Norm-establishing and norm-following in autonomous agency

Xabier E. Barandiaran[1*], Matthew D. Egbert[2]
[1] IAS-Research Centre for Life, Mind, and Society
Dept. of Logic and Philosophy of Science
UPV/EHU University of the Basque Country, Spain
*xabier.academic@barandiaran.net
[2] Center for Computational Neuroscience and Robotics
University of Sussex, Brighton, U.K.

June 25, 2012

Abstract

One of the fundamental aspects that distinguishes *acts* from mere *events* is that actions are subject to a normative dimension that is absent from other types of interaction: natural agents behave according to intrinsic norms that determine their adaptive or maladaptive nature. We briefly review current and historical attempts to naturalize normativity from an organism-centred perspective that conceives of living systems as defining their own norms in a continuous process of self-maintenance of their individuality. We identify and propose solutions for two problems of contemporary modelling approaches to viability and normative behaviour in this tradition: 1) How to define the topology of the viability space beyond establishing normatively-rigid boundaries, so as to include a sense of gradation that permits reversible failure; and 2) How to relate, in models of natural agency, *both* the processes that establish norms and those that result in norm-following behaviour. In this paper we present a minimal metabolic system that is coupled to a gradient climbing chemotactic mechanism. Through studying the relationship between metabolic dynamics and environmental resource conditions, we identify an emergent *viable region* and a *precarious region* where the system tends to die if environmental conditions remain the same. We introduce the concept of *normative field* as the change of environmental conditions required to bring the system back to its viable region. *Norm-following*, or *normative action*, is defined as the course of behaviour whose effect is positively correlated with the normative field. We generalize from our minimal model to a wider range of theoretical and modelling approaches, and close with some discussion about the limitation of our model, possible extensions and

some final reflections around the nature of norms and teleology in agency.

Contents

1 Introduction

The issue of biological agency and natural norms is attracting increasing attention (Frankfurt, 1978; Burge, 2009; Di Paolo, 2005; Skewes and Hooker, 2009; Barandiaran et al., 2009; Silberstein and Chemero, 2011; Rietveld, 2008; Kauffman, 2000, 2003; Enoch, 2006) and Artificial Life is very well suited to make some conceptual progress on key aspects of agency and its origins. In fact, minimal models of agency have been a recurring topic in the field, from protocellular models to robotics—see Barandiaran et al. (2009) for references.

From bacteria to humans, the way in which living systems actively regulate their relationship with their environments strongly contrasts with inanimate objects. Whereas the behaviour of living systems generally qualifies as agency, the inanimate world is a world of events. Part of this distinction between actions and events lies on the normative character of the former. Actions are subject to a *normative judgement*: they are either good or bad, appropriate or inappropriate, adaptive or maladaptive. A bacterium might be described as failing to move up a sugar gradient, but it makes little (if any) sense to say that a planet has failed to follow an orbit. This normative dimension of *agency* is widespread in nature and it continues to capture the attention of philosophers, theoretical biologists, psychologists and roboticists alike, for it has proven to be a difficult property to define, naturalize, model or synthesise.

In a recent paper, Barandiaran et al. (2009) described normativity as one of the essential features of an agent which is defined as: "an autonomous organization capable of adaptively regulating its coupling with the environment according to the norms established by its own viability conditions." (Barandiaran et al., 2009, p. 376). The main goal of this paper is to make explicit what is meant by the expression "according to the norms established by its own viability conditions". Similar expressions have been used by Kauffman (2003); Bourgine and Stewart (2004); Di Paolo (2005); Barandiaran and Moreno (2008); Skewes and Hooker (2009) but no model has yet been developed to illustrate and describe in detail the meaning of this expression and others closely associated with it.

We shall start by introducing the problem of normativity and agency in more detail and identifying the two most prominent theories that are currently available for naturalizing a notion of proper or non-derived normativity: the evolutionary and the organizational (or organismic) accounts of normative function. We then focus on those developments of the organizational approach that have had more impact in the fields of cybernetics, artificial life and adaptive behaviour. Having set up the context and identified some key problems of previous approaches we describe our minimal model. The model covers "norm-establishing" and "norm-following" aspects. First we describe a minimal metabolic system that allows us to define *norm-establishing* at the protocellular scale. A parametric analysis of changes in environmental conditions of metabolism allows us to define and quantify the notions of "precariousness", "viability" and "normativity"; in terms of regions inside the viability space. Next, we move into *norm-following* by adding a chemotactic behavioural mechanism to the metabolic system just described. By projecting the effect of chemotactic behaviour

into the viability space we can precisely define "normative behaviour" and characterize an adaptive region that emerges from the relational dynamics between agent and environment. Then, we provide a set of generalized definitions that are applicable beyond our model. We end up with some possible extensions of our model and some discussion about the nature of norms and teleology in agency.

2 Agency, adaptivity and normativity

2.1 The framework: events vs actions, descriptive vs normative and the challenge of natural norms

In a world of events that science describes as governed by necessary laws and principles, the concept of *agency* can appear in tension with the very method and assumptions of science. The notion of agency has been traditionally (Davidson, 1963) formulated as behaviour caused or motivated by reasons, desires or intentions and as opposed to just antecedent events: to run home so as to protect yourself from a storm is not the same as being pushed by the wind. In a world of fundamental particles governed by laws there seems to be no room for agency; no way to formulate a concept of what processes or interactions *ought-to-be*, or are *intended-to-be* as different from what they *actually* are; no room for norms that can be violated such that an agent might be said to *fail*. And yet we have the experience of acting, of following or *failing* to follow certain norms (be them biological, psychological or social).

Ultimately, filling in the gap between fundamental physics and our own experience of agency might require to include concepts concepts such as "will", "awareness" or "reasons" into the explanatory picture. But, minimal forms of natural behaviour (like chemotaxis) encapsulate some of the most important properties of "higher" levels of agency (Barandiaran et al., 2009) and their explanation can arguably be considered a pre-requisite to ground and naturalize the very notion of agency and its more complex manifestations. It has long been argued for the continuity between biological and human agency (Frankfurt, 1978; Kauffman, 2000; Burge, 2009; Skewes and Hooker, 2009; Silberstein and Chemero, 2011; Bickhard, 2009; Ruiz-Mirazo and Moreno, 2012) claiming that human agency is a natural complexification of lower forms of agency. Where "natural complexification" implies no fundamental discontinuity nor the addition of essentially different principles. Without explicitly attaching to a strong continuity thesis, however, it seems reasonable to assume that the modelling and conceptualization of biological agency (even of the simplest forms) constitutes a valuable conceptual and methodological resource to make progress towards higher forms of agency. Moreover, the case of biological agency seems more amenable to scientific experimentation, modelling and formalization than psychological, ethical or social forms. It therefore provides a solid departure point to investigate the nature of agency.

One of the key properties of agency that shows up across different levels is *normativity*, the dimension of behaviour in which *value* comes into play. Actions are good or bad, adaptive or maladaptive, appropriate or inappropriate (Christensen and Bick-

hard, 2002; Barandiaran and Moreno, 2008; Burge, 2009). The issue of normativity has several manifestations and a long philosophical and theoretical history. One of its earlier formulations comes from David Hume who noted the "impossibility" of deriving *ought* from *is*, *prescriptive* statements from *descriptive* ones. There seems to be no justified manner in which (no matter how many) statements about "X being or operating-as A" can lead to a valid inference about "X having to-be or having to-operate-as A". This has important consequences for the fixation of epistemological norms (those concerning how knowledge or truth *must* be achieved or justified) and ethical norms (how we *ought* to behave). Immanuel Kant provided an additional decisive contribution: he proposed that epistemological and ethical norms where to be found and justified as *conditions of possibility*. The Humean problem could thus be solved: it is not that norms about how X *should be* are to be derived of inferred from what X *actually is* but, instead, from *what makes possible the very existence* of X[1]. In this way the normative shows itself not as an empirical fact, but as what makes possible that we apprehend empirical phenomena or ascertain reasons to act in the first place.

The issue of norms had an enormous development in the areas of epistemology and ethics since then but this is out the scope of the present paper. What is relevant to us is the naturalist turn that some of these developments took, in contrast to those trends that argued merely in terms of abstract reasons or the very structure of logic or language use. Naturalism tries to bring philosophical problems down to the grounds of natural sciences (Quine, 1969): i. e. something that can be measured. And there is an empirical manifestation of epistemic (and ethical) normativity: behaviour. So, if we are to assume that normativity does not belong to a supernatural sphere, then the *behaviour* of living systems might be a good entry point for trying to explain and model it. Of course, moving into the realm of biology does not automatically solve the theoretical challenges posed by normativity, but it does re frame the questions: when we model the mechanisms of natural behaviour and they lead to a specific outcome, how can we justify the claim that the animal has *failed* or that it *should* had done something else? Where do the norms against which natural behaviour is evaluated or judged come from? How do they relate to behaviour generating mechanisms? Can new norms appear in nature? How does the *actual functioning*, here-and-now, of a behavioural mechanism differ from how it *ought to function*? Is it possible to address these questions in a manner that is centred on the organism and not on aspects of its evolutionary history or on other agent's intentions?

2.2 Intrinsic and derived norms in natural and artificial systems

While artificial systems can be judged to operate in relation to norms, these norms have (thus far) always been pre-defined by the designer of the artificial system or interpreted by an external observer or user. In other words, what is good or bad

[1] Some contemporary approaches to social norms (Apel, 1980; Habermas, 1990) might help to make this philosophical point more clear: "do not lie" is a social norm for linguistic behaviour because being "truthful" is a condition of possibility of linguistic communication, or inversely, if everybody was to lie all the time it would be impossible to communicate.

functioning for a robot, a car or a coffee machine has been a matter of the design specifications which are largely independent from the ongoing operation of the artefact. In a sense we can talk of *derived* or *extrinsic normativity*: it is not the artefact that posses or defines any purpose on and by itself but always in relation to some external agent or social context from which such a purpose or norm is derived. So, for instance, the decision of when and how to fix the artefact (i. e. what counts as broken) depends on the design specifications of the designer, the social context of use or simply the pragmatic intentions of the end user, but not on any intrinsic property of the system itself. The same physical object can be either broken if intended to be used for a particular purpose or perfectly fine in a different context. This is unlike biological organisms that seem to "respond" to norms that are more closely related to their organization.

