineness of grain.

- The concept of affordance can be applied to any system for which the concept of ability, that is, function, has application; thus it extends beyondanimals. In particular it applies to systems displaying closure of constraints.

Field of Affordances

The previous definition of an affordance has explained how possession of an ability entails reliability and adaptivity. This does not fully address the difficulty with Moreno and Mossio's concept of action—the flexibility in selecting alternative means or in changing ends. However, as we shall see, the basis for an understanding of the flexibility of the agent is already laid down. To bring this out, it is helpful to contrast the present definition of affordance with an alternative one. One influential approach to affordances present them as drawing the animal towards engagement (or prompting the animal to recoil); in other words, the affordance, when perceived, is supposed to at least incline the animal towards actually performing the afforded behaviour. Thus Dreyfus and Kelly (2007) describes affordances as soliciting the animal: an affordance, once perceived, is a solictation to act. Yet our above account defines an affordance as purely a relation where such a relation, even when instantiated, need not imply any solicitation.

One might think that a theory of affordances which eliminates the motivational aspect of affordances would reduce its utility as a theory of action. Yet, on the present conception, the distinction between solicitation and affordance is important. First, it accords well with the realistic attitude taken towards affordances: affordances are there even when not noticed: the stairs are climb-able remains true, and the relation affords-climbing is instantiated, although the animal has no urge to climb the stairs. Instantiating the affordance relation simply means the affordance is available, not that it solicits. Second, this distinction allows us to grasp the source of the solicitation. On the present account, solicitation arises from the self-maintaining dynamic of the system, which drives the system towards engaging selectively with available affordances. Thus, for instance, the self-maintaining dynamic requires energy or some food source, which drives the system towards selectively engaging with relevant affordances of, say, foraging. Thus, strictly speaking, the affordance by itself does not solicit; it is the way the dynamics of self-maintenance is coupled to the environment that drives the system to be motivated to engage with relevant affordances.

Focusing on this dynamic aspect of affordances allows us to do justice to flexibility. A system possesses an interconnected system of abilities and thus faces not just one isolated affordance at a time, but as Rietveld and Kiverstein (2014) put it, a rich landscape of affordances which it has to selectively engage with. Any realistic system has to solve the task of selecting which affordances to engage with, and in what order and combination. For systems capable of some form of learning or even simply error minimisation through feedback, repeated engagement with sets and sequences of affordances might lead to preferred paths through the landscape of affordances. Indeed, this could lead to the development of new affordances. This in turn alters the landscape of affordances, making available yet further affordances. Thus, for sufficiently complex systems, the affordances available are not fixed; they co-evolve with the development of new abilities or functions of the system.

An adequate conceptualisation of the dynamical interaction between the system and the evolving landscape of affordances is provided by a cybernetic approach, proposed by a number of neuroscientists (Cisek, 2007; Cisek and Kalaska, 2010; Pezzulo and Cisek, 2016). The landscape of affordances can be formalised as follows as a state space. First, consider abilities. Each action to be generated by the system is specified by a number of action parameters (accounting for the fineness of grain); this space of parameters define the space of possible actions of, say, particular body parts. Potential actions are modelled as regions within this space (Pezzulo and Cisek, 2016). A region that the agent can reach and explore reliably and adaptively represents an ability of the agent. Second, although the features of the environment are placed in physical spacetime, the agent in the environment is best modelled not as a physical object in three dimensional physical space, but a system placed within this abstract space of affordances, where each affordance has an extra spatiotemporal location (recall the cybernetic setup of von Uexkull). Given this setup, at each moment, several affordances are salient and compete with each other to be executed. Such competition could simply involve mutual inhibition—engaging with one affordance inhibits engaging with others, and the affordances compete based on the benefits engaging in the activity has for the system as a whole at the time (Cisek (2007) calls this the 'affordance competition hypothesis). But one can envisage more complex ways of engaging with affordances. Pezzulo and Cisek (2016) thus introduce a hierarchical system supporting 'hierarchical affordance competition': competition occurs in parallel at each layer, and influence each other through top-down and bottom-up signals. Typically, in such hierarchical models, factors at the higher levels code more abstract features, and often represent predictions of a system about future environmental states that unfold at a longer timescale.

