Stapleton & Froese - penultimate draft for: Murillo, García-valdecasas and Barrett (Eds.),
Biology and Subjectivity: Philosophical Contributions to a Non-Reductive Neuroscience (Dordrecht: Springer).
First submitted March 14th 2014, revised version submitted June 5th 2015.

Chapter 8

The enactive philosophy of embodiment: from biological foundations of agency to the phenomenology of subjectivity

Mog Stapleton [1 *] and Tom Froese [2, 3]
[1] Institut für Philosophie, Universität Stuttgart, Germany
[2] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Mexico
[3] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico

8.1. Introduction

Following on from the philosophy of embodiment by Merleau-Ponty, Jonas and others, enactivism is a pivot point from which various areas of science can be brought into a fruitful dialogue about the nature of subjectivity. In this chapter we present the enactive conception of agency, which, in contrast to current mainstream theories of agency, is deeply and strongly embodied. In line with this thinking we argue that anything that ought to be considered a genuine agent is a biologically embodied (even if distributed) agent, and that this embodiment must be affectively lived. However, we also consider that such an affective agent is not necessarily also an agent imbued with an explicit sense of subjectivity. To support this contention we outline the interoceptive foundation of basic agency and argue that there is a qualitative difference in the phenomenology of agency when it is instantiated in organisms which, due to their complexity and size, require a nervous system to underpin their physiological and sensorimotor processes. We argue that this interoceptively grounded agency not only entails affectivity but also forms the necessary basis for subjectivity.

To begin with, we introduce an emerging movement in cognitive science and related fields, known as enactivism. The enactive approach to cognitive science brings together several fields of study into one coherent research program of life, mind and sociality (Thompson 2007; Di Paolo & Thompson 2014). It thereby inherits the interdisciplinary perspective that is characteristic of the cognitive sciences, but puts special emphasis on a number of additional fields that have been neglected by the mainstream. In particular, the enactive approach stands out by bringing together two venerable traditions of continental philosophy, Husserlian phenomenology and the philosophy of the organism, with cutting edge research in the sciences of complexity, as formalized by dynamical systems theory (Weber & Varela 2002). Perhaps surprisingly, the insights gained by phenomenology lend themselves to being described as structures in a temporal flow using dynamical systems theory (Varela 1999), and Husserl's method of using imaginative variation to reveal a phenomena's essential characteristics is not far removed from computer-aided systems modeling of minimal cognition (Froese & Gallagher 2010). Similarly, key philosophical claims about the self-organizing and self-producing nature of organisms, going back at least to Kant, are starting to find expression in the fields of AI, systems biology and artificial life modeling (Froese & Ziemke 2009; Di Paolo 2010).

This confluence of approaches puts enactivism in a privileged position to investigate the relationship between biological embodiment and phenomenological subjectivity (e.g. Desmidt et al. 2014). It does so by conceptualizing the objectively living body (*Körper*) and the subjectively lived body (*Leib*), following Husserl's ([1952] 1989)[1] terminology, as two sides of the same coin. The traditional mind-body problem is therefore converted into the more tractable "mind-body-

[1] Stylistic note: we cite the most recent English edition of our sources whenever this is possible for ease of reference. However, in order to avoid giving a false impression of the original date of publication, we also always provide the year of publication of the first edition in square brackets.

body problem" (Hanna & Thompson 2003). And we are given an additional mediating concept that has been almost completely neglected in analytical philosophy of mind, but which has taken center stage in enactive theory, namely life itself:

> The Mind-Body-Body Problem […] is how to understand the relation between (i) one's subjective consciousness, (ii) one's living and lived body (*Leib*), that is, one's animate body with its "inner life" and "point of view;" and (iii) one's body (*Körper*) considered as an objective thing of nature, something investigated from the theoretical and experimental perspective of natural science (physics, chemistry, and biology). (Hanna & Thompson 2003)

This constellation of phenomena has long fascinated the phenomenological tradition of philosophy. For example, Scheler ([1928] 2008), Plessner ([1928] 1975), Merleau-Ponty ([1942] 1983), Sartre ([1960] 2004) and Jonas ([1966] 2001) were all in their own ways interested in grounding the origins of human subjectivity in the most basic principles of organic being and behavior. To be sure, accepting that the human condition is partially constituted by forms of organic and animal life does not entail that these forms explain all there is to being human. For example, Heidegger ([1929] 1995) was inspired by the biologist von Uexküll (1909, [1934] 1957), who argued for the existence of an organism's own point of view (i.e. its *Umwelt*), to contrast the restricted point of view of the animal from the conceptual world of the human. Nevertheless, Heidegger recognized that our being-in-the-world (*Dasein*) depends both on *Dasein*'s understanding of being as well as on *Dasein*'s living embodiment (Kessel 2011).

This tradition of relating subjectivity with biological embodiment is not a mere historical relic; it is continued by modern phenomenological philosophers such as Barbaras (2005), Gallagher (2005) and Zahavi (1999), who also engage with the ongoing development of enactive theory (e.g., Barbaras 2002, 2010; Zahavi 2011; Gallagher 2012). In the following we complement these efforts by sketching an enactive approach to the question of how a physical living body (*Körper*) can be an affectively lived body (*Leib*) and also a reflectively lived consciousness. Specifically, first we will describe what kind of embodiment is an essential prerequisite for affective agency and then we will consider some additional biological constraints that are imposed by phenomenological subjectivity. Here we are not concerned with giving the biological or the phenomenological domain metaphysical priority but, following Varela's (1996) working hypothesis of neurophenomenology, with putting insights from both domains into a relationship of mutual constraints to further our understanding of the phenomenon of life as a whole.

