Chapter 12
Is Collective Agency a Coherent Idea? Considerations from the Enactive Theory of Agency

Mog Stapleton and Tom Froese

12.1 Introduction: The Enactive Approach

The enactive approach[1] to cognitive science is characterized by being grounded in biology, phenomenology and principles like self-organization and autonomy that were developed in the second-order cybernetics movement. Whereas cybernetics focused on the observation and study of feedback systems, second-order cybernetics tried to also account for the possibility of the observer herself (Froese 2010). Enactivism is an inherently interdisciplinary approach to cognition rather than a

[1] Note that the term "enactivism" has recently come to be used in several ways. Here we use it to refer to the paradigm heavily influenced by Maturana and Varela (1987) and formally instigated with the introduction of the term in Varela et al. (1991). This has been described as "autopoietic enactivism" by Hutto in order to distinguish it from his theory which he calls "radical enactivism" (Hutto and Myin 2013) and from sensorimotor enactivism (Noë 2004). While it is useful to distinguish these streams of research, the term "autopoietic enactivism" is somewhat misleading as although the theory of autopoiesis has been a strong inspiration for researchers in this paradigm, not all accept that autopoiesis is necessary and/or sufficient for cognition (for this debate see Froese and Di Paolo 2011; and the discussions in Thompson 2011; Wheeler 2011). It is therefore perhaps better to refer to it as "biological enactivism" in order to distinguish it from the other streams. For the purpose of this paper we do not draw on these other streams and will use the term "enactivism" as it was originally introduced and as it continues to be used by the main propagators of this approach (Varela et al. 1991; Thompson 2007; Di Paolo 2005; 2009a; Di Paolo and Thompson 2014).

M. Stapleton (✉)
Institute of Philosophy, University of Stuttgart, Seidenstraße 36, Stuttgart 70174, Germany
e-mail: stapleton@philo.uni-stuttgart.de

T. Froese
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas/Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City, Mexico

© Springer International Publishing Switzerland 2015

C. Misselhorn (ed.), *Collective Agency and Cooperation in Natural and Artificial Systems*, Philosophical Studies Series 122,
DOI 10.1007/978-3-319-15515-9_12

mere bringing together of several disciplines. This is reflected in the variety of disciplines in which the enactivist paradigm is used to guide research, such as artificial life and robotics (Di Paolo 2003; 2005; Barandiaran et al. 2009; Morse et al. 2011), social and developmental psychology (Reddy 2008; De Jaegher et al. 2010; Froese and Fuchs 2012) psychiatry (de Haan et al. 2013), sociology (Protevi 2009), and philosophy of mind (Thompson 2007; Gallagher and Zahavi 2008).

The enactivist paradigm differs from the received view in these disciplines because it does not condone starting our investigation into the mind by abstracting away from our biological body, as has been the standard (cognitivist and functionalist) approach in the cognitive sciences broadly construed. Instead, our biology is taken seriously, as the basis from which to start an investigation into the nature of mind. Cognition is thus, for enactivists, a fundamentally embodied phenomenon in the strongest sense of the term. As we will see, agency is likewise grounded in this "deep embodiment".[2] And, herein lies the crux: if cognition and agency are fundamentally embodied phenomena how could there be such a thing as "collective agency"?

While enactivism has its roots in the theoretical framework of second-order cybernetics (see Froese 2010), the starting point for enactivism proper was the research into the autopoietic organization of living cells (e.g., Maturana and Varela 1987). It was argued that the cell is the minimal living (and cognitive) system because it metabolically produces itself as an individual in its own right along within that individual's domain of possible interactions. What this means is that the processes within the cell create the very boundary which enables these processes to continue to produce both themselves and the boundary, while also maintaining viable interactions with the environment. It is this self-organising and self-producing organization of matter that is defined as "autopoiesis." Abstracting from the material instantiation of this organization – but maintaining the property of operational closure, i.e., conditions whereby processes depend on each other for their continuation – yields a form of organization which is defined as "autonomous." This organization was first derived by Varela from a study of cellular metabolism and other biological networks, but it does not need to exclusively be instantiated in natural biological systems (for an introductory overview of the concepts of autopoiesis and autonomy see Di Paolo and Thompson 2014). The upshot of this is that despite the enactive approach being grounded in the theory of autopoiesis and thus being a fundamentally biological, and embodied, approach, it remains an open question as to whether artificial systems (or indeed a collection of biological systems) could instantiate an autonomous or autopoietic organization (Froese and Ziemke 2009).