These higher levels predictions should not be understood as complete plans or models of the future. They are a "nested cascade of expectations or reference signals" of different levels of fineness of grain that are communicated to lower layers, thus influencing how the actual engagement with the environment is to be carried out. This makes possible the execution and planning of more abstract or temporally extended activity. In particular, this makes possible the planned creation of new affordances or destruction of existent affordances, as well as the prospective search for new affordances. It is the responsiveness to a whole field of affordances, and the possibility of navigating this field through the hierarchical competition that takes into account information about environmental features that unfold at various timescales, that allows one to see how flexibility becomes possible.

Let us take stock. The basic elements of the present framework is an agent located in an environment provided with a field of affordances, and the trajectory through such affordances, a trajectory driven by the disequilibrium characterising the dynamical self-maintenance of the agent. This provides the foundation for our attribution and explanation of action, as we shall now see.

4.3 Actions and Action Explanations Revisited

We started out with the problem of discovering the underlying conceptual structure of the attribution of actions to an agent, as well as the explanation of action. We now have all the concepts necessary to revisit the problem. As noted in the previous chapter, an agent is a system displaying closure of constraints. What about the attribution of an action to such an agent? It is tempting, I think, to identify an action with a certain event; after all, in attributing actions we do so on the basis of observations of the agent, and what we observe seems to be events. Yet this does not imply that actions are events; at most, this shows they are observable as events. Indeed, whether actions are events is not the key question. In the present theory, we take as primitive the existence of an agent who traces a trajectory through its field of affordances, a scenario described through a state space. Actions are particular parts of these trajectories. They are thus strictly speaking particular sets of states of the system. As discussed in chapter 3, states in general represent both structural properties of the system and tendencies of the system. Thus, one may conceptualise actions as manifestations of particular dispositions of the agent, where such manifestations come with intrinsic tendencies. I will set aside the question whether an action can technically be described as an event in this framework. The important fact is rather that the observable event indicates the existence of certain dispositions in the agent connected to the self-maintaining physical constitution of the agent.

In this framework, the fact that actions come with a certain standard of completeness reflected in the progressive can be understood as follows. We have previously accounted for the temporality of an agent by appealing to its specific self-maintaining dynamic: the system is in a present state of disequilibrium, which can only be minimised through an appropriate future trajectory. The standard of completeness arises from this as follows. Each action is a manifestation of a disposition which reflects the underlying state of disequilibrium. The particular state of disequilibrium driving the action defines the completeness: only by engaging with the field of affordances in a particular way can the disequilibrium of the system be appropriately minimised.

The same fact also explains the synthesis of two actions or activities implied in a naive action explanation. Such explanations apply to systems which face a field of affordances. Such systems have to combine various affordances flexibly in a manner described in the previous section in order to minimise disequilibrium. In other words, when one gives an explanation, 'X is doing A because it is doing B', one is viewing A and B as related in virtue of the fact that, first, X has the ability to do A and to do B, and second, X does A because it is attempting to minimise the disequilibrium in its present state, a minimisation that takes place through doing B.

Viewing action explanations in this light entail that such explanations are also naturalistic explanations. This can be seen most clearly if one borrows from the philosophical discussion concerning different kinds of explanations in science. Philosophers have noted the existence and importance of non-causal explanations besides causal, mechanistic explanations. One particularly important type of explanation, due to its generality, is explanation by constraint (Lange, 2016, chapter 2). The contrast between these two kinds of explanations is best illustrated via examples. Consider Lange's example: Why are gravitational and electric interactions alike in conserving energy? One explanation proceeds causally. A gravitational interaction, governed by laws of gravity, conserves energy; an electric interaction, governed by the laws of electricity, conserves energy as well. But another explanation proceeds via constraints: conservation of energy is a general constraint on natural processes, and both gravitational and electric phenomena fall under such a constraint. That such phenomena fall under such a constraint is not itself a causal fact.

Stated generally, the contrast between the two kinds of explanation is that in an explanation through a causal process, the explanation in general has the following form: a certain state is explained through the fact that it is brought about through a law of evolution given initial conditions. An explanation through constraints has another form: a certain state is explained through the fact that the actual state is taken from a space of possible states, where the existence of the constraint eliminates the other possibilities, leaving the actual state we observe. Both forms of explanation are valid; a case might even be made that explanations through constraints is more fundamental (given the importance of principles of stationary action, its extension through path integrals, etc.).