There is a possible analogy between the norms that a designer or context imposes upon an artefact and the norms that natural evolutionary "design" might impose upon organisms. And this analogy has inspired what currently stands as the mainstream approach to naturalizing norms. The most celebrated exponent of the so called *selected-function* or *evolutionary approach* is Millikan (1989). She defends a whole philosophical program to naturalize biological, cognitive and epistemological norms in evolutionary theory. Roughly speaking, behaviour is here considered to be normative or adaptive if it has been selected by evolution. Under this view adaptation is ultimately a result of natural selection, and it is only as a result of a process of selection that a character or process (e. g. a pattern of behaviour) can be said to be adaptive or maladaptive. What sets up the norm is the history of population dynamics. In this view an organ or a particular behavioural mechanism might be said to be broken, be maladaptive or to fail to carry out its proper function if it is currently working in a manner that was not the mode of functioning that evolution selected-for. Artificial Life models have contributed to clarify and quantify this notion of normativity and teleology (Bedau and Norman, 1996)[2].

An alternative conception of normativity that is *intrinsic* or *non-derived* can be developed with criteria that are independent from history, or the social context of design and use, and are instead grounded on the current organization of the system and its ongoing dynamics. This is precisely the motivation underlying the main alternative approach to normativity and adaptation. The *organizational approach* (as it might be called) puts the idea of autonomy (from the Greek *autos* = self, and *nomos* = norm) at its centre (Varela, 1979; Ruiz-Mirazo and Moreno, 2004; Di Paolo, 2004; Barandiaran and Ruiz-Mirazo, 2008). This approach sees the norms of a system as being determined by its present organization. This results in an inversion of the selected-function approach: a trait does not acquire a normative function because it has been selected, but it has been selected because it came to fill a normative function within a living system's organization.

[2] A detailed critique of evolutionary accounts of function, purpose, teleology or intentionality, and its relation to *intrinsic* or *non-derived* purpose can be found elsewhere (Christensen and Bickhard, 2002; Mossio et al., 2009).

2.3 The organismic approach to biological autonomy: a short historical introduction

The origins of this organism-centred or autonomous-organizational approach can be traced back to the works of Aristotle and, particularly, to Kant (*Critique of Judgement*) and his account of teleology as emerging from the self-organization (a term that is due precisely to him) of living systems. This trend of thought was further developed along German idealism and the so called *naturphilosophie* school, but it was not until the late XIX century that it started to take a more detailed empirical taste, with the rise of cell theory and experimental physiology—e.g., the work of Claude Bernard and his conception of the *milieu interieur* (Bernard, 1865). We can observe an influential integration of this organismic trend in biology in the work of Walter B. Cannon and his development of the concept of *homeostasis* to name the "coordinated physiological processes which maintain most of the steady states in the organism" (Cannon, 1932).

But it is perhaps the work of Canguilhem (Canguilhem, 1966) that most directly addressed the issue of normativity in biology. As noted by Gayon (1998), Canguilhem further develops some of the central ideas already anticipated in Kurt Goldstein's masterpiece *Der Aufbau des Organismus* where the issue of normativity is directly linked to the individual organism: "there is only one relevant norm; that which includes the total concrete individuality; that which takes the individual as its measure" (Goldstein, 1934, p. 269) [3].

Canguilhem rejects a statistical notion of *normality* and the idea that the deviation from a populational mean is a valid departure point for approaching pathologies in organisms. Instead, he introduces the notion of normativity at the very core of the living individual in a sense that connects directly with the contemporary debate on norms:

> [L]ife is in fact a normative activity. The normative, in philosophy, includes every judgment which evaluates or qualifies a fact in relation to a norm, but this mode of judgment is essentially subordinate to that which establishes norms. The normative, in the fullest sense of the word, is that which establishes norms. And it is in that sense that we plan to talk about biological normativity. (Canguilhem, 1991, p. 126-127)

A fundamental distinction introduced by Canguilhem is that between *norm-establishing* and *norm-following* (Rand, 2011). Organism do not only operate according to norms, i. e. they do not only seem to follow norms, but they also establish these norms autonomously. How exactly this occurs (and how it contrasts with the processes described by physics) was a central contribution of Hans Jonas' philosophical biology. Metabolism defines the organism as an individual that escapes the determination of its constituent parts by renewing them continuously, leaving the *form* or organization of the system as the persistent reference identity. The more complex it is, the more precarious and the more threatened by the environment a living system becomes; and yet,

[3] Quote borrowed from Gayon (1998).

the more in *need* of its environment. To *do* is both the consequence and the condition of possibility of the very existence of an organism.

> "To be" is its intrinsic goal. Teleology comes in where the continuous identity of being is not assured by mere inertial persistence of a substance, but is continually executed by something done, and by something which has to be done in order to stay on at all: it is a matter of to-be-or-not-to-be whether what is to be done is done. (Jonas, 1968, p. 243)

In this way, Jonas shows how individuality and normativity get intertwined: norms apply to natural identities which, unlike inanimate objects, depend on their own, continuous action of metabolic renewal and resource seeking for their ongoing existence. This recursive nature of living organization allows to derive the *norms*, not directly from isolated *events* (as the problem was originally posed by Hume), but from the set of organized and interdependent processes that constitute an organism (i.e. its conditions of possibility).

These arguments suggest inspiring and convincing research avenues, but there remains a large gap between this philosophy and scientific modelling and experimentation. Notions of "identity", "precariousness", "preservation" or "need" require far more detailed development to be consistently ensambled to deliver a clear scientific notion of normative action. It is through the relatively modern development of cybernetics, theoretical biology and the physics and chemistry of far-from-equilibrium systems that this tradition is progressively entering, albeit partially and not without difficulties, into a proper scientific treatment of normative action.

A branch of contemporary *theoretical biology*, under the labels of autopoietic theory or enactivism (Maturana and Varela, 1980; Varela, 1979; Varela et al., 1991; Bourgine and Stewart, 2004; Di Paolo, 2005) and autonomous systems (Kauffman, 2000; Christensen and Hooker, 2000; Collier, 2000; Ruiz-Mirazo and Moreno, 2000, 2004; Bechtel, 2007; Barandiaran and Moreno, 2008) have further elaborated some of the central tenets of the organicist school, connecting them with contemporary biology, artificial life and cognitive science. Of particular relevance is the work by Christensen and Bickhard (2002), who provide a detailed philosophical discussion around the notion of normative function within the context of contemporary philosophy of biology and cognitive science–see also Mossio et al. (2009) for a more recent version of "normative function" in biology. The central claim of these developments is that the normative nature of a biological functions stems from the far-from-equilibrium and self-sustaining nature of the networked processes that constitute living organization.

However, putting this theoretical approaches to work in scientific modelling requires bridging the gap that separates conceptual and philosophical contributions from mathematical and computational modelling. The goal of this paper is to bring this organicist development of the notion of normativity to a formulation that can shed light on some of its most intricate problems.

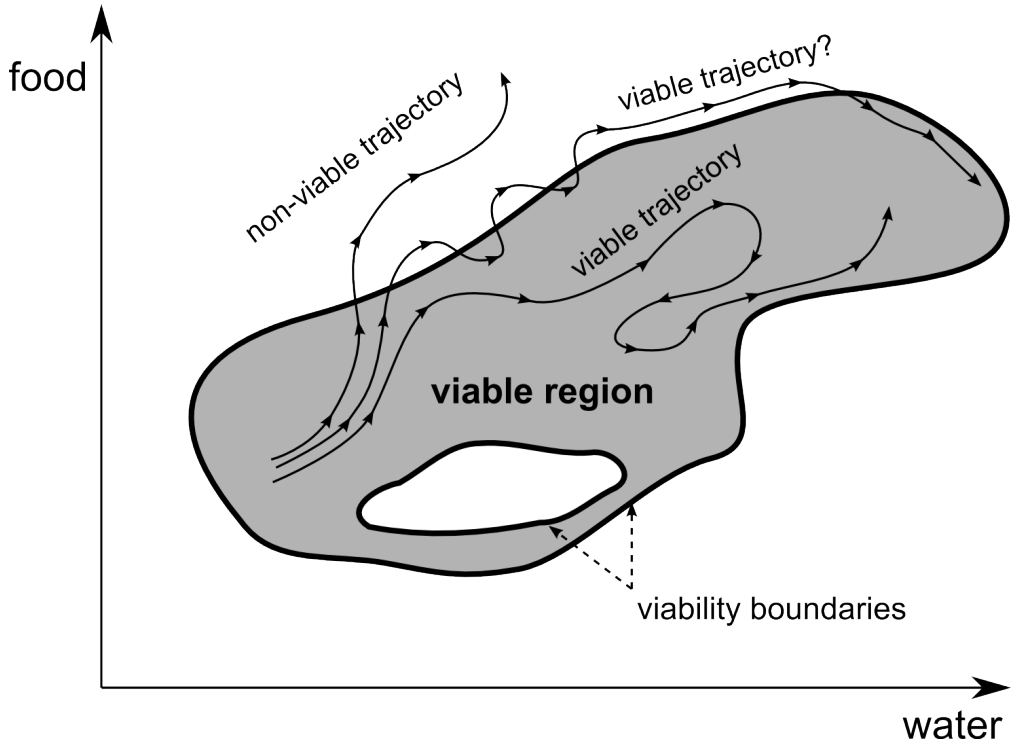

Figure 1: Standard representation of viability space: A viable region (grey area) and adaptive behaviour (trajectories) are illustrated for two essential variables (food and water). Outside the viable region the system will die. Viable trajectories are those that remain within the viability boundaries.

2.4 Adaptive behaviour and viability: contemporary theories and modelling frameworks

The contemporary conception of the organisational approach contends that norms are to be found as conditions of viability of a system, conditions of self-maintenance or closure-conditions. Examples and theoretical tools are often drawn from the adaptive behaviour literature where such norms as pictured as "viability constraints" or "viable region(s)" within a "viability space" defined by different physiological variables (see Figure 1). Ashby's influential work (Ashby, 1952) contributed to the formalization of this framework. Fist, he translated into the concept of "essential variables" what Claude Bernard called "internal milieu" or Walter Cannon's "homeostatic variables". Within the multidimensional space defined by those variables he termed "viable or homeostatic region" the subspace where the system was to remain alive or viable.