[2] The term "deep embodiment" is taken from Ezequiel Di Paolo's (2009) ShanghAI Lecture available at http://shanghailectures.org/guest-lectures/43-presenter/177-ezequiel-di-paolo. It refers to the fact that embodiment is taken as ontologically essential for mind, rather than as just a contingent functional extension of mind that could be separated from it, like a tool.

12.2 The Enactive Theory of Agency

What is the relevance of this self-organising and self-producing organization for agency? To see this, consider our pre-theoretical notion of agency. At the very least the term 'agent' implies (1) an individual, and (2) a capacity for action. This minimal notion is the one often used in informatics and robotics where 'agent' is standardly used to refer to any robot (or indeed software) that has a particular function and 'acts' so as to achieve this. There are two things to note here in regard to the use of 'agent' to refer to these systems. Firstly, the demarcation of the system in question is heteronomous, because its boundaries are only defined externally. That is to say, what counts as the individual agent depends upon an outside perspective and the interests of the observer. Secondly, such "agents" have of course been designed and programmed by humans to act in whatever way they do, be it by explicit design or by artificial evolution. They do not act in order to satisfy intrinsic needs and so their existence as agents does not directly depend on what they do; whether their movements happen to satisfy the conditions of successful behavior is decided by an external observer. Thus, even though nobody may be touching or controlling them remotely as they 'act', these 'actions' can still be seen as something external to that system (see Froese et al. 2007, for a related discussion of how the concept of 'autonomy' is used in robotics).

Enactivism offers a principled way of grounding individuality and action in a system. As outlined above, an autopoietic organization can be defined as a system's capacity to produce its own boundary. This constitutes a system as an individual in its own right with its own domain of interactions. Take the paradigmatic case of the bacterium. Even though, depending on what our explanatory project is, we can zoom in to look at parts within its body, or zoom out to include parts of the environment such as the whole Petri dish in which it is swimming, it is nevertheless the case that the bacterium constitutes itself as an individual in a way that the systems of the other two perspectives do not. There is nothing that indicates those other systems as such to be anything but dependent on our distinctions. Only when we distinguish the bacterium as being delimited by its self-produced network of metabolism and boundary does the bacterium appear to us as self-distinguishing; it is ontologically individuated from the environment. Ontological individuation in this respect refers to more than an appropriate posit relative to an explanatory project; it is a strong claim about the fundamental status of minimal living systems. On this view, ontological individuation is necessarily based on self-individuation, of which autopoiesis is one fundamental example.

In theory at least, it is possible that the material constituents of an autopoietic system could come together with the right kind of organization, and if all its metabolic needs were provided for in its direct environment then that organization could be maintained without the need for it to move. Such a system however needs an extremely stable environment to survive (an example of this in practice are some endosymbiotic bacteria, see Barandiaran and Moreno 2008). As soon as the external environment is anything short of benevolent the system

will disintegrate and no longer exist. This sharp binary between life and death in autopoiesis has been addressed by Di Paolo (2005) who introduced the notion of adaptivity to enactive theory.

Following Di Paolo (2005), if we understand the viability set of a system to be the set of changes that can happen in the environment and within the system itself without the system's organization breaking down, then adaptivity is the property that a system has of being both sensitive to and able to regulate these changes such that if they are leading towards dissipation the system can change itself (adapt) in order that it can evade dissolution. This could happen in two ways. The system could adapt itself internally such that it, for example, improves its chemotactic ability by incorporating a new compound into its metabolism (Egbert et al. 2012). Or, and this is the property of interest to us here, it could regulate its interactions by moving itself away from a noxious environment and/or towards a beneficial environment (Barandiaran and Moreno 2008). This is not a mere passive movement of a system that is subject only to external modulations but the action of a system moving itself according to its intrinsic needs. This is the foundation of the enactive account of agency.