In the present framework, an action explanation is precisely an explanation through constraints. An explanation of the form, agent X did A/is doing A because it did/is doing B explains the observation of two events as observations of particular dispositions of the agent. It assumes the existence of a system with a self-maintaining dynamic, and explains the occurrence or manifestation of A as a result of the constraints on this dynamics imposed by B. As an explanation through constraint, it does imply the existence of a causal pathway that realises the constraint, although the teleological explanation by itself does not make explicit this pathway; nor would making explicit this pathway add to the explanatory force. Nevertheless, viewing teleological explanations as explanations through constraints makes the connection to the causal production of an action clear: a proper teleological explanation does not only give a possible end which an action fulfills, but points out the existence of constraints that eliminate other possibilities of action of the system, leading to the actual production of a particular action.

Conclusion

In this work we have attempted to sketch a naturalistic foundation for Thompson's theory of agency. We have interpreted the agent generally as any system possessing closure of constraints, and thus a system possessing functions and ultimately facing a field of affordances. An action is then understood as a manifestation of a disposition placed within the trajectory the system takes through the field of affordances. An action explanation draws out the fact that such a trajectory is selected on the basis of some principle of optimisation under constraints, and thus constitutes an explanation through constraints which at the same time imply the existence of a causal pathway leading to the production of the action.

Let us return to the starting point of this work—Thompson's theory as an incarnation of a broadly Aristotelian position. The above framework has incorporated Aristotle's theory of form, elucidating the role of this form in the attribution and explanation of action (points 1-4) in the introduction, in particular clarifying the relation between action explanations and the causal production of an action. This naturalistic understanding of Aristotle's theory raises two questions.

First, Aristotle's theory is more generally a theory of action/passion for all of nature. In the present work, following Thompson, we have restricted our attention to agents displaying some sort of purposiveness or animacy. However, there is no reason why the concept of action as a naturalistic phenomena cannot be extended beyond animate agents. A development of this idea would require investigation into the metaphysical nature of the physical world.

Second, we have not discussed the more complex forms of agency usually associated with systems possessing greater cognitive capacities as well as a richer social life. In particular, we have not been able to examine the enrichments of naive action explanations to psychological states, or to locate the role of such concepts as desire or belief within the present framework. As remarked, this connects to the difficult problem of the naturalisation of content, thereby linking the higher forms of cognition to the naturalistic basis of a flux of matter and energy sketched here.

It is however hoped that we have indicated a framework that provides a language applicable to these domains as well.

Bibliography

Anscombe, E. (2000). *Intention.* Cambridge: Massachusetts: Harvard University Press.

Anscombe, E. (2005). Has mankind one soul: an angel distributed through many bodies? In M. Geach and L. Gormally (Eds.), *Human Life, Action and Ethics: Essays by G.E.M. Anscombe*, St Andrews studies in philosophy and public affairs. Exeter, UK: Imprint Academic.

Aristotle (2020). *De anima.* Clarendon Aristotle series. Oxford: Oxford University Press.

Armstrong, D. M., C. B. Martin, U. T. Place, and T. Crane (1996). *Dispositions: a debate.* International library of philosophy. London: Routledge.

Ashby, W. R. (1957). *An Introduction to cybernetics.* London: Chapman and Hall.

Bansho, Y., T. Furubayashi, N. Ichihashi, and T. Yomo (2016). Host–parasite oscillation dynamics and evolution in a compartmentalized rna replication system. *Proceedings of the National Academy of Sciences - PNAS 113*(15), 4045–4050.

Bird, A. (2016). Overpowering: How the powers ontology has overreached itself. *Mind 125*(498), 341–383.

Bishop, R. C. (2019). *The Physics of Emergence.* London: Morgan and Claypool Publishers.

Bishop, R. C., M. Silberstein, and M. Pexton (2022). *Emergence in Context: A Treatise in Twenty-First Century Natural Philosophy* (1 ed.). Oxford: Oxford University Press.

Boyd, A. B., J. P. Crutchfield, M. Gu, and F. C. Binder (2024). Thermodynamic overfitting and generalization: Energetic limits on predictive complexity. *arXiv.org*.

BURGE, T. (2009). Primitive agency and natural norms. *Philosophy and phenomenological research 79*(2), 251–278.

Burge, T. (2022). *Perception: First Form of Mind*. Oxford: Oxford University Press.

Carlson, G. N. and F. J. Pelletier (1995). *The generic book*. Chicago: University of Chicago Press.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, Mass. ;: MIT Press.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical transactions of the Royal Society of London. Series B. Biological sciences 362*(1485), 1585–1599.

Cisek, P. and J. F. Kalaska (2010). Neural mechanisms for interacting with a world full of action choices. *Annual review of neuroscience 33*(1), 269–298.