This way of conceiving and formalizing adaptivity had considerable influence. In ethology, animal behaviour has been modelled as responding to the maintenance of these essential variables (physiologically measured or measurable: e. g. temperature,

food or hydration levels) within their viability boundaries (McFarland, 1999). A similar approach is taken by Beer's research in adaptive behaviour, where he defines a system's behaviour as adaptive "only so long as it succeeds in maintaining its trajectory within this viability constraint, that is, only so long as it succeeds in maintaining the conditions necessary for its continued existence" (Beer, 1997, p. 266). Abstractions of these conditions (or functionally specified boundaries of viability) have been used to design robot controllers (Bersini, 1994). And a detailed formal development of some of these issues is due to Jean-Pierre Aubin's *viability theory*—see Aubin et al. (2011) for the latest version.

2.5 Current problems with modelling viability boundaries and the nature of norms

In terms of modelling, the organizational or organismic account of natural norms has translated into the notion of boundaries of viability (following the work of Ashby). The fundamental norm is "do not cross the boundary", the viability constraints *have* to be respected. The origin of these boundaries is given by chemical rates (energetic, kinetic, etc.) and physical constraints and their interdependences in metabolism and, for more complex multicellular organism, by the physiological conditions of recursive self-maintenance of the organism, such as blood circulation and respiration[4].

There is however a problem with how this notion of boundaries of viability have been modelled and conceptualized so far:

> **1. The problem of boundaries** If the notion of viability constraint or boundary is *normatively-rigid* (meaning the system is dead once crossed) then the norm is all-or-nothing and there is no room for reversible failure, no gradation. If the boundary is not rigid, then how can we quantify its "flexibility"? Are there alternative conceptions of the viability space that can accommodate gradation, intermediate spaces or wide "boundaries" before an irreversible state is reached? How can we quantify and qualify a richer topology of the viability space?

Ezequiel Di Paolo has identified a similar problem of the autopoietic theory and he attempted to solve it providing what remains perhaps the most complete and explicit of current definitions of adaptivity (Di Paolo, 2005):

> a system's capacity, in some circumstances, to regulate its states and its relation to the environment with the result that, if the states are sufficiently close to the boundary of viability, 1. tendencies are distinguished and acted upon depending on whether the states will approach or recede from the

[4] So, for example: imagine that above 57°C reaction rates increase so that waste products cannot be released out at sufficient speed and the cell bursts. The chemical dynamics of the system define this limit of 57°C (should the type and rate of the constituent metabolic reactions be different the system would have had a different boundary of viability). So the behaviour of the organism is adaptive or normative if it avoids its physiological processes move beyond the 57°C boundary of viability.

> boundary and, as a consequence, 2. tendencies of the first kind are moved closer to or transformed into tendencies of the second and so future states are prevented from reaching the boundary with an outward velocity. (Di Paolo, 2005, p. 436)

The issue of normativity can be explicitly re-framed under this definition as the requirement to "prevent" those tendencies of essential variables that might approximate the boundaries "with an outward velocity". There remain however some ambiguities in Di Paolo's definition that prevent a straight-forward quantification and measurement of normative action: how much is "sufficiently close" to the boundaries of viability so that a regulation becomes adaptive? How are "velocities" and "tendencies" defined and modelled? Are there different types of boundaries within the viability space? Are cases like the trajectory that crosses repeatedly the boundary of viability in Figure 1 possible?

The crucial contribution of Di Paolo's definition is that it introduces a dimension of temporality and velocity into the notion of adaptivity and norms that deserves more attention than what previous conceptions of viability have dedicated. Essential variables have tendencies or velocities in relation to which normative action has to be defined. Normativity is not only a question of boundaries or constraints that need just to be respected but, so to speak, a question of forces and tendencies that have to be compensated for.

In addition, most of the models that have been previously developed suffer a different type of problem (yet somewhat related to the previous one):

> **2. The problem of dissociation between norm-establishing and norm-following processes**: Existing models of viability appear dissociated into two categories: i) those where viability boundaries appear as given or defined from without and where the models focus on how to shape adaptive dynamics to maintain the trajectories of essential variables within those boundaries; and ii) those where a set of viability conditions emerge out of metabolic or physiological process but do not address behaviour in relation to them. As a consequence, the organismic dynamics that define viability and the dynamics that control adaptive behaviour remain dissociated in existing models.

This problem is rarely acknowledged in the literature but even when recognized it is assumed as a necessary pragmatic compromise and a valid assumption. So, for instance, Beer states that:

> [T]his explicit separation between an animals behavioural dynamics and its viability constraints is fundamentally somewhat artificial. (...) However, if we are willing to take the existence of an animal for granted, at least provisionally, then we can assume that its viability constraint is given a priori, and focus instead on the behavioural dynamics necessary to maintain that existence. (Beer, 1997, p. 265)

To be fair, this is a reasonable simplification to be done when the main focus of interest is the exploration of behavioural dynamics. However, making progress on the conceptual development of the notion of normative action and adaptive agency might require to integrate these two aspects.

In previous work (Egbert et al., 2009, 2010) we have explored the relationship between the viability boundary determining metabolic dynamics and the dynamics that drive organismic behaviour. But a further problem remained on these and other related models: although the boundaries of viability were directly linked to the modelled system, they were only defined by the system in a relatively trivial way. The boundaries of our models and similar efforts by others (see e. g. Ruiz-Mirazo and Mavelli, 2008) were the result of rough physical magnitudes: disappearance of the protocell due to complete lack of catalysts or bursting disintegration of the protocell marked by the upper limit of the tension of the membrane[5]. The boundaries were not intrinsically emergent from interactions between system processes in the holistic system-interdependent manner that characterizes integrity and systemic identity in real organisms.

In natural systems, the limits of viability do not map with the physical disintegration of a system but rather with the loss of the capacity of the system to sustain itself. To lose viability is not to disappear altogether, but to cross a much more subtle boundary where self-maintenance is overwhelmed by processes of degradation—where the system is on a trajectory towards death that can only be averted by changing the environmental conditions. Often, once viability has been lost, it becomes increasingly difficult to regain viability, sometimes reaching a point of no return and becoming irreversible. This problem connects back to the first one, the nature and rigidity of the boundaries as standardly conceived by the adaptive behaviour literature needs to be revisited. Additional regions within the space defined by essential variables might be required and the notion of boundary complemented with that of a field.

On what follows we shall introduce a minimal model of agency capable to address the problems just presented in the previous section. We first introduce a model of a minimal protocell-like metabolism whose dynamics define a precise topology of its viability space. This allows us to identify and illustrate *norm-establishing* as emerging from metabolism. For fixed concentrations of available resources, we can plot a bifurcation diagram of the chemodynamics that indicates the intrinsic boundaries of viability of the system. Different regions within the viability space will be identified and the adaptive *norms* of the system clearly defined and quantified. Next we add chemotactic capacities to the protocell, we put it in an environment with a chemical metabolic resource gradient and we study how behaviour maps into the viability space and consider how can it be interpreted as an instance of *norm-following* behaviour. We then abstract away from our model to more general cases and we finally come back to

[5] This is not to say that existing models assume a naive conception of such boundaries, in the case of Ruiz-Mirazo and Mavelli (2008) the viability boundary of the membrane is transformed by the addition of peptides that are produced by metabolism and is therefore "modulated" by the system. However, the only relevant viability boundary considered (although elastic and influenced by the system) is the maximum tension that the membrane can accommodate before bursting.

the problems and theoretical discussion opened above.

3 Norm-establishing: The emergence of norms from metabolic dynamics

3.1 A minimal model of metabolic autocatalysis

3.1.1 Model design

The metabolic organisation is one of the most fundamental properties of living systems and has been studied as such by many under the form of autocatalytic networks (Kauffman, 1986), autopoietic or autonomous systems (Maturana and Varela, 1980; Varela, 1979; Ruiz-Mirazo and Moreno, 2004; Bourgine and Stewart, 2004), fluid machineries like the *chemoton* model (Ganti, 2003), or systems closed to efficient causation (Rosen, 1991; Letelier et al., 2006; Piedrafita et al., 2010)— to mention but some of the most relevant contributions to a systemic conception that takes metabolic self-organization as the core principle of living organization.

Real metabolisms are complex networks involving hundreds of metabolites and even more reactions. But we are interested in general abstract properties of metabolic networks, so we do not need to model all of the intricacies. We start with the simplest conceivable model, involving the simulation of the concentrations of two categories of chemical components: autocatalysts (A) and resources or 'food' chemicals (F). The abstraction here is high-level; $[A]$ is intended to represent the concentration of all of the different chemicals involved in an autocatalytic metabolic network, conceptually equivalent to the "biomass" of metabolites or an order parameter globally representing metabolic dynamics. Similarly, $[F]$ represents the concentration of all material and energetic resources necessary for the ongoing functioning of metabolism.

We consider two processes: (i) the autocatalysis of A, whereby it transforms F into more A, and (ii) the spontaneous degradation of A into a quickly dissipating, non-reactive waste that has no subsequent effect upon the system. These two processes are described by the following reaction equations:

$$\begin{aligned} 2A + F &\leftrightarrow 3A \\ A &\rightarrow \emptyset \end{aligned} \tag{1}$$

The following differential equation describes how the concentration of A changes over time.

$$[\dot{A}] = \frac{-k_b[A]^3}{6} + \frac{k_f[F][A]^2}{2} - k_d[A] \tag{2}$$

The first term represents the 'backward reaction' of $3A \rightarrow 2A + F$, the second term represents the 'forward reaction' of $2A + F \rightarrow 3A$ and the third term represents the degradation of A into a non-reactive waste ($A \rightarrow \emptyset$).

The constants ($k_b = 0.45$, $k_f = 1.0$, $k_d = 1.0$) were assigned such that, for low $[A]$, the forward autocatalytic reaction occurs more rapidly than the backward reaction.

The stoichiometry of the autocatalysis was selected as it results in a bistable system; when the concentration of A is high, the backward reaction runs more quickly than the forward reaction.