8.2. Biological foundations of agency

What is it about biological cognitive systems that makes us want to talk in terms of them being agents? It cannot merely be that they move around in and make changes to the environment - after all a robotic vacuum cleaner achieves this very effectively. If we admit such a reactive robot into the class of agents, then we have to include the humble thermostat as well, and the concept of agency eventually ends up being so inclusive so as to be theoretically useless (Froese 2014). Nor can it be that a system must have a thematised *feeling* of agency, explicitly experiencing themselves as the author of their actions to the point of reflective self-awareness. Such a strong stipulation would plausibly rule out most non-human animals and quite possibly even infants: biological systems which - even though they may not be aware of it - intuitively instantiate some sort of basic subjecthood beyond its mere attribution by others.[2] There is something deep in the

[2] The attribution of subjecthood by others may in fact be an essential element in infants' development of an explicit awareness of their own subjectivity (Reddy 2003), but on the view we are promoting here they were already agents on

notion of agency, such that we know that merely attributing agency to a system, from our perspective as external observers, is not sufficient for it to genuinely be an agent (Rohde & Stewart 2008). Agency is intrinsic to the system itself, but in virtue of what?

The most obvious feature of systems which one might consider as uncontroversially agentive is that they have wants and needs. The kinds of robots which are commonly termed 'agents' in informatics seem - at least on the surface - to have these, typically because they have been programmed, designed or evolved so as to move as if they are satisfying needs (e.g. Parisi 2004). Biological systems, of course, do not need to act *as if* they have them. As Weber and Varela (2002), following Jonas ([1966] 2001), argue, their needs are fundamental to their continued existence: a living system that no longer has any needs to satisfy is in fact no longer a living system. As an aside, we note that the phenomenon of an agent's wants is a bit more difficult to address, because fulfilling a desire goes beyond the mere satisfaction of an existing, well-defined lack to an open-ended quest for something that is not yet present (Barbaras 2002, 2010). As such, desires are probably dependent on a more explicit awareness of one's own subjectivity.

The enactive theory of agency has its roots in these biological foundations of cognitive systems. A precursor to its theoretical framework started taking shape in the cybernetics movement (Froese 2010), for example in attempts to explain organisms' purposeful behavior in terms of feedback dynamics. Another important milestone was Maturana and Varela's (1987) work on the self-organising and self-producing properties of minimal living systems such as the cell. They argued that the cell is the minimal living system because it forms itself as an identity - that is to say it forms itself as a system distinct from its environment. Its particular organization allows it to be self-producing - the processes that go on inside the cell produce the boundary of the cell which distinguishes it as an identity, and this boundary allows these processes within the cell to keep going and producing it - a circularly causal process (Varela, Maturana, & Uribe 1974; Varela 1979). This form of organization of the living system was termed *autopoiesis*, a concept that is closely related to a number of other technical concepts.

While autopoietic systems are material systems, what is key to the formation of their own systemic identity is the specific organizational nature of the metabolic processes rather than the particular material processes with which that organization is realized (for an introductory overview, see Di Paolo and Thompson 2014). The autopoietic organization is defined as autonomous because it has the property of operational closure, which means that the organization subserves a dynamic process of self-generation under far-from-equilibrium conditions. In other words, an autonomous system is organized as a network of processes that mutually depend on each other, and on the organization of the whole network, for their continued existence. Although there are similarities between these ideas and the cybernetics of feedback systems, the crucial difference is that autonomy ensures that the system is genuinely self-determining from the bottom up and not just self-maintaining a set of externally defined conditions (Froese & Stewart 2010).

These requirements are not trivial, but nevertheless an autopoietic system could spontaneously emerge and continue to exist and self-generate without having to be a sensorimotor system - providing that it exists in optimal conditions that provide everything it needs for its continued self-production. In addition, there are some interesting examples of self-organizing material systems that may or may not be autopoietic systems, such as a tornado or dust devil. Here we need to be careful to distinguish self-maintenance from self-production. Autopoiesis refers to a network of processes of *production* of new components, whereas a tornado and a dust devil presumably are only *rearranging* pre-existing components. Things are not quite as simple as this (see, e.g., the critical discussion by McGregor and Virgo 2011), but the difference between

---

their own terms even before they were born. This is what allows much of the body schema to develop *in utero* rather than post-partum (Lymer 2011).

synthesis of new components compared to mere re-organization of existing components provides a useful heuristic.

Nevertheless, a spontaneously emerging autopoietic system - even if it were autonomous, would not be *agentive* - its movements and indeed entire existence are at the mercy of external factors and its survival as a system is just a matter of luck that the right conditions for its continued existence happen to occur. In actual fact, even the simplest autopoietic systems, such as reaction-diffusion systems, are not *entirely* passive; they are capable of self-movement and a limited range of interactions with their environment (Froese, Virgo, & Ikegami 2014). But the important point here is that these basic autopoietic systems are not capable of actively regulating their behavior in relation to their needs. Although in the early formulations of the enactive approach an equivalence between autopoietic, living, and cognitive systems was assumed (e.g., Stewart 1992), this has started to be questioned. Some enactivists currently argue that autopoiesis is necessary but not sufficient for life and life necessary but not sufficient for cognition (Froese & Di Paolo 2011) while others argue that it is *autonomy* rather than autopoiesis that is necessary for cognition (Thompson 2007; Thompson & Stapleton 2009). The precise relation between autopoiesis, autonomy, and life remains an open question for future research (see in particular the discussions in Thompson 2011; Wheeler 2011).

Di Paolo (2005) enhanced the concept of autopoiesis with that of adaptivity in order to yield an organizational structure which subserves the kind of systems that we consider to be agents. An autopoietic system, however it is instantiated, is going to have limits to what kind of changes can happen in the environment and within its own systems that still allow the system to continue. At some point, however robust the system is, if the changes from the organization are too great it is not going to be able to self-produce and self-maintain. The set of changes that can happen within these limits are its viability set. According to Di Paolo (2005), in order for an autopoietic system to be able to continue its existence under changes in the environment, rather than just cease to exist as a system, it needs to "(i) be capable of determining how the ongoing structural changes are shaping its trajectory within the viability set, and (ii) have the capacity to regulate the conditions of this trajectory appropriately" (Froese & Di Paolo 2011, p. 8). This is the property of *adaptivity* (Di Paolo 2005; see also Barandiaran & Moreno 2008). There is a growing consensus in enactivism that autopoiesis and adaptivity are necessary and sufficient for life, and that therefore living is sense-making because the underlying adaptive processes are normative (Thompson 2011).