Comrie, B. (1976). *Aspect : an introduction to the study of verbal aspect and related problems*. Cambridge textbooks in linguistics. Cambridge: Cambridge University Press.

Conant, R. and W. R. Ashby (2024). Every good regulator of a system must be a model of that system. In D. C. Krakauer (Ed.), *Foundational Papers in Complexity Science*, Volume 2, pp. 1061–1074. SFI Press.

Cover, T. M. and J. A. Thomas (2006). *Elements of information theory* (2nd ed. ed.). Hoboken, N.J: Wiley-Interscience.

Curto, C., V. Itskov, A. Veliz-Cuba, and N. Youngs (2013). The neural ring: An algebraic tool for analyzing the intrinsic structure of neural codes. *Bulletin of mathematical biology 75*(9), 1571–1611.

Davidson, D. (1963). Actions, reasons and causes. In *Problems of Rationality*, pp. 12–24. Oxford: Oxford University Press.

Dreyfus, H. and S. D. Kelly (2007). Heterophenomenology: Heavy-handed sleight-of-hand. *Phenomenology and the cognitive sciences 6*(1-2), 45–55.

Dupleix, S. (1626). *La Physique, ou science des choses naturelles.* Paris: Claude Sonnius.

Durand, P. M. (2021). *The Evolutionary Origins of Life and Death.* Chicago: University of Chicago Press.

Eigen, M. and P. Schuster (1977). A principle of natural self-organization: Part a: Emergence of the hypercycle. *Die Naturwissenschaften 64*(11), 541–565.

Eigen, M. and P. Schuster (1978a). The hypercycle: A principle of natural self-organization part b: The abstract hypercycle. *Die Naturwissenschaften 65*(1), 7–41.

Eigen, M. and P. Schuster (1978b). The hypercycle: A principle of natural self-organization part c: The realistic hypercycle. *Die Naturwissenschaften 65*(7), 341–369.

Eustache de Saint Paul (1609). *Summa philosophiae quadripartita: De Rebus dialecticis, moralibus, physicis et metaphysicis. I-II.* Paris: C. Chastelain.

Foot, P. (2001). *Natural goodness.* Oxford: Clarendon Press.

Furubayashi, T., K. Ueda, Y. Bansho, D. Motooka, S. Nakamura, R. Mizuuchi, and N. Ichihashi (2020). Emergence and diversification of a host-parasite rna ecosystem through darwinian evolution. *eLife 9*, 1–15.

Ganti, T. (2003a). *Chemoton Theory: Volume 1: Theoretical Foundations of Fluid Machineries.* New York: Kluwer Academic.

Ganti, T. (2003b). *Chemoton Theory Volume 2: Theory of Living Systems.* New York: Kluwer Academic.

Ganti, T. (2003c). *The principles of life.* Oxford: Oxford University Press.

Gardenfors, P. (2014). *Geometry of meaning: semantics based on conceptual spaces.* Cambridge, Massachusetts: The MIT Press.

Garson, J. (2016). *A Critical Overview of Biological Functions.* Cham: Springer.

Garson, J. (2017). Against organizational functions. *Philosophy of science 84*(5), 1093–1103.

Gelman, S. A. (2003). *The essential child : origins of essentialism in everyday thought / Susan A. Gelman.* Oxford series in cognitive development. Oxford: Oxford UP.

Ghiselin, M. T. (1997). *Metaphysics and the Origin of Species.* Albany: State University of New York.

Gibson, J. J. (2015). *The ecological approach to visual perception* (Classic edition. ed.). Psychology Press classic editions. New York: Psychology Press.

Godfrey-Smith, P. (2009). *Darwinian populations and natural selection.* Oxford: Oxford University Press.

Godfrey-Smith, P. (2013). Darwinian individuals. In *From Groups to Individuals*, pp. 17–. The MIT Press.

Godfrey-Smith, P. (2016). Individuality, subjectivity, and minimal cognition. *Biology and philosophy 31*(6), 775–796.

Hanczyc, M. M. and T. Ikegami (2010). Chemical basis for minimal cognition. *Artificial life 16*(3), 233–243.

Hannon, E. and T. Lewens (2018). *Why we disagree about human nature* (First edition. ed.). Oxford scholarship online. Oxford: Oxford University Press.

Hattab, H. (2009). *Descartes on forms and mechanisms.* Cambridge: Cambridge University Press.

Heft, H. (2003). Affordances, dynamic experience, and the challenge of reification. *Ecological psychology 15*(2), 149–180.