3.1.2 Dynamics

Figure 2A shows the influence of each of the three terms on the right hand side of Equation 2 when the concentration of F is fixed at 1.4. The backward reaction decreases the concentration of A at a rate proportional to $[A]^3$. The forward reaction increases it at a rate proportional to $[F][A]^2$ and the degradation reaction decreases $[A]$ at a rate proportional to $[A]$.

As shown in Figure 2B, when $[F] = 1.4$, the combined effects of the three terms result in two stable equilibria, at $[A] = 0$ and $[A] = 7.57$. In between these two stable equilibria lies an unstable equilibrium at $[A] \approx 1.76$. For $[F] = 1.4$, only when the concentration of A is above this unstable equilibrium point, is the forward autocatalytic reaction sufficient to counteract the two processes that decrease the concentration of A.

Figure 3 shows the equilibria for fixed values of $[F]$ between 0 and 2. As could be expected, when there is a low concentration of F, no matter how high the concentration of A, the system is incapable of self-producing at a rate sufficient to compensate for its degradation. This means that for these low concentrations of F, there is only a single stable equilibrium where the concentration of autocatalyst, $[A] = 0.0$. The system bifurcates at $[F] \approx 1.09$. For fixed concentrations of F greater than this value, the system has two stable equilibria, one where $[A] > 0.0$, one where $[A] = 0.0$, and an unstable equilibrium that lies in-between. To calculate the value of $[F]$ at the bifurcation point, we solved Equation 2, identifying the value of $[F]$ where there are exactly two equilibria, as this is only the case at the bifurcation point. To do so, we first factored out an A from Equation 2.

$$[\dot{A}] = [A]\left(\frac{-k_b[A]^2}{6} + \frac{k_f[F][A]}{2} - k_d\right) \tag{3}$$

It is clear in this equation that one equilibrium occurs when $[A] = 0$, and that any other equilibria occur when what is inside the parentheses equals 0. When we apply the quadratic formula,

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{4}$$

to identify the roots of the quadratic inside the parentheses of 3, it is apparent that there is only a single additional root when $\sqrt{b^2 - 4ac} = 0$. Plugging the coefficients into this equation and then solving for $[F]$, we find the equation that specifies the value of $[F]$ at the bifurcation point.

$$[F] = \frac{2}{k_f}\sqrt{\frac{2 \cdot k_b \cdot k_d}{3}} \tag{5}$$

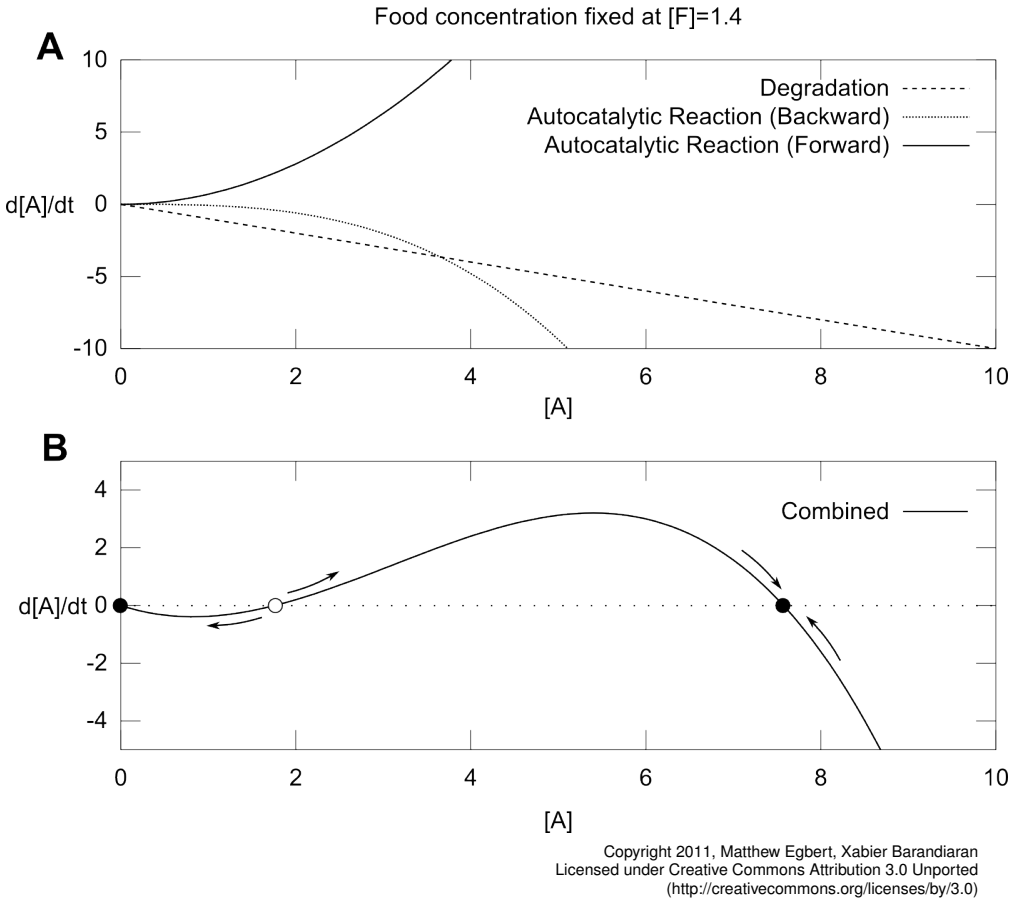

Figure 2: Phase portraits the model system for fixed $[\mathbf{F}] = \mathbf{1.4}$. A. The influence of each of the three terms of the autocatalytic equation, forward autocatalytic reaction (solid), backward autocatalytic reaction (dotted) and degradation (dashed). B. The combined influence of all three terms defines three equilibrium points: left and right stable equilibria in black and unstable equilibrium in white—arrows indicate dynamic tendencies of $[A]$ towards or away from fixed points.

When we substitute the constants specified above, we find that at the bifurcation point, $[F] \approx 1.09$.

3.2 Interpretation

3.2.1 The topology of the viability space: living, dead, precarious and viable regions

Such a simple model of metabolism suffices to generate a **viability space** where, **living**, **viable**, **precarious**, and **dead** regions can be clearly identified. These are highlighted in Figure 4. The **dead region** corresponds to the the zero concentration of the required metabolites (a complete disintegration of the system). The arrows in figure 4 indicate the general tendency of metabolic dynamics for different regions of the

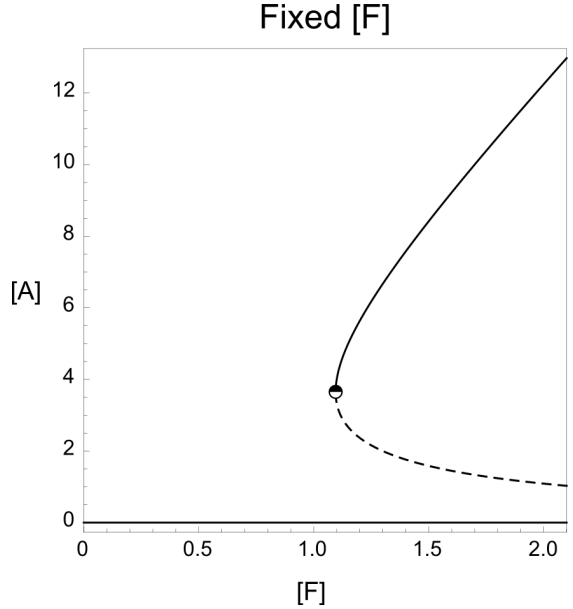

Figure 3: Bifurcation diagram showing equilibria for different (fixed) concentrations of F. Parametric analysis of $[A]$ equilibrium points for different values of $[F]$. Unstable equilibria are shown with a dashed line, and two stable equilibra with solid lines (a "trivial" equilibrium point at $[A] = 0$ and a non-trivial one above the unstable equilibrium). See text body for details.

viability space, superimposed upon the bifurcation diagram shown in Figure 3. A **viable region** can be precisely defined for a range of the parameter $[F]$ and a range of initial conditions $[A]$ as the subregion of the living region where, for each point, the evolution of the system will tend toward the stable living equilibrium. The unstable equilibrium at the bottom of the viable region defines a boundary of viability below which, the system tends to the **dead state**.

For small values of $[A]$ and $[F]$ we can distinguish what we have termed a **precarious region** (medium grey area in Figure 4), where the system is still alive but will tend to die if the parametric condition $[F]$ is kept constant, but could still recover if $[F]$ is appropriately modulated. Underneath the precarious region a **terminal region** can also be distinguished (dark grey area in Figure 4). If $[A]$ falls in this region the system will be "alive" for some time, but will irreversibly die. The reasons for this impossibility of recovering viability are due to intrinsic limiting factors on the dynamics of change of $[F]$. So, for instance, in a protocell, the maximum increase in $[F]$ could be limited by the maximum permeability of the membrane. Note that at this point we are still not considering any limitations coming from the structure of the environment itself or the capacities of the system to change those conditions (e. g. by moving around a gradient of $[F]$).

There is a significant difference between the standard representation of a viability

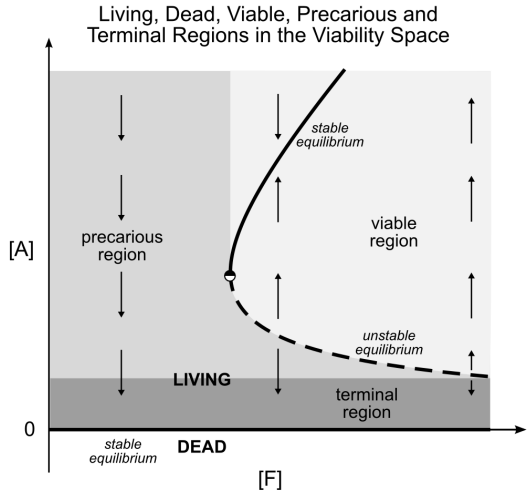

Figure 4: Regions in viability space. Living, dead, viable, precarious and terminal regions of the viability space. The *dead region* or state lies at $[A] = 0$, above which the *living region* appears. Inside the living region three different sub-regions are distinguished: the *viable region* (light grey) where the system will remain alive if environmental conditions don't change, the *precarious region* (medium grey) where the system is still alive but tends towards death unless environmental conditions change and the *terminal region* (dark grey) where the system will irreversibly fall into the dead region. See text body for detailed explanation.

space in Figure 1 and our representation in Figure 4. The first thing to note is that the viability space in our approach is not only defined by a set of *essential variables* or just a constraint, but by a relational property between internal variables ($[A]$) and an environmental parameters or boundary conditions ($[F]$). In addition, a viable region is not specified by a set of "ad-hoc" or experimentally determined boundaries: it is the result of the internal dynamics of the system in relation to the environment. The bifurcation point and the boundary between two basins of attraction (the viable and the dead one) are the result of the dynamics specified by metabolism. If we further explore the dynamics of the system within this regions we can move beyond a description of normativity in terms of *boundaries* and *regions* and enrich it with concept of *field*.