The processes necessary for adaptivity can occur within the system, i.e. by means of modification of the internal milieu, but only to a certain extent. In order to be more adaptive, a system must be able to adjust its relationship with its environment, such as by moving its position to a more favorable location, in order to change the effect this environment has on its viability. And because these changes in its environment caused by its moving are related to what the system needs in order to continue its existence, this movement is not a mere passive motion but is realized according to some goal or norm (survival), and is thus defined as an action. This gives the system intrinsic agency.

Developing this idea, Barandiaran, Di Paolo and Rohde (2009) propose an operational definition of agency. After critically reviewing existing definitions of agency, they argue that "we can generalize that agency involves, at least, *a system doing something by itself according to certain goals or norms within a specific environment*" (p. 369), and flesh this out with the necessary and sufficient conditions/aspects of agency, namely individuality, interactional asymmetry, and normativity. More specifically,

> (i) there is a system as a *distinguishable entity* that is different from its environment, (ii) this system is *doing* something by itself in that environment and (iii) it does so according to a certain goal or *norm*. (Barandiaran et al. 2009, p. 369).

Let us consider these three requirements of agency in more detail.

(1) Individuality: Individuality needs to arise from the structure of the system itself rather than be attributed from an outside observer. Thus an artificial system could conform to this requirement, but only if the processes which constitute it as a system bind it together in a coherent way and by doing so distinguish it from the environment with which it interacts (Froese & Ziemke 2009). It does not suffice to just distinguish a part of the world and view it as a system for our explanatory purposes. It may be useful to think of some non-self-individuated parts of the world as active systems, but to do so is to think of them from *our perspective* and not as a result of any particular dynamics of the system in question *as an individual*. To think in this way therefore does not imply that such systems have an identity of their own (Froese 2014). A mobile robotic 'agent' of the kind standardly found in robotics labs therefore does not fulfill this criterion - it is we who demarcate the mechanisms of the robot as the relevant system for our explanatory project. We could just as well demarcate the level of its interactions with other robots as the level of the system - as we do in swarm robotics. Or even 'agent + environment' as the system (Beer 2000). It all comes down to what our explanatory project is. There does not seem to be a basis for a more or less 'correct' attribution of the system's boundaries. Compare this to a living cell. In the case of the cell the tight interactions between its components feed back to each other and enable each other to exist and to continue as such, i.e. the system consists of a network of operationally closed processes. It is - ontologically - an individual because when we, in our role as external observers, distinguish it as a system appropriately, then the system will reveal itself to us as actually autonomously distinguishing *itself*, that is, as existing independently of our epistemological choices and distinctions.

(2) Interactional asymmetry: for agency it is not sufficient for an individual system to just be a moving system, nor to merely be in interaction with the environment or other systems. Nor is it sufficient for it to rely on a subsystem (which is not relevantly interconnected with the rest of the system) that drives its movement (e.g., the idea of metabolism-independent chemotaxis, see Egbert, Barandiaran, and Di Paolo 2012). The movement must arise at the agent level as a whole so that the agent uses the movement to modulate its coupling with the environment. The key point here is that agency requires that the adaptive regulation of agent-environment interaction is realized by the agent rather than resulting only from contributions of the environment. This distinction is important because some types of behavior can be adaptive at the level of the population, i.e. they lead to enhanced reproductive success, while still remaining reactive at the level of the individual, i.e. they are driven mostly by the environment (Froese, et al. 2014). Active regulation of interaction does not need to be happening all of the time in order for us to attribute agenthood but it must be a relevant aspect of the system. It is this aspect that underpins the system's greater adaptability as its regulated movement allows it to find an environment that better suits its viability set.

(3) Normativity: The concept of normativity should not be misunderstood as applying only in relation to human cultural conventions that guide action, although in that context it certainly finds a particularly elaborated expression (Torrance & Froese 2011). What is of interest here is something more fundamental, namely the biological norms that guide adaptive behavior. For a movement to be an action of an agent it must be a movement of an individual, have interactional

asymmetry (arise from the agent modulating its coupling) *and* be relevant to some goal - a goal which it either achieves or fails to achieve. This goal should not be externally given and the system arbitrarily directed toward its realization, such as when designing an artifact to behave in certain ways. Rather the goal should arise from and be relevant to the system's self-producing and self-maintaining activity. In a manner of speaking the system must aim at a goal in order for its movement to be an action, and it must be possible to fail at achieving this goal. However, we should beware of letting our manner of speaking mislead us into reifying the basis of this biological normativity into hypothetical entities of some sort, such as explicit representations of goals and norms. These are clearly operative in the cultural domain, for example in the legal system (Gallagher 2013), but they get in the way of operational explanations aimed at the subpersonal level (Di Paolo, Rohde & De Jaegher 2010). Such goals or norms emerge within the living system as a result of the autonomous or adaptive dynamics (metabolic or otherwise) attempting to keep the system within its boundaries of viability (Barandiaran & Egbert 2014).