Heider, F. and M. Simmel (1944). An experimental study of apparent behavior. *The American journal of psychology 57*(2), 243–259.

Hordijk, W. (2013). Autocatalytic sets: From the origin of life to the economy. *Bioscience 63*(11), 877–881.

Hölldobler, B. and E. O. Wilson (2009). *The superorganism: the beauty, elegance, and strangeness of insect societies.* New York: W.W. Norton and Company.

Joyce, G. F. (2007). Forty years of in vitro evolution. *Angewandte Chemie (International ed.) 46*(34), 6420–6436.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering 82*(Series D), 35–45.

Kauffman, S. A. (1993). *The origins of order: self-organization and selection in evolution.* New York: Oxford University Press.

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

Korsgaard, C. M. (2018). *Fellow creatures: our obligations to the other animals* (First edition. ed.). Uehiro series in practical ethics. Oxford: Oxford University Press.

Korsgaard, C. M. C. M. (2009). *Self-constitution : agency, identity, and integrity / Christine M. Korsgaard.* Oxford scholarship online. Oxford: Oxford University Press.

Kuznetsov, P. I., R. L. Stratonovich, and V. I. Tikhonov (1965). *Non-linear transformations of stochastic processes* (1st ed. ed.). Oxford: Pergamon Press.

Lagerspetz, K. Y. H. (2001). Jakob von uexküll and the origins of cybernetics: Jakob von uexküll: A paradigm for biology and semiotic. *Semiotica 134*(1-4), 643–651.

Lange, M. (2016). *Because without cause: non-causal explanations in science and mathematics.* Oxford studies in philosophy of science. New York, NY: Oxford University Press.

Lavin, D. and M. Boyle (2010). Desire and the good. In S. Tenenbaum (Ed.), *Desire, practical reason and the good.* Oxford: Oxford University Press.

Lewin, K. (1926). Vorsatz, wille und bedurfnis. *Psychologische Forschung 7*, 330–385.

Lewontin, R. C. (1970). The units of selection. *Annual review of ecology and systematics 1*(1), 1–18.

Maddy, P. (2022). *A Plea for Natural Philosophy: And Other Essays* (1 ed.). Oxford: Oxford University Press, Incorporated.

Martin, C. B. (2008). *The Mind in Nature*. Oxford: Oxford University Press.

Maturana, H. R. and F. J. Varela (1980). *Autopoiesis and Cognition: the Realization of the Living*. Boston Studies in the Philosophy of Science, 42. Dordrecht: Springer Netherlands.

Mele, A. (2010). Teleological explanations of actions: Anticausalism versus causalism. In J. H. Aguilar and A. A. Buckareff (Eds.), *Causing Human Actions: New Perspectives on the Causal Theory of Action*, pp. 183–198. The MIT Press.

Merleau-Ponty, M. (1995). *La Nature. Notes Cours au College de France.* Paris: Editions du Seuil.

Michaels, C. F. (2003). Affordances: Four points of debate. *Ecological psychology 15*(2), 135–148.

Michotte, A. (1963). *The perception of causality.* Methuen's manuals of modern psychology. London: Methuen.

Millikan, R. G. (1984). *Language, thought, and other biological categories: new foundations for realism.* Cambridge, Mass: MIT Press.

Mizuuchi, R., T. Furubayashi, and N. Ichihashi (2022). Evolutionary transition from a single rna replicator to a multiple replicator network. *Nature communications 13*(1), 1460–1460.

Montévil, M. and M. Mossio (2015). Biological organisation as closure of constraints. *Journal of theoretical biology 372*, 179–191.

Moreno, A. and M. Mossio (2015). *Biological Autonomy: a Philosophical and Theoretical Investigation.* Berlin: Springer.

Mossio, M. and L. Bich (2017). What makes biological organisation teleological? *Synthese (Dordrecht) 194*(4), 1089–1114.

Mossio, M., C. Saborido, and A. Moreno (2009). An organizational account of biological functions. *The British journal for the philosophy of science 60*(4), 813–841.

Myers, D. J. (2022). Categorical systems theory. Manuscript.

Neander, K. (2017). *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge, MA.: MIT Press.

Nishimori, H. and G. Ortiz (2011). *Elements of phase transitions and critical phenomena*. Oxford graduate texts. Oxford: Oxford University Press.

Nolte, D. D. (2018). *Galileo Unbound: A Path Across Life, the Universe and Everything* (First edition ed.). Oxford: Oxford University Press.