3.2.2 The Normative Field

We can now introduce the notion of a **normative field**, defined by the minimal constant increase of a behaviourally influenced variable, in this case $[F]$, that is required at each point of the precarious region in order to move the state of the system into a viable region; before the system reaches the terminal region. Figure 5 illustrates this field. If the values of $[F]$ and $[A]$ are low (bottom-left side of the figure) the required increase of $[F]$ is very big, since the tendency of $[A]$ will soon push the system into

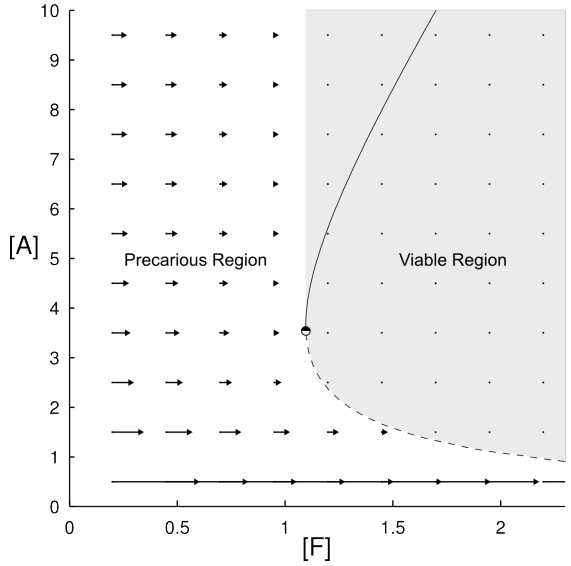

Figure 5: Normative vector field. Vectors indicate direction and minimum amount of constant increase in $[F]$ that is required to get into the viable region before reaching dying. See text for explanation.

the terminal region. If the concentration of F is low but there is a lot of A the required constant increase in $[F]$ is low because the system has sufficient time to reach the viability boundary before the tendency to die becomes irreversible. There are still rigid boundaries of a binary nature, like the dead-living boundary at $[A] = 0$, but the notion of normative field along the precarious region permits to define and quantify norms in a graded manner.

We have provided a minimal model of *norm-establishing* processes and how they define different normative regions and dynamic normative fields. The goal of the next section is to see how *norm-following* behaviour can occur and how it maps into the normative field.

4 Norm-following: Chemotaxis and normative behaviour

We now consider the possibility influencing or modulating $[F]$ through behaviour. There are many ways in which this could be accomplished. We consider two cases, both of which involve a motile protocell (or bacterium) performing chemotaxis on a gradient of F. The first mechanism is a stochastic mechanism inspired by the "run and tumble" behaviour observed in bacteria such as *Escherichia coli*. The second is a non-stochastic, 'perfect' gradient-climbing mechanism in which the simulated organism

always moves directly up the $[F]$ gradient [6].

4.1 Chemotactic models: design and dynamics

4.1.1 Stochastic gradient climbing

In this section, we extend the model described above, simulating the metabolism as being located within a bacterium-like body that can move around a 2D environment. In the environment, F is distributed according to a Gaussian gradient specified by the following formula (also, see Figure 6A).

$$[F]_{(x,y)} = \exp(x^2 + y^2) \tag{6}$$

The stochastic gradient climbing behaviour is inspired by the chemotaxis mechanism of *E. coli* (Eisenbach, 2007), which is based on the on the modulation of two types of motility: "running", where the organism moves in a straight line and "tumbling", a random reorientation of the organism. Running, is a linear, constant velocity motion defined by the following differential equations $\frac{dx}{dt} = \cos(\alpha)$, $\frac{dy}{dt} = \sin(\alpha)$, $\frac{d\alpha}{dt} = 0$, where α represents the orientation of the organism. A 'tumble' is simulated by selecting a new value for α from a flat distribution between 0 and 2π. Each tumble is always followed by a short 'cool down' period ($0.1t$) of running.

The run/tumble behaviour of the organism is modulated by change in the concentration of F. When $[F]$ is increasing, the organism tends to run and when it is decreasing the organism tends to tumble. This is simulated by calculating, every iteration, ρ, an approximation of the recent change in the concentration of F that includes an error term.

$$\rho := [F]_t - [F]_{(t-0.01)} + \epsilon \tag{7}$$

When $\rho \geq 0$, the simulated organism runs, otherwise (as long as it is not 'cooling down' from a recent tumble) it tumbles. The calculation of ρ includes the error term ϵ which is a random number selected from a normal distribution with a mean of 0 and a standard deviation of 0.01. This error term means that the organism does not always make a correct evaluation of the change in the concentration of $[F]$.

In most cases, if the organism starts reasonably close to the gradient it performs a statistical chemotactic behaviour, moving up the resource gradient. Figure 6A shows typical paths of three agents, each started from a different distance from the peak resource. Figure 6B shows the evolution of essential variables as a result of agent's behaviour, through the *viability space*, the same space as in Figures 3, 4, 5 and 7. We see in Figure 6B that Trajectory 2 (in green) starts in the precarious region, moves for some time in correlation with the normative field, then moves away from the peak

[6] Note that the mathematical formalisms and results we are about to interpret as chemotaxis could, under some circumstances and constraints, be interpreted in terms of more basic forms of agency in protocells that might not required motility: e.g. the regulation of membrane permeability and transport that affects the amount of metabolically available $[F]$— more realistic models of protocell metabolic dynamics under different permeability conditions can be found at Ruiz-Mirazo and Mavelli (2008).

(marked with *) and then turns into the viable region. Once there, it comes out of it (marked **) to return back into the viable region and stays there. It has survived a state that, in the absence of behavioural influence on $[F]$, would have led to its death, but it can be said to have made *mistakes* (marked * and **) on its way towards regaining viability. It is also the case that sometimes, despite performing chemotaxis, the simulated agents fail to reach sufficiently high concentrations of $[F]$ before they fall into the terminal irreversible region. Such is the case of Trajectory 1 in Figure 6, at the bottom line, where after having moved slightly against the gradient, it reaches a point around $[A] = 2$; there, despite being positively correlated with the normative field, the overall behaviour has been insufficient to reach the viable region, the agent cannot behaviourally compensate for its precariousness and enters the terminal region.

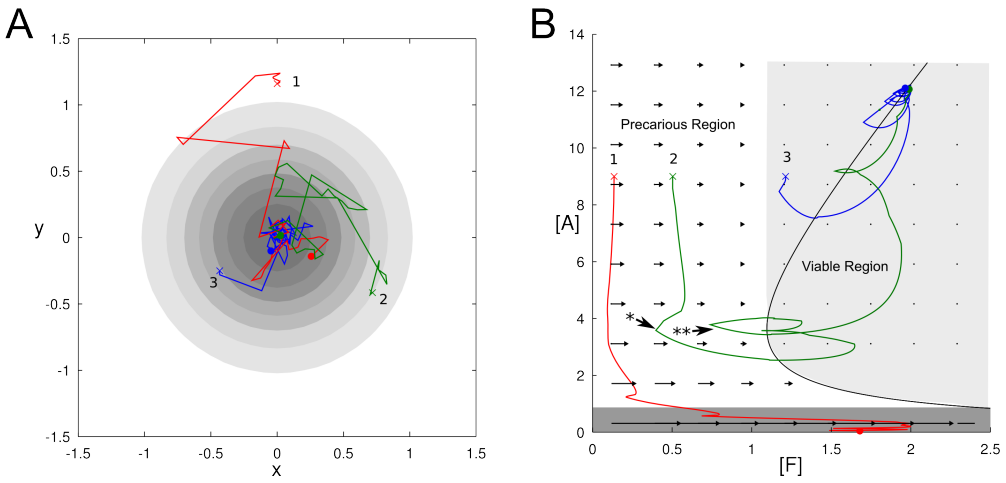

Figure 6: Trajectories of three stochastic gradient-climbing agents on 2D space (A) and the projection of the effect of their behaviour on the viability space. A. Three agents start at different distances from the peak of the resource. These typical trajectories ran for $5t$ (500 iterations), start at the Xs and end at the closed circles. B. Projection of the same agent's trajectories within the viability space. Agent 1 fails to behaviourally compensate for its precariousness and dies, whereas agents 2 and 3 remain alive within the viable region. Episodes of *failure* of agent 2 to correlate positively with the normative field are marked with * and **.