An account of agency that requires individuality, interactional asymmetry, and normativity gets us some of the aspects which are key to our intuitive notion of agency, in particular its being the source of action and intrinsic intentionality. While these requirements may not be sufficient for realizing agency in all its forms, they seem to at least be instantiated in systems to which we intuitively do attribute agency. And, of particular import, such an account is operationalisable and thus, even though it is grounded in biology, it is not biologically chauvinist. To say that an account of agency is 'not biologically chauvinist' is to say that it does not rule out the possibility of a non-biological instantiation of agency (see also the discussion by Thompson 2011). The enactive conception of agency - grounded in the systemic principles of autonomy and adaptivity - is nevertheless embodied in a non-trivial way, because it is not only that the body instantiates properties that make it an agentive system but that these functions are by their very nature grounded in biology: the biology of value. What this means is that even if it is operationalisable and realizable in an artificial system, this artificial system will be, in a sense, an artificial biological system because the values that arise out of the system and hence its normativity are grounded in its autonomous (in this case *artificial-autopoietic*) self-production and adaptive self-maintenance, hence it would match our criteria of a living system – even if it is not composed of typical organic material.

Agency is one of the ways in which enactivism is in tension with orthodox (functionalist) philosophy of mind and cognition. If the agential system were a purely functionalist system then it ought to be realizable in any kind of different "hardware". Yet if values arise from the self-creation and self-maintenance of a particular system, then if you abstract that system from the physical basis that it creates and with which it maintains itself, then those values cease to exist for that system (Di Paolo 2009). This might seem at odds with the fact that our definition of agency is based on organizational criteria and therefore ought to be able to be instantiated in a variety of different systems - not only living ones (for a more detailed discussion of related concerns, see Wheeler, 2011 and Thompson 2011). However, while the organization of such an agent is indeed relatively independent with respect to its particular physical realization, this does not mean that anything goes, as would be the case for functionalism. For example, the essential role of mortality in meaning-generation entails that agency cannot be completely divorced from a precarious existence in some material substrate (Di Paolo 2009; Froese, in prep.), and this necessity of far-from-equilibrium self-production and self-maintenance imposes even more specific material constraints. For example, the material of the components of the system cannot be inert, but robots are typically built from such material (Moreno & Etxeberria 2005). And it is the particular material instantiation we suggest - what Stapleton (2012) dubs a system's "particular embodiment" - that gives rise to the particular values and norms of that system.

Advocating that it is a system's particular embodiment that matters for agency, it should be noted, is not *in principle* at odds with the functionalist approach to embodiment. For example in Clark's (1989) work on "microfunctionalism" he argues that functionalism need not be identified with formal descriptions pitched at a gross level, but that what is essential to functionalism is merely that the "structure, not the stuff, counts" (Clark 1989, p. 31). In a similar vein Wheeler, drawing on the University of Sussex evolutionary hardware and robotics paradigm of A. Thompson (e.g. Thompson 1995, develops Clark's (1997) notion of "continuous reciprocal causation" arguing that in evolved systems (both biological and artificial) the "low-level" properties of the hardware are relevant to adaptive success (Wheeler 2005, pp. 267-68)). For these reasons at least one of us (Stapleton) is inclined towards what she calls a "nanofunctionalist" paradigm (see Stapleton 2012, Chapter 5 and Conclusion). This is the position that the relevant level for understanding cognition in natural cognitive systems is very close to (and in some cases entwined with) the implementational level. Nevertheless, what is important about this implementation is the (nano-)functional role it plays for the system. Such a position however runs the risk of irritating both traditional (functionalist) embodiment theorists who have explicitly rejected radical embodiment (see for example Clark 1999; Wheeler 2010) for not abstracting from implementation enough, and enactivists who reject the functionalist tradition for abstracting from implementation too much.

It is however indeed the case for enactivists that what we intuitively consider as cognition involves a certain amount of autonomy with regard to its bodily basis, for example the relative operational autonomy of the nervous system (Barandiaran & Moreno 2006). Another mark of the cognitive may be that it serves to decouple an agent from its environment by mediating its sensorimotor interactions (Fuchs 2011). This decoupling can be achieved in a variety of manners, and enactivism is currently exploring how such mediation can help to bridge the cognitive gap from basic agency to more full-blown forms of human action (see review by Froese 2012).

8.3. From affective agency to subjective self

Given that the enactive conception of agency already involves an appeal to the normativity of adaptive behavior, it also provides a useful foundation for grounding the affective dimension of animal existence. Colombetti (2010), following Weber and Varela (2002), argues that the processes which subserve the biological (autopoietic) organization of living systems not only establish a "point of view" and locus of agency but also the enaction of meaning. The idea is that those parts of the world that are relevant to the self-production and self-maintenance of a system have meaning for the system. Here we are not concerned with linguistic semantics. Meaning is another aspect of the values and norms discussed earlier: all are generated from within the system as a result of its relation to those parts of the world it interacts with (i.e. its "Umwelt", according to von Uexküll's terminology). For something to be meaningful to the system is for it to have value for it, and thus for it to have a normative character. Colombetti notes that Di Paolo's (2005) addition of adaptivity to the previously binary (alive or dead) notion of autopoiesis allows us to account for the grades of meaning (or "degrees of value") offered by a system's environment, and thereby "makes room for a notion of organismic preferences" (2010, p. 149).

When we understand cognition in these terms, the self-regulatory metabolic (homeostatic) functions from which the values and meaning of a particular living system emerge are also those that ground emotion for neurobiological theorists such as Damasio (1999) and Panksepp (1998). Colombetti therefore concludes that "[o]n this view of emotion, the account of natural purposes developed by Weber & Varela (2002) and Di Paolo (2005) as a theory of bodily sense-making is as much a theory of emotion as it is a theory of cognition" (2010, p. 150). In general, enactivism provides a fitting framework for the tight integration of emotion, cognition and perception (e.g.

Colombetti 2007; 2014; Colombetti & Thompson 2008; Thompson & Stapleton 2009; Varela & Depraz 2005; Ward & Stapleton 2012; Bower & Gallagher 2013).