Opfer, J. E. (2002). Identifying living and sentient kinds from dynamic information: the case of goal-directed versus aimless autonomous movement in conceptual change. *Cognition 86*(2), 97–122.

Pezzulo, G. and P. Cisek (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in cognitive sciences 20*(6), 414–424.

Phillips, R. and N. Orme (2020). *The molecular switch: signaling and allostery*. Studies in physical biology. Princeton, New Jersey ;: Oxford Princeton University Press.

Quarteroni, A., A. Manzoni, and C. Vergara (2017). The cardiovascular system: Mathematical modelling, numerical algorithms and clinical applications. *Acta numerica 26*, 365–590.

Rietveld, E. and J. Kiverstein (2014). A rich landscape of affordances. *Ecological psychology 26*(4), 325–352.

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

Rothstein, S. (2016). Aspect. In M. Aloni and P. Dekker (Eds.), *The Cambridge Handbook of Formal Semantics*, Cambridge Handbooks in Language and Linguistics, pp. 342–368. Cambridge: Cambridge University Press.

Ruler, J. A. v. (1995). *The crisis of causality: Voetius and Descartes on God, nature, and change.* Brill's studies in intellectual history, volume 66. Leiden ;: E.J. Brill.

Saborido, C., M. Mossio, and A. Moreno (2011). Biological organization and cross-generation functions. *The British journal for the philosophy of science 62*(3), 583–606.

Schrödinger, E. (1944). *What is life? The physical aspect of the living cell.* Canto classics. Cambridge: Cambridge University Press.

Searle, J. (1959). Determinables and the notion of resemblance. *The Aristotelian Society Supplement 33*, 141–158.

Sehon, S. (2016). *Free Will and Action Explanation: a Non-Causal, Compatibilist Account.* Oxford: Oxford University Press.

Sellars, W. (1963). Philosophy and the scientific image of man. In *Science, Perception and Reality.* Ridgeview Publishing Company.

Simondon, G. (2013). *Cours sur la perception.* Paris: Presses Universitaires de France.

Smith, E. and H. J. Morowitz (2016). *The origin and nature of life on Earth : the emergence of the fourth geosphere.* Cambridge: Cambridge University Press.

Sobolev, S. L., A. I. Kitov, and A. A. Lyapunov (1955). Main features of cybernetics. *Problems of Philosophy*, 136–148.

Spelke, E. (2022). *What babies know : core knowledge and composition. Volume 1.* Oxford cognitive development series. New York: Oxford University Press.

Spelke, E. S. and K. D. Kinzler (2007). Core knowledge. *Developmental science 10*(1), 89–96.

Still, S. (2020). Thermodynamic cost and benefit of memory. *Physical review letters 124*(5), 050601–050601.

Still, S., D. A. Sivak, A. J. Bell, and G. E. Crooks (2012). Thermodynamics of prediction. *Physical review letters 109*(12), 120604–120604.

Stoffregen, T. A. (2003). Affordances as properties of the animal-environment system. *Ecological psychology 15*(2), 115–134.

Szathmáry, E. and J. M. Smith (1995). The major evolutionary transitions. *Nature (London) 374*(6519), 227–232.

Thompson, M. (2008). *Life and action : elementary structures of practice and practical thought.* Cambridge, Mass: Harvard University Press.

Turvey, M., R. Shaw, E. Reed, and W. Mace (1981). Ecological laws of perceiving and acting: In reply to fodor and pylyshyn (1981). *Cognition 9*(3), 237–304.

Varela, F., H. Maturana, and R. Uribe (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems 5*(4), 187–196.

Wei, F., T. Zhong, Z. Zhan, and L. Yao (2021). Self-assembled micro-nanorobots: From assembly mechanisms to applications. *ChemNanoMat : chemistry of nanomaterials for energy, biology and more 7*(3), 238–252.

Wharton, D. A. (2002). *Life at the limits: organisms in extreme environments.* Cambridge: Cambridge University Press.

Wiener, N. (1961). *Cybernetics or Control and Communication in the Animal and the Machine.* The MIT Press. Place of publication not identified: The MIT Press.

Wilson, D. S. and E. Sober (1989). Reviving the superorganism. *Journal of theoretical biology 136*(3), 337–356.

Wilson, J. (2023). Determinables and Determinates. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 ed.). Metaphysics Research Lab, Stanford University.

Zachos, F. E. (2016). *Species Concepts in Biology.* Cham: Springer.