4.1.2 Direct gradient climbing

The mechanism just described has an overall statistical effect of moving the protocell up the resource gradient. To simplify the analysis, we abstract the stochasticity of the mechanism, replacing it with a continuous behaviour that performs a direct gradient climbing. When performing this behaviour, the organism always moves directly up the resource gradient at a rate proportional to its distance from the peak; an approximation of the overall effect of the stochastic gradient climbing behavior. The removal of

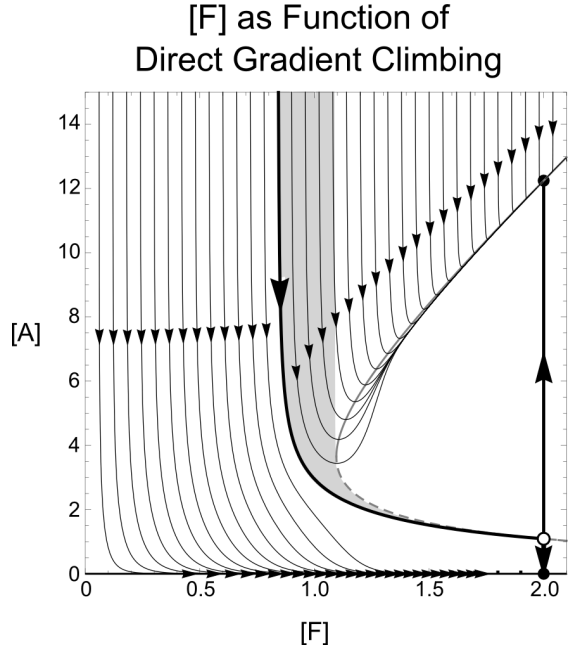

Figure 7: Superposition of direct gradient climbing behaviour upon the viability space: Whereas Figure 3 showed the bifurcation diagram of $[A]$ dynamics for fixed values of $[F]$, here we show the effect of a perfect gradient climbing behaviour (specified in Equation 9) on the viability space. In light grey, the space of the precarious region for which behaviour manages to reach the viable region. See text for details.

stochasticity simplifies the dynamical analysis and allows us to more easily visualise the effect of behaviour upon the viability of the simulated agent.

The radial symmetry of the F gradient and the fact that the direct gradient climbing behaviour only involves axial motion allow us to reduce the number of spatial dimensions in our model. Now, a single variable, x, representing the distance of the organism from the peak of the resource gradient, suffices. The following equation describes the gradient, relating $[F]$ to x.

$$[F] = 2e^{-x^2} \tag{8}$$

We define $\frac{dx}{dt} = -k_m x$, formalising our definition of direct gradient climbing mechanism. We then derive the following differential equation in terms of $[F]$, eliminating x from the model.

$$[\dot{F}] = 2k_m \cdot \ln(\frac{2}{[F]})[F] \tag{9}$$

In our analysis, we consider k_m, a constant indicating the speed of motion of the simulated organism, to be 0.1.

To explore the influence of the direct gradient climbing behaviour on the viability space of the simulated organism, we have superimposed streamlines indicating the

trajectories with perfect gradient climbing behaviour over the bifurcation diagram for fixed $[F]$ (see Figure 7). With perfect gradient climbing, the entire system has two stable equilibria, ("dead" and "living") that are indicated with solid circles, and a saddle node that is indicated with an open circle. All of these points are on the line $[F] = 2.0$. Also indicated on this diagram by a bold curve is the streamline that ends at the saddle-node. This path lies along the basin boundary between initial conditions that lead to the living equilibrium and those that lead to the dead equilibrium. The bold vertical arrows indicate the implication of a small perturbation from the saddle node.

The effect of the direct gradient climbing behaviour upon the viability space of the simulated organism can be seen in Figure 7. With $[F]$ no longer fixed, but increasing as a result of behaviour, the space of initial values of $[F]$ and $[A]$ that fall into stable "living" equilibrium point is greater than that of the non-behaving model.

4.2 Interpretation

4.2.1 Normative behaviour

Since $[F]$ can be modulated by behaviour (provided that the environment displays a gradient of $[F]$), a sense of *normative behaviour* can be precisely defined for every state of the system in the precarious region: the amount of increase of $[F]$ that behaviour should achieve to compensate for its precariousness; i.e. the required movement in space that increases available $[F]$ in accordance with the normative field. To remove the ambiguity on the expression "in accordance with" we can more precisely define norm-following behaviour as *the course of behaviour whose effect on the viability space positively correlates with the normative field*. Quantitative and qualitative aspects of these correlation can be distinguished. At the qualitative level, the projection of behavioural effects on the normative field can take three different forms of correlation: positive, negative or null. Whereas norm-following behaviour is that which is positively correlated with the normative field, the negatively correlated behaviour can be said to be *against the norm*, and the behaviour that has no correlation can just be said to be *neutral* (only in a sense relative to the correlation with the norm, since this "neutrality" is deadly for the system)[7]. At the quantitative level one can measure *how much* a behaviour matches the normative field. Note that the system can fail to meet the norm, i. e. to adapt, for a variety of causes (e. g. because there is not enough $[F]$ in the environment, because it cannot move sufficiently fast or does not manage to move up the gradient—like the case of the experiments illustrated in Figure 6). And yet, for some of these cases of failure, behaviour can be said to be in accordance with the norm, i. e. *a norm-following* behaviour, if it positively correlates with the normative field. Failure might be quantitative in these cases: despite the correlation with the normative field being positive the agent does not move sufficiently fast, behaviour matches the direction of the normative vectors but not their length.

[7] A proper sense of "neutrality" is found within the viable region where the normative field is null.

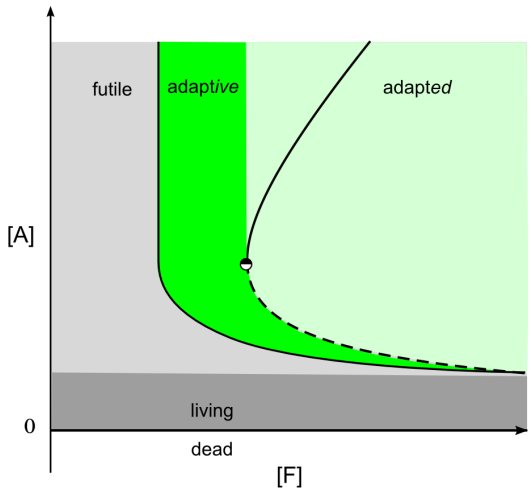

Figure 8: Adapt*ed*, Adapt*ive*, and futile regions of the viability space. The projection of the agent-environment coupling upon the viability space defines an *adaptive region* where the system is capable to compensate for its precariousness for that particular coupling and a *futile region* where the agent will inevitably die given its behavioural capacities and the structure of the environment.

4.2.2 The Adapted, Adaptive and The Futile Regions

We have projected the relational agent-environment coupling effect on the viability space (as shown in Figure 7) for a specific and stable environment and a deterministic and perfect behavioural capacity of the agent. The resulting "extension" of the viable region we have termed the **adaptive region**, because this is the area within the precarious region where behaviour is capable of compensating for decay (see Figure 8). Whereas the viable region can be said to be the space where the system is adapt*ed* (there is nothing that the system *has* to do), this is the region where the system is adapt*ive* (there is something that the system *has* to do and it *can* do). Unlike the viable region, which is defined for all possible environmental conditions, the adaptive region has to be defined in relation to both the structure of the environment (the gradient define by Equation 8) and the behavioural capacities of the system (defined by Equation 9).

An additional region can be added if we consider the limits of the agent's behavioural capacities and the structure and limits of the environment (the space left to the grey are in Figure 7 where all trajectories lead to death). We call this the **futile region**: no matter what it does the agent will die because it lacks the appropriate behavioural capacity to modulate its environmental conditions; or simply because the environment does not allow for it. So, for instance, if there is insufficient $[F]$ in the environment or if the chemotactic mechanism does not allow the protocell to move up

the gradient of $[F]$ fast enough, the agent will die. Note that we have defined normative behaviour as the behaviour whose effects positively correlate with the normative field. As a result, within the futile region the agent will be behaving normatively but without a real chance to survive (that is, out of the adaptive region). The futile region is specified by the organization of the metabolic system as well as behavioural or environmental dynamics. This is unlike the terminal region, which is specified by factors that are intrinsic to the organization of the system in a way that is independent of behavioural capacities and environmental structure.

5 Recapitulation: the topology of the viability space, some generic definitions

Abstracting away from our model and bringing together the set of distinctions we have made along the paper, we can now provide a set of generic definitions that could be applied to other models of adaptive or normative agency and behaviour.

Viability space: the space defined by the relationship between: a) the set of *essential variables* representing the components, processes or relationships that determine the system's organization and, b) the set of *external parameters* representing the environmental conditions that are necessary for the system's self-maintenance.

Living region: the region of the state space where a system maintains its ongoing organization as a unity.

Viable region: a region of the viability space where the evolution of essential variables, *ceteris paribus* (i.e. given fixed environmental conditions), will remain within the living region. That is: $\forall x(t) \in V \subset L; x(t+1) \in V$ (where L and V stand for *living* and *viable* regions respectively).

Precarious region: a subregion of the living region where the system will reach death if the environmental conditions do not change.

Terminal region: a subregion of the living region where the system will reach death no matter how the environmental conditions change.

Normative field: for each point of the precarious region, the required minimal change of environmental conditions necessary to bring the system back to a viable region or to a precarious region of an opposing tendency.

Normative action or norm following behaviour: system-driven modulation of the its coupling with the environment whose effect on the viability space positively correlates with the normative field.

Adaptive region: An adaptive region is defined as a projection of the agent-environment coupling dynamics into the viability space and specifies a subregion of the precarious region where for each point of that region at least one possible trajectory leads to viability.

Futile region: A futile region is defined as a projection of the agent-environment coupling dynamics into the viability space and specifies a subregion of the precarious region where all trajectories lead to death.

Some aspects of the above definitions remain in need of further development, we shall briefly mention some of them. The definitions of *viability space* and *living region* where provided for completeness, but a more precise and justified definition of these terms is out of the scope of the this paper. Similarly the term "essential variable" demands further explanation. For a protocellular system where a membrane or some enclosing property (e. g. phase separation between oil-droplet and water) distinguishes a self-sustaining network of reactions from the environment, essential variables represent those observables of a system that correspond to the capacity of the system to keep producing or maintaining itself as a system distinguishable from its environment. The expression "ongoing organization as a unity" also remains open to discussion and further clarification. For the purpose of this paper it suffices to indicate that at some point the essential variables of the system cease to represent an observable *of* the system. It could be just because the system disappears as a result of the observables disappearing altogether ($[A] = 0$ in our model) or for more complex reasons. So, for instance, if beyond a certain level of $[A]$ the protocell bursts then $[A]$ is not any more the amount of metabolites inside the membrane of an individual cell but a measure of the concentration of A in the "environment" (e. g. the Petri dish or pond where the bacteria was found before the bursting). We have committed to a specific conception of life —inspired by (Jonas, 1966; Maturana and Varela, 1980; Ganti, 2003; Ruiz-Mirazo et al., 2004)— but the definitions just provided do not strictly depend on it, other attempts to define life might be equally valid as long as different processes can define and be mapped into the viability space. One could even proceed, to some extent, without a definition of life and establish essential variables, environmental conditions and their boundaries experimentally (e. g. heart-rate = 0).