This integration of affect and agency can be seen clearly when we consider how value and action are integrated in organismic systems. In a very simple system such as a single cell the internal metabolic dynamics and those underpinning the cell's sensorimotor functions are not segregated very well - although precisely how independent they are from each other is debatable (Egbert, Barandiaran, & Di Paolo 2012). Nevertheless the internal value -- from the physiological condition of the cell -- is entwined with the action of the cell to maintain itself within its viability set. So even though it may not be a consciously feeling system, it is nevertheless right to think of it as an affective system: the world for the cell is shaped by this affect - it is what imbues the world with value for it, and thus what imbues the cell with normativity.

Once the agentive system is constituted by multiple cells, two issues arise: the complexity of the modulations required, which depend on a larger ensemble of physiological states, and the time it takes for processes in one part of the system to affect more distal processes. These increases in internal spatiotemporal distances and their bridging in terms of increased bodily sensing and regulation go hand in hand with an increase in decoupling from the environment, or at least an expansion of the role of self-mediation in action and feeling. We are therefore approaching the transition from mere affective agency to full-blown subjectivity. In the following we consider essential aspects of the internal organization of this more specific form of embodiment in some detail.

What is needed for multicellular systems then is some sort of homeostatic mechanism, which is both sensitive to the animal's internal milieu and able to regulate it effectively, while at the same time providing the basis for adaptive interaction with the world. In animals these functions are provided by the nervous system. While typically the term 'interoception' is used to refer to the sensitivity or awareness of internal, visceral changes as mediated by the autonomic part of the peripheral nervous system, we suggest that this is a contingent fact based on the predominance of research into humans. Interoception as the sense of the internal body after all, may be mediated through molecular communication networks in, for example, the endocrine and immune systems (Cameron 2001). That complex multicellular creatures like us require a functioning interoceptive nervous system as well does not detract from the basic structure of interoception being essentially a sensitivity to internal modulation which is needed in order to effect internal and external changes for the purposes of maintaining or increasing the adaptivity of the system. While this may just be another way of specifying the internal dynamics of autonomous, adaptive systems, we believe that understanding the mechanisms by which these organisational criteria are brought about in biological systems will yield an increased understanding of the requirements for more complex forms of agentive systems in general.

Two questions in particular seem to fall out of this. Firstly, how is interoception realized in humans and can this specific mechanism be operationalized beyond our particular embodiment? And secondly, how does the instantiation of an interoceptive "nervous" system in an agentive system qualitatively alter the system in terms of its agency? Finding answers to these questions is crucial for developing an enactive approach to agency that goes beyond an account of organismic agency in general, toward an account of animality more specifically and, most importantly, of human subjectivity. In the following we offer some considerations about how to begin answering this open challenge.

The term 'interoception' was introduced by Sherrington in the early 20th century and used by him to distinguish the sense of the visceral body from the sense of limb position (proprioception), the senses of touch, pain, and temperature (exteroception), sight and hearing (telereception) and taste and smell (chemoreception) (Sherrington 1948, cited in Craig 2002). Nowadays vision, hearing, smell and taste, like touch, tend to be categorized as exteroceptive,

which refers to them being concerned with information external to the body. On this kind of distinction, proprioception (and kinesthesia) ought to count as interoceptive, and they are often referred to as such. However under the orthodox categorization of the senses it would be more correct to think of proprioception and kinesthesia as belonging to the "somatic senses" which additionally include pain, temperature, itch and vestibular balance (Kandel, Schwartz & Jessell 2000).

Interoception is distinct from the other senses which take the internal body as their object (proprioception and kinesthesia) as it is typically used to refer to the afferent sensory information from the autonomic nervous system, such as heart muscle, other smooth muscle (but not skeletal muscle which is included in the somatic nervous system), and the exocrine glands (i.e. sweat glands, saliva glands, stomach, liver, pancreas). Craig (2002; 2003a; 2003b) however has argued against this traditional way of categorizing the senses, advocating that "interoception should be redefined as the sense of the physiological condition of the entire body, not just the viscera" (2002, p. 655). The reason for this is that recent research suggests that pain, temperature, and light touch are mediated by the same tracts in the spinal cord and subcortex as visceral information (Craig 2002; 2003b; Craig & Blomqvist 2002).

These two different ways that interoception is defined, i.e., as the sense of the visceral body and the sense of the entire physiological body, are underpinned by different mechanisms in organisms that have developed nervous systems for mediating the sensorimotor signals. After all, a cell's 'entire physiological body' just *is* its visceral body. In humans, however, this difference in the term's scope is important. While changes in the viscera that threaten the system's viability are dealt with by adapting internally, on their own they do not seem to intrinsically motivate action, at the very least not within the time scales with which we would normally judge agency. The receptors that mediate pain, light touch, and temperature, however, are hooked up so that not only can activation induce a spinal reflex (in very short time-scales) but their shared afferent pathway (with the visceral neurons) projects to motor areas in the brain such as the limbic motor cortex (ACC) for activation of movement over (still short but) longer time-scales (see Craig 2002, for a comparison of his 'spinothalamocortical pathway' with the conventional pain pathways). This interoceptive-motor loop grounds what Craig calls 'homeostatic emotions' which arise from threats to the viability of the system which homeostasis on its own cannot resolve (Craig 2003a). Craig considers these to be homeostatic 'emotions' because both motivations *and* sensations are generated. In addition to projections from this pathway to the limbic motor cortex there are also projections to the limbic sensory cortex (insula) and in primates to the interoceptive cortex and from there to the right anterior insula (Craig 2003a, 2003b). This is of particular interest to us here because activity in the anterior insula cortex is consistently associated with subjective feelings and may even provide a basis for Sherrington's 'material me' (Craig 2003b). This supports our argument that agency and affectivity are fundamentally entwined and not only at the level of (first-order) autopoietic systems but also in complex organisms made up of various intersecting autonomous networks like ourselves – a meshwork of selfless (i.e., non-reified) selves (Varela 1991).

These insights from the neurobiology of interoception in primates allow us to gesture towards how it might be operationalized. An abstraction from its particular instantiation may be used to understand the different levels of complexity of agency in biological systems and their intrinsic affectivity. Furthermore it can guide us in modeling agency for the purposes of artificial cognitive systems research and robotics. For now we wish to emphasize one particular point: while the instantiation of an interoceptive nervous system does not necessarily give an organism *more* agency than simpler agentive organisms (at least from what we have proposed so far – either you are an agent or you are not), it does give it a qualitatively different kind of agency. This kind of agency is not merely affective but instantiates a reflective stance that is discussed by

phenomenologists in terms of subjectivity. The subject's lived body is no longer always transparently lived through as the background for activity in the world, because this agential world-directed perspective can be turned on itself – inwardly – such that the body becomes an explicit part of the subject's experiential world rather than its implicit mode of revealing that world. In this way a new horizon of actions opens up, involving the development of new forms of mediated self-regulation. Thus, if we wanted to think of agency as admitting of degrees, we could emphasize this more pronounced diversity and asymmetry in the domain of agent-environment relations, which puts even more weight on the side of the body of the agent, as an enhancement of agency (Stapleton & Froese, in press).

According to the enactive theory of agency, all agents are affective systems because of the normativity involved in their adaptivity and sense-making. But it takes a special form of decoupled or mediated regulation of internal and interactive processes in order for an agent to be a cognitive agent (Barandiaran & Moreno 2006). Following what we have said about interoception, it seems plausible to suggest that such a cognitive agent must at the same time be a feeling agent, i.e. an agent that is able to make sense of the status of its internal processes. Jonas ([1966] 2001) proposed that emotions co-emerge with distal perception in order to enable the mediated regulation of actions across the greater spatiotemporal distances that perceptual experience reveals in the world, and which can no longer be bridged by mere reactive behaviour. The same logic applies to the greater internal distances revealed by interoception, whereby I perceive the state of my body at the same time as I become distanced from merely being my body, and my actions are guided by an assessment of how I feel.

It is at this point that we can start talking about subjectivity, rather than just agency, since the self, at least in a minimal embodied form, has become an explicit part of the concerns that motivate and guide actions. Much more needs to be said about this transition from the biological foundations of agency to the phenomenological conditions of subjectivity, in particular with regard to full-blown self-awareness, but we hope to have shown that enactivism is in a good position to make progress on this difficult and deep question. What is clear is that the nervous system will play a central role in this story because its electrical activity liberates an agent's capacity of regulation from underlying metabolic and developmental constraints, thereby opening up a much wider range of internal and interactive actions and perceptions, while at the same time more tightly integrating the multicellular agent as a whole (Arnellos & Moreno, in press; Barandiaran & Moreno 2006). With respect to the question of the operational requirements of specifically human subjectivity, a deeper consideration of the constitutive role of social interactions and the pre-existing cultural context will become indispensable (Stewart 2010; Kyselo 2014). Yet the biological foundations of subjectivity will remain essential even as the socio-cultural extensions of the self begin to play a significant role, which is why we are sceptical of the possibility of a genuinely collective subject – an agent consisting of a social network of other agents without material integration but having its own first-person experience (Stapleton & Froese, in press). The complex material and functional requirements for the emergence of such an integrated subjective 'self' from the perspective of a multicellular system as a whole (Arnellos & Moreno, in press) are unlikely to be repeated at the purely social level due to its looser interactions.

8.4. Conclusion

For an autopoietic system to constitute an agent there must be sufficient internal communication to allow the system to be sensitive to its own internal dynamics and move in response to them and to perturbations in the environment that threaten their boundaries of viability. Subjectivity requires something more, namely an interoceptively grounded form of sense-making. We have briefly described how this interoception is realized in the case of human agents. It is an open challenge for enactivism to operationalize this physiology in such a way that we can better grasp the essential

dynamics that are being realized by the autonomic nervous system, and to determine in what manner these dynamics could be realized in other forms of embodiment. At the same time the enactive approach points beyond physiology to the essential socio-cultural dimensions of selfhood, which suggest that the emergence of human subjectivity on the basis of organismic agency cannot be fully understood in terms of changes in biological embodiment alone. The self certainly cannot be reduced to the brain alone, but neither is it limited by the boundaries of the body: there can be no self without others.

Acknowledgements

The authors thank Nathanial Barrett for his helpful comments on, and suggestions for, the early drafts of this paper.

References

Arnellos, A. and A. Moreno. Forthcoming. Multicellular agency: an organizational view. *Biology & Philosophy*.

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

Barandiaran, X. and M.D. Egbert. 2014. Norm-establishing and norm-following in autonomous agency. *Artificial Life, 20*(1): 5-28.

Barandiaran, X. and A. Moreno. 2006. On what makes certain dynamical systems cognitive: a minimally cognitive organization program. *Adaptive Behavior, 14*(2): 171-185.

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

Barbaras, R. 2002. Francisco Varela: a new idea of perception and life. *Phenomenology and the Cognitive Sciences, 1*: 127-132.

Barbaras, R. 2005. *Desire and Distance: Introduction to a Phenomenology of Perception*, trans. P.B. Milan. Stanford, CA: Stanford University Press.

Barbaras, R. 2010. Life and exteriority: the problem of metabolism. In *Enaction: Toward a New Paradigm for Cognitive Science*, ed. J. Stewart, O. Gapenne and E.A. Di Paolo. Cambridge, MA: The MIT Press.

Beer, R.D. 2000. Dynamical approaches to cognitive science. *Trends in Cognitive Sciences, 4*(3): 91-99.

Bower, M. and S. Gallagher. 2013. Bodily affects as prenoetic elements in enactive perception. *Phenomenology and Mind, 4*(1): 108-131.

Cameron, O.G. 2001. Interoception: the inside story—a model for psychosomatic processes. *Psychosomatic Medicine*, *63*(5): 697–710.