We are aware that the conceptual framework of these definitions could be applied to biological function more generally and not only to the issue of agency and normative behaviour. Because our focus of interest is agency we have decide to narrow down the framework making some simplifications that could lead to confusion. So, for instance, we defined the viability space and its topology in relation to "environmental conditions". But what we consider to be environmental conditions could perfectly be internal conditions and the relationship between agent and environment might be translated into part/whole relationship (e.g. the heart vs. the conditions that the rest of the body create for it, like oxygen supply, hormonal and neuronal activity, etc.).

We have modelled normativity at a single level of description/organization, one involving a macroscopic description of metabolic dynamics and its coupling with the

environment. What was defined as a viable region at this level could perfectly correspond to a precarious region at a lower level of description/organization if observed in more detail. Most cases of natural normativity are cases where various nested hierarchies of precarious processes constitute and organism and so non-precarious viable regions might only be defined by fixing or assuming the normative functioning of lower levels.

In our case the normative field was represented as a vector field, but could equally had been a scalar field, since there was a single viability parameter. More complex cases might require a tensor field to express different routes to viability that are equally valid when different viable regions co-exist within the same viability space. It could also be the case that the viable space be very small or non-existent, depending on the particular topology of the viability space. For such cases, and others where viable regions might still be in place, we have adapted the definition of normative field including the condition "to bring the system *to a precarious region of an opposing tendency*". Although in our model it turns out impossible to reach an opposing tendency without directly falling into the viable region, more complex cases might permit to remain indefinitely alive within the precarious region by navigating opposing tendencies of essential variables. For more complicated models where there are multiple environmental variables that influence the viability of the organism, identifying the normative field might not be trivial. We do not address the problem of finding a universal way for determining a normative field in this paper. More complex models will be required to make progress in this direction.

6 Discussion

The model we have presented here is still limited in a number of ways that we consider important. However we are now in a position to address them within the framework we have sketched. We have no space to deal with these topics in detail but we shall point them out briefly.

6.1 Virtuality, multidimensionality and plasticity of norms

The virtual nature of norms It follows from our analysis that norms have a virtual character. Norms are not mechanisms operating on the system. There is nothing on the *current* state of the system that specifies the norm. Rather, it is in the space of the *possible* dynamic evolution of the system for different environmental conditions that the normative vector field is found, in a conditional manner: "what-would-happen-if" the system were to evolve under such-and-such conditions and how it would have to compensate such an evolution if it were to avoid death in the future. The organization of the system is the source of norms, but it makes no sense to claim that norms themselves have any causal or regulatory effect.

Multidimensionality of the viability space The present model benefits from its low dimensionality in that it is easier to understand, but is also suffers, perhaps, from being over simplified in that there is really only two ways that the system can vary. Real organisms are of course much more complex and would display a multi-dimensional normative field and viability boundaries or surfaces. Richer topologies of the viability space might be able to give rise to different phenomena. One of them is the existence of different viability regions (in the form of islands in a viability space, separated by precarious regions) for the same organism, that could define "forms or regimes of viability" and their associated "forms of adaptation" (that is, modes in which behaviour can compensate for precariousness around those viable regions).

Plasticity of norms We have considered only fixed organizations that have no potential for change. Living systems are plastic, meaning that certain organic functions can be expanded or reduced in terms of their speed, tolerance, rate, resistance, etc. It is well known that animals, if progressively exposed to a poisonous environment can significantly increase their tolerance and achieve levels of exposure that would kill then quickly were those to arise suddenly. We have not considered a protocell that can change its metabolism or its relationship with the environment under different conditions. But our model could be expanded to address such cases. Examples of metabolic plasticity could include the activation of different metabolic routes under different conditions so that the viable region might change (e. g. $[F]$ is metabolised faster). An example of plasticity at the system-environment relationship level could be the alteration of membrane properties: e. g. if the protocell is exposed repeatedly to some chemical in the environment that could, in turn, increase (or decrease) the inflow of $[F]$, the terminal region could be expanded or reduced. Ruiz-Mirazo and Mavelli (2008) provide a perfect illustration of this notion of plasticity of boundaries, their model of a protocell includes the production of peptides that change the elastic properties of the membrane, modifying the upper limit of the internal concentration of metabolites that the protocell can reach before it bursts.

6.2 Moving beyond normative behaviour: agency and teleology

Up to this point we have studied *norm-establishing* dynamics on the one hand, behaviour on the other and we were able to define behaviour as normative or *norm-following* by mapping the effects of behaviour on the viability space and its normative field. Behavioural mechanisms in our model were directly sensitive to $[F]$, meaning they were assumed to be causally correlated with some environmental quantity or value (like transmembrane receptors do in bacteria). In a sense, the mechanisms and dynamics that produced behaviour in our model were still dissociated from norm-establishing processes. In our minimal system it happens to be the case that an increase in $[F]$ is always good. As a result, mechanisms as simple as those that can transduce variation in $[F]$ into the appropriate motor output were capable of following the norms (i. e. generating behaviour that positively correlated with the normative field) independently of the dynamics of $[A]$. A crucial point that we had to leave

out of this paper is the requirement that some kind of causal-dynamical integration between *norm-establishing* and *behaviour-generating* mechanisms might be necessary to achieve a full sense of agency and teleology. A system can follow a norm, yet it may do so blindly, without any causally relevant effect of norm-establishing processes upon norm-following processes. Can we really say that such systems are agents? In cases like bacterial chemotaxis it is assumed that natural selection has tuned sensor transduction mechanisms, sensorimotor chemical pathways and flagellar rotation speed and probabilities so that behaviour turns out to be adaptive. But is there any possibility to address intrinsic teleology other than the blind watchmaker creating organisms that behave normatively but are blind to their own norms?

A deeper dynamic integration between norm-establishing and norm-following processes is indeed possible. We can use the present model to explain this possibility. So, for instance, instead of just "tracking" the increase of $[F]$ in the environment it is possible to make behaviour be dynamically "sensitive" to $[A]$ (or even some combination of $[A]$ and $[F]$). In this way the system would be capable to modulate its behaviour in direct causal correlation with its viability dynamics and not just by responding to external conditions. Note that this cannot happen if viability boundaries are externally defined or contingently imposed on the viability space. This is why modelling the emergence of intrinsic norms is crucial. In our model, the unstable equilibrium boundary that separates the viable from the precarious region implies radically different dynamics for metabolism. Only when norms emerge in this manner can norm-establishing and norm-following processes be connected in a dynamically integrated manner.

The goal of this paper has been to clarify the notion of "a system following norms that are generated by itself". This is part of a larger body of research, where a great deal of interesting and exciting work lies ahead. We have published on and continue to study the relationship between metabolic dynamics and behavioural control in what we refer to as "self-sensitivity" —see Egbert (2011) for a comprehensive view. This research explores some of the consequences of what we have called *metabolism-based chemotaxis* in bacteria (Egbert et al., 2010) as opposed to the inherited concept of *metabolism-independent* forms of chemotaxis—a view that has recently been challenged by experimental confirmation of many species of bacteria performing various types of metabolism-dependent chemotaxis, see (Vegge et al., 2009; Alexandre, 2010). The evolutionary and adaptive potential of integratin metabolism and behaviour was also explored in Egbert et al. (2012).

Also, previous work by Barandiaran and Moreno (2006); Barandiaran (2008) has focused on the possible limitations of biological agency in terms of integrating norm-establishing processes with behaviour generating mechanisms, exploring instead the concept of autonomy and viability at the level of neurodynamics and behaviour as a new domain where properly behavioural or cognitive norms are to be found. We believe that exploring a deeper connection between *norm-establishing* and *norm-following* processes will provide a insightful research avenue to clarify the nature of teleology and agency.

Finally, we want to clarify that issues of normativity only relate to one of the three

conditions that, according to Barandiaran et al. (2009), define a system as an agent: i) individuality, ii) interactional asymmetry and iii) normativity. The model presented along this paper is therefore partial and is not intended to capture agency as such but just to address some of its constitutive properties (those related with normativity) [8].

7 Conclusions

One of the central challenges to naturalize and synthesize natural agents is the issue of normativity. This paper started with a short summary of some of the most relevant contributions to the characterization of natural norms, focusing on the organizational or organismic approach. This school of thought contends that norms emerge from the precarious and self-sustaining nature of living individuals, whose existence depends on the satisfaction of some environmental or boundary conditions. But further progress is needed to articulate a scientifically fruitful theory that can be formalized and modelled.

We identified two problems of previous modelling approaches to norms via the notion of viability boundaries. The first problem has to do with the adhoc imposition of norms on modelled systems and the resultant rigidity of viability boundaries and the absence of a normative gradation. In a circumvention of this problem, we modeled the *norm-establishing* processes explicitly, allowing us to generate a viability space with intrinsic boundaries and dynamic regions. This made possible us to derive the notion of a normative field, enabling the quantification of normative behaviour in a dynamical way that is richer than the trespassing of a binary "life or death" boundary—as suggested by Di Paolo (2005). In this sense the precarious region avoids the standard rigidity of boundaries, generating a middle ground between the boundary that separates viable and precarious regions and the boundary that separates precarious and terminal regions. It is along this precarious middle-ground where tendencies of essential variables are explicit and can be compensated. Although our model did not address the possible plasticity or elasticity of such boundaries we considered possible expansions of the model to accomodate this phenomena.

We then introduced behaviour generating mechanisms and addressed the second identified problem: the problem of dissociation between norm-establishing and norm-following dynamics. By mapping the effect of behaviour upon metabolism, we were capable to assess the normative nature of particular actions. We finally acknowledged the limitation of our current model, identifying the need to make further progress by integrating the dynamics of norm-establishing processes directly with those generating behaviour; so as to avoid a normative-blindness and make progress towards a better understanding of natural agency. Many open questions lie ahead. But the

[8] It can even be argued that our model fails to satisfy the conditions of individuality and interactional asymmetry: a) a single autocatalytic reaction does not suffice to capture the relevant conditions for the emergence of a biological individuality and b) a direct gradient climbing mechanism or a stochastic chemotaxis that is directly correlated with $[F]$ does not satisfy the kind of asymmetry that caracterizes agential forms of sensorimotor coupling (the system is fully and deterministically driven by the current state of its environment).

minimal model presented here (and the conceptual definitions and distinctions it has made possible) allow us, now, to formulate those questions more precisely and move to beyond the inspiring, but also limited, theoretical and philosophical approaches that put the autonomy of the life, its capacity to create norms, at the centre of living phenomenology.