Clark, A. 1989. Microfunctionalism: Connectionism and the Scientific Explanation of Mental States. Research Paper. Retrieved July 17, 2011. http://www.era.lib.ed.ac.uk/handle/1842/1332

Clark, A. 1997. *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Clark, A. 1999. An embodied cognitive science? *Trends in Cognitive Sciences*, *3*(9): 345–351.

Colombetti, G. 2007. Enactive appraisal. *Phenomenology and the Cognitive Sciences, 6*: 527-546.

Colombetti, G. 2010. Enaction, sense-making, and emotion. In *Enaction: Toward a New Paradigm for Cognitive Science*, ed. J. Stewart, O. Gapenne and E.A. Di Paolo. Cambridge, MA: The MIT Press.

Colombetti, G. 2014. *The Feeling Body: Affective Science Meets the Enactive Mind*. Cambridge, MA: The MIT Press.

Colombetti, G. and E. Thompson. 2008. The feeling body: toward an enactive approach to emotion. In *Developmental Perspectives on Embodiment and Consciousness*, ed. W.F. Overton, U. Müller and J.L. Newman. New York, NY: Lawrence Erlbaum.

Craig, A.D. 2002. How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*(8): 655–666.

Craig, A.D. 2003. A new view of pain as a homeostatic emotion. *Trends in Neurosciences*, *26*(6): 303–307.

Craig, A.D. 2003. Interoception: the sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13*(4): 500–505.

Damasio, A. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. London, UK: Vintage.

Damasio, A. 2010. *Self Comes to Mind: Constructing the Conscious Brain*. Knopf Doubleday Publishing Group.

Desmidt, T., M. Lemoine, C. Belzung and N. Depraz. 2014. The temporal dynamic of emotional emergence. *Phenomenology and the Cognitive Sciences, 13*(4), 557-578.

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

Di Paolo, E.A. 2009. Extended life. *Topoi, 28*(1): 9-21.

Di Paolo, E.A. 2010. Robotics inspired in the organism. *Intellectica, 1-2*(53-54), 129-162.

Di Paolo, E.A., M. Rohde & H. De Jaegher 2010. Horizons for the enactive mind: values, social interaction, and play. In *Enaction: Toward a New Paradigm for Cognitive Science*, ed. J. Stewart, O. Gapenne and E.A. Di Paolo. Cambridge, MA: MIT Press.

Di Paolo, E. and E. Thompson. 2014. The Enactive Approach. In *The Routledge Handbook of Embodied Cognition*, ed. L. Shapiro. Routledge Press.

Egbert, M.D., X. Barandiaran and E.A. Di Paolo. 2012. Behavioral metabolution: the adaptive and evolutionary potential of metabolism-based chemotaxis. *Artificial Life, 18*: 1-25.

Froese, T. 2010. From cybernetics to second-order cybernetics: a comparative analysis of their central ideas. *Constructivist Foundations, 5*(2): 75-85.

Froese, T. 2012. From adaptive behavior to human cognition: a review of *Enaction*. *Adaptive Behavior, 20*(3): 209-221.

Froese, T. 2014. Bio-machine hybrid technology: a theoretical assessment and some suggestions for improved future design. *Philosophy & Technology*, *27*(4): 539-590.

Froese, T. (in prep.). Life is precious because it is precarious: death and the problem of meaning for computationalism.

Froese, T. and E.A. Di Paolo. 2011. The enactive approach: theoretical sketches from cell to society. *Pragmatics & Cognition, 19*(1): 1-36.

Froese, T. and S. Gallagher. 2010. Phenomenology and artificial life: toward a technological supplementation of phenomenological methodology. *Husserl Studies, 26*(2): 83-106.

Froese, T. and J. Stewart. 2010. Life after Ashby: ultrastability and the autopoietic foundations of biological individuality. *Cybernetics & Human Knowing, 17*(4): 83-106.

Froese, T., N. Virgo, and T. Ikegami. 2014. Motility at the origin of life: its characterization and a model. *Artificial Life, 20*(1): 55-76.

Froese, T. and T. Ziemke. 2009. Enactive artificial intelligence: investigating the systemic organization of life and mind. *Artificial Intelligence, 173*(3-4): 366-500.

Fuchs, T. 2011. The brain - a mediating organ. *Journal of Consciousness Studies, 18*(7-8): 196-221.

Gallagher, S. 2005. *How the Body Shapes the Mind*. New York, NY: Oxford University Press.

Gallagher, S. 2012. *Phenomenology*. Basingstoke, UK: Palgrave Macmillan.

Gallagher, S. (2013). The socially extended mind. *Cognitive Systems Research, 25-26*, 4-12

Hanna, R. and E. Thompson. 2003. The mind-body-body problem. *Theoria et Historia Scientarum, 7*(1): 23-42.

Heidegger, M. [1929]. 1995. *The Fundamental Concepts of Metaphysics: World, Finitude, Solitude*. Bloomington, IN: Indiana University Press.

Husserl, E. [1952]. 1989. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. Second Book: Studies in the Phenomenology of Constitution,* trans. R. Rojcewicz and A. Schuwer. Dordrecht, Holland: Kluwer Academic Publishers.

Jonas, H. [1966]. 2001. *The Phenomenon of Life: Toward a Philosophical Biology*. Evanston, IL: Northwestern University Press.

Kandel, E.R., J.H. Schwartz and T.M. Jessell. 2000. *Principles of neural science*. McGraw-Hill, Health Professions Division.

Kessel, T. 2011. *Phänomenologie des Lebendigen: Heideggers Kritik an den Leitbegriffen der neuzeitlichen Biologie*. Freiburg, Germany: Karl Alber.

Kyselo, M. (2014). The body social: an enactive approach to the self. *Frontiers in Psychology, 5*(986). doi: 10.3389/fpsyg.2014.00986

Lymer, J. 2011. Merleau-Ponty and the affective maternal-foetal relation. *Parrhesia, 13:* 126-143.

Maturana, H.R. and F.J. Varela. 1987. *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston, MA: Shambhala Publications.

McGregor, S. and N. Virgo. 2011. Life and its close relatives. In *Advances in Artificial Life: 10th European Conference, ECAL 2009*, ed. G. Kampis, I. Karsai and E. Szathmáry. Berlin, Germany: Springer-Verlag.

Merleau-Ponty, M. [1942]. 1983. *The Structure of Behavior*. Pittsburgh, PA: Duquesne University Press.

Moreno, A., & Etxeberria, A. (2005). Agency in natural and artificial systems. *Artificial Life, 11*, 161-175.

Panksepp, J. 1998. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. New York, NY: Oxford University Press.

Parisi, D. (2004). Internal robotics. *Connection Science, 16*(4), 325-338.

Plessner, H. [1928]. 1975. *Die Stufen des Organischen und der Mensch: Einleitung in die philosophische Anthropologie*. Berlin, Germany: Walter de Gruyter & Co.

Reddy, V. 2003. On being the object of attention: implications for self-other consciousness. *Trends in Cognitive Sciences, 7*(9): 397-402.

Rohde, M. and J. Stewart. 2008. Ascriptional and 'genuine' autonomy. *BioSystems, 91*(2): 424-433.

Sartre, J.-P. [1960]. 2004. *Critique of Dialectical Reason. Volume One: Theory of Practical Ensembles*, trans. A. Sheridan-Smith. London, UK: Verso.

Scheler, M. [1928]. 2008. *The Human Place in the Cosmos*, trans. M.S. Frings. Evanston, IL: Northwestern University Press.

Sherrington, C. 1948. *The Integrative Action of the Nervous System*. Cambridge, UK: Cambridge University Press.

Stapleton, M.L. 2012. *Proper embodiment: the role of the body in affect and cognition* (PhD Dissertation). University of Edinburgh. Retrieved from Edinburgh Research Archive: http://hdl.handle.net/1842/6396.

Stapleton, M.L. and T. Froese. Forthcoming. Is collective agency a coherent idea? Considerations from the enactive theory of agency. In *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation*, ed. C. Misselhorn. Berlin: Springer.

Stewart, J. 1992. Life = Cognition: The Epistemological and Ontological Significance of Artificial Life. In *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, ed. F.J. Varela and P. Bourgine. Cambridge, MA: MIT Press.

Stewart, J. 2010. Foundational Issues in Enaction as a Paradigm for Cognitive Science: From the Origin of Life to Consciousness and Writing. In *Enaction: Toward a New Paradigm for Cognitive Science*, ed. J. Stewart, O. Gapenne and E.A. Di Paolo. Cambridge, MA: The MIT Press.

Thompson, A. 1995. Evolving electronic robot controllers that exploit hardware resources. In *Advances in Artifical Life: Third European Conference on Artificial Life*, eds. F. Morán, A. Moreno, J.J. Merelo and P. Chacón. Berlin: Spinger.

Thompson, E. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Thompson, E. 2011. Reply to commentaries. *Journal of Consciousness Studies, 18*(5-6): 176-223.

Thompson, E. and M. Stapleton. 2009. Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi*, *28*(1): 23–30.

Torrance, S. and T. Froese. 2011. An inter-enactive approach to agency: participatory sense-making, dynamics, and sociality. *Humana.Mente, 15*: 21-53.

Varela, F.J. 1979. *Principles of Biological Autonomy*. New York, NY: Elsevier North Holland.

Varela, F.J. 1991. Organism: a meshwork of selfless selves. In *Organism and the Origins of Self*, ed. A.I. Tauber. Dordrecht: Kluwer Academic Publishers.

Varela, F.J. 1996. Neurophenomenology: A Methodological Remedy for the Hard Problem. *Journal of Consciousness Studies, 3*(4): 330-349.

Varela, F.J. 1999. The specious present: a neurophenomenology of time consciousness. In *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, ed. J. Petitot, F.J. Varela, B. Pachoud and J.-M. Roy. Stanford, CA: Stanford University Press.

Varela, F.J. and N. Depraz. 2005. At the source of time: valence and the constitutional dynamics of affect. *Journal of Consciousness Studies, 12*(8-10): 61-81.

Varela, F.J., H.R. Maturana and R. Uribe. 1974. Autopoiesis: the organization of living systems, its characterization and a model. *BioSystems, 5*: 187-196.

von Uexküll, J. 1909. *Umwelt und Innenwelt der Tiere*. Berlin, Germany: Julius Springer.

von Uexküll, J. [1934]. 1957. A stroll through the worlds of animals and men: a picture book of invisible worlds. In *Instinctive Behavior: The Development of a Modern Concept*, ed. C.H. Schiller. New York, NY: International Universities Press.

Weber, A. and F.J. Varela. 2002. Life after Kant: natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences, 1*: 97-125.

Ward, D. and M. Stapleton. 2012. Es are good: Cognition as enacted, embodied, embedded, affective and extended. In *Consciousness in Interaction: The role of the natural and social context in shaping* consciousnes, ed. F. Paglieri. Amsterdam: John Benjamins Publishing Company.

Wheeler, M. 2005. *Reconstructing the cognitive world the next step*. Cambridge, MA: MIT Press.

Wheeler, M. 2010. Minds, things and materiality. In *The Cognitive Life of Things: Recasting the Boundaries of the MInd*, ed. L. Malafouris and C. Renfrew. Cambridge: McDonald Institute for Archaeological Research.

Wheeler, M. 2011. Mind in life or life in mind? Making sense of deep continuity. *Journal of Consciousness Studies, 18*(5-6): 148-168.

Zahavi, D. 1999. *Self-Awareness and Alterity: A Phenomenological Investigation*. Evanston, IL: Northwestern University Press.

Zahavi, D. 2011. Mutual enlightenment and transcendental thought. *Journal of Consciousness Studies*, *18*(5-6): 169-175.