Acknowledgements

We would like to thank Ezequiel Di Paolo, Alvaro Moreno and Kepa Ruiz-Mirazo for discussion on early results and theoretical design of the present model and Randy Beer for his help with the dynamical analysis of the model. Special thanks to Kepa Ruiz-Mirazo for his careful and insightful revision of the manuscript.

Xabier E. Barandiaran currently holds a postdoctoral position funded by FP7 project eSMCs IST-270212 (EU 7th Framework through "ICT: Cognitive Systems and Robotics") and also hold a Postdoc with the FECYT foundation (funded by Programa Nacional de Movilidad de Recursos Humanos del MEC-MICINN, Plan I-D+I 2008-2011, Spain) during the development of this work. XEB also acknowledges funding from "Subvencion General a Grupos de Investigacion del sistema universitario vasco. Grupo Filosofia de la Biologia" from Gobierno Vasco IT 505-10.

References

Alexandre, G. (2010). Coupling metabolism and chemotaxis-dependent behaviours by energy taxis receptors. *Microbiology*, 156(8):2283–2293.

Apel, K.-O. (1980). *Towards a transformation of philosophy*. Marquette University Press.

Ashby, W. R. (1952). *Design for a Brain: the origin of adaptative behaviour*. J. Wiley, London, 2nd edition.

Aubin, J., Bayen, A., and Saint-Pierre, P. (2011). *Viability Theory: New Directions*. Springer, 2nd edition. edition.

Barandiaran, X. and Moreno, A. (2006). On what makes certain dynamical systems cognitive: A minimally cognitive organization program. *Adaptive Behavior*, 14(2):171.

Barandiaran, X. E. (2008). *Mental Life: A naturalized approach to the autonomy of cognitive agents*. PhD thesis, University of the Basque Country, San Sebastian.

Barandiaran, X. E., Di Paolo, E. A., and Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5):367–386.

Barandiaran, X. E. and Moreno, A. (2008). Adaptivity: From metabolism to behavior. *Adaptive Behavior*, 16(5):325–344.

Barandiaran, X. E. and Ruiz-Mirazo, K. (2008). Modelling autonomy: Simulating the essence of life and cognition. *Biosystems*, 91(2):295–304.

Bechtel, W. (2007). Biological mechanisms: Organized to maintain autonomy. In *Systems Biology: Philosophical Foundations*.

Bedau, M. A. and Norman, P. H. (1996). Measurement of evolutionary activity, teleology, and life. In Langton, I., Taylor, C., Farmer, D., and Rasmussen, S., editors, *Artificial Life II*, volume X of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 431–461. Addison-Wesley, Redwood City, CA.

Beer, R. D. (1997). The dynamics of adaptive behavior: A research program. *Robotics and Autonomous Systems*, 20:257—289.

Bernard, C. (1865). *Introduction l'tude de la mdecine exprimentale*. JB Baillire et fils.

Bersini, H. (1994). Reinforcement learning for homeostatic endogenous variables. In *Proceedings of the third international conference on Simulation of adaptive behavior : from animals to animats 3: from animals to animats 3*, page 325333, Cambridge, MA, USA. MIT Press.

Bickhard, M. H. (2009). The biological foundations of cognitive science. *New Ideas in Psychology*, 27(1):7584.

Bourgine, P. and Stewart, J. (2004). Autopoiesis and cognition. *Artificial life*, 10(3):327–345.

Burge, T. (2009). Primitive agency and natural norms. *Philosophy and Phenomenological Research*, 79(2):251–278.

Canguilhem, G. (1966). *Le normal et le pathologique*, volume 2. Presses universitaires de France.

Canguilhem, G. (1991). *The Normal and the Pathological*. Zone. Original in French 1966.

Cannon, W. (1932). *The wisdom of the body*. WW Norton & Company, inc.

Christensen, W. D. and Bickhard, M. H. (2002). The process dynamics of normative function. *The Monist*, 85(1):329.

Christensen, W. D. and Hooker, C. A. (2000). Autonomy and the emergence of intelligence: Organised interactive construction. *Communication and Cognition*, 17(3-4):133157.

Collier, J. (2000). Autonomy and process closure as the basis for functionality. *Annals of the New York Academy of Sciences*, 901(1):280–290.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60(23):685–700. ArticleType: primary_article / Issue Title: American Philosophical Association, Eastern Division, Sixtieth Annual Meeting / Full publication date: Nov. 7, 1963 / Copyright 1963 Journal of Philosophy, Inc.

Di Paolo, E. A. (2004). Unbinding biological autonomy: Francisco varela's contributions to artificial life. *Artificial Life*, 10(3):231233.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, 4(4):429–452.

Egbert, M. D. (2011). *Adaptation From Interactions Between Metabolism and Behaviour: Self-Sensitive Behaviour in Protocells*. PhD thesis, University of Sussex.

Egbert, M. D., Barandiaran, X. E., and Di Paolo, E. A. (2012). Behavioral metabolution: The adaptive and evolutionary potential of Metabolism-Based chemotaxis. *Artificial Life*, 18(1):1–25.

Egbert, M. D., Barandiaran, X. E., and Paolo, E. A. D. (2010). A minimal model of Metabolism-Based chemotaxis. *PLoS Comput Biol*, 6(12):e1001004.

Egbert, M. D., Di Paolo, E. A., and Barandiaran, X. E. (2009). Chemo-ethology of an adaptive protocell: Sensorless sensitivity to implicit viability conditions. In *Advances in Artificial Life, Proceedings of the 10th European Conference on Artificial Life, ECAL.*, pages 242—250. Springer.

Eisenbach, M. (2007). A hitchhiker's guide through advances and conceptual changes in chemotaxis. *Journal of Cellular Physiology*, 213(3):574–580.

Enoch, D. (2006). Agency, shmagency: Why normativity won't come from what is constitutive of action. *The Philosophical Review*, 115(2):169–198.

Frankfurt, H. G. (1978). The problem of action. *American Philosophical Quarterly*, 15(2):157–162.

Ganti, T. (2003). *The Principles of Life*. Oxford University Press, USA.

Gayon, J. (1998). The concept of individuality in canguilhem's philosophy of biology. *Journal of the History of Biology*, 31(3):305325.

Goldstein, K. (1934). *Der Aufbau des Organismus*. M. Nijhoff.

Habermas, J. (1990). *Moral consciousness and communicative action*. MIT Press.

Jonas, H. (1966). *The Phenomenon of Life. Toward a Philosophy of Biology*. Chicago-London.

Jonas, H. (1968). Biological foundations of individuality. *International Philosophical Quarterly*, 8(2):231–251.

Kauffman, S. (2003). Molecular autonomous agents. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 361(1807):1089–1099.

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

Kauffman, S. A. (2000). *Investigations*. Oxford University Press, USA, 1 edition.

Letelier, J., Soto-Andrade, J., Abarzua, F. G., Cornish-Bowden, A., and Cardenas, M. L. (2006). Organizational invariance and metabolic closure: Analysis in terms of (M,R) systems. *Journal of Theoretical Biology*, 238(4):949–961.

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and cognition: the realization of the living*. Springer.

McFarland, D. D. (1999). *Animal Behaviour: Psychobiology, Ethology and Evolution*. Longman, 3 edition.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science*, 56(2):288–302. ArticleType: research-article / Full publication date: Jun., 1989 / Copyright 1989 Philosophy of Science Association.

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

Piedrafita, G., Montero, F., Morn, F., Crdenas, M. L., and Cornish-Bowden, A. (2010). A simple Self-Maintaining metabolic system: Robustness, autocatalysis, bistability. *PLoS Comput Biol*, 6(8):e1000872.

Quine, W. (1969). Epistemology naturalized. *Ontological Relativity and Other Essays*, 13(3):6990.

Rand, S. (2011). Organism, normativity, plasticity: Canguilhem, kant, malabou. *Continental Philosophy Review*, 44(4):341–357.

Rietveld, E. (2008). Situated normativity: The normative aspect of embodied cognition in unreflective action. *Mind*, 117.

Rosen, R. (1991). *Life Itself*. Columbia University Press, 1 edition.

Ruiz-Mirazo, K. and Mavelli, F. (2008). On the way towards 'basic autonomous agents': Stochastic simulations of minimal lipid-peptide cells. *Biosystems*, 91(2):374–387.

Ruiz-Mirazo, K. and Moreno, A. (2000). Searching for the roots of autonomy: The natural and artificial paradigms revisited. *Communication and CognitionArtificial Intelligence*, 17(3-4):209228.

Ruiz-Mirazo, K. and Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life*, 10(3):235–259.

Ruiz-Mirazo, K. and Moreno, A. (2012). Autonomy in evolution: from minimal to complex life. *Synthese*, 185(1):21–52.

Ruiz-Mirazo, K., Pereto, J., and Moreno, A. (2004). A universal definition of life: Autonomy and Open-Ended evolution. *Origins of Life and Evolution of Biospheres*, 34(3):323–346.

Silberstein, M. and Chemero, A. (2011). Dynamics, agency and intentional action. *Humana Mente*, 15:1–19.

Skewes, J. and Hooker, C. (2009). Bio-agency and the problem of action. *Biology and Philosophy*, 24(3):283–300.

Varela, Thompson, E., and Rosch, E. (1991). *The embodied mind: cognitive science and human experience*. MIT Press.

Varela, F. J. (1979). *Principles of biological autonomy*. North Holland New York.

Vegge, C. S., Brondsted, L., Li, Y., Bang, D. D., and Ingmer, H. (2009). Energy taxis drives campylobacter jejuni toward the most favorable conditions for growth. *Appl. Environ. Microbiol.*, 75(16):5308–5314.