Barandiaran et al. (2009) develop this idea of intrinsic agency in adaptive systems and propose an operational definition of agency on which there are three necessary and (jointly) sufficient conditions: individuality, interactional asymmetry, and normativity. The requirement of individuality falls out of what we discussed in the preceding paragraphs; the claim is that to be an agent a system must distinguish itself as an individual rather than merely having individuality thrust upon it from our perspective. The enactive theory of agency argues that systems with autopoietic organization are a paradigm case of genuine individuality because they generate their own boundaries which allow the very processes which generate these boundaries to continue, in a circularly causal spiral. The second condition, interactional asymmetry, is the requirement that the system must be able to change its relationship to the environment, and that changes be actively generated more from within rather than being a passive result of external forces. Finally, the third condition, normativity, is the requirement that action is guided by internal norms. These internal norms are the goals that arise as a result of the intrinsic needs of the system. A bacterium, for example, may be guided to move up a sugar gradient because of its metabolic needs; to continue existing and not succumb to dissolution the bacterium must metabolise the sugar. The normativity lies in the intrinsic value that these goals (in this case metabolizing sugar) have for the continued existence of the system.

In summary, an enactive theory of agency (following Barandiaran, Di Paolo & Rohde) proposes that an agentive system is one which creates itself as an individual, adapts to changes in the environment by manipulating itself in that environment more than the environment manipulates it, and is moved to do so in order to satisfy needs that arise internally, needs that result from the system undergoing self-generation under far-from-equilibrium conditions, i.e., conditions which continually threaten its existence.

12.3 Multi-agent Systems, Multi-system Agents, and Multi-agent Agents

The enactive theory of agency provides us with conditions for a system being an agent. It might seem that a minimal agentive system, according to these criteria, must be autopoietic and adaptive; in short: a biological organism (even if a mere single-celled one). While this may turn out to be the case, it does not directly follow from the Barandiaran et al. proposal, indeed they specifically state that "agency does not have to be subordinated to biological/metabolic organization but can appear at different scales responding to a variety of autonomous processes…" (pp. 8–9). While all autopoietic systems instantiate an autonomous organization there might also be systems which instantiate this autonomy without necessarily being grounded in autopoiesis in the chemical domain. We can therefore see that even though the enactive conception of agency is grounded in biological embodiment it nevertheless allows for the possibility of 'collective agents'. That is to say, on the face of it there does not seem to be a contradiction between the conception of collectivity or distributedness and the enactive conception of agency: as long as the collective system fulfills the conditions of individuality, interactional asymmetry, and normativity it may be considered to be a genuine agent.

12.3.1 Multi-agent Systems vs. Multi-system Agents

Recall that we are here using the term collective agency in the sense of a system made up of agents, which itself – as a system – is agentive. We must therefore distinguish collective agency from two other similar concepts that are easy to conflate: multi-agent systems and multi-system agents. Multi-agent systems are systems which are composed of multiple independent software 'agents', each of which has its own goal or function to fulfill. These 'agents' interact – or at least communicate – with each other in order to negotiate their activity and make a contribution to (what at least on the surface might seem to be) the 'goal' of the larger system. In informatics, because the term 'agent' is used so broadly, there is no problem in identifying such a system as a multi-agent agent – or 'collective agent'.

But, if we insist upon a deep notion of agency, such as the enactive one proposed by Barandiaran et al., then these systems will fail to be agents at both levels: neither the subsystems nor the group 'agent' appropriately distinguish themselves ontologically from the environment so as to fulfill the criteria of autonomous (rather than heteronomous) individuality, nor are their interactions and goals endogenously created, serving their persistence as a system. This is not to say that it is in principle impossible for us to design the conditions of a multi-'agent' system whose emergent properties satisfy the enactive conditions of agency, but in practice this has not been achieved so far (see Froese and Di Paolo 2008 for an initial attempt). The problem is that we are faced with the task of engineering second-order emergence (Froese

and Ziemke 2009): the emergent behavior of the interacting 'agents' has to provide the conditions for the emergence of a new agent. And even if we managed to overcome this problem in practice, we would still only end up with a multi-system agent, i.e., a genuine agent at the emergent group level but composed of systems which do not fulfill the criteria for agency. Such a system would fall short of genuine collective agency.

12.3.2 Swarms as Multi-agent Systems

Consider however a case where the individual agents of a multi-agent system really are agents in the full (enactive) sense of the term, as for example in the case of swarms of insects, birds, or fish. Seemingly very complex behavior can emerge at the group level as a result of local interactions between these individual agents. Do we have evidence however, to think that the swarm as a whole is a well-defined entity, that it regulates its domain of interactions, and that it generates its own norms, according to which it's activity is guided? Although it may be possible to view swarms as generating themselves (it is after all the own dynamics of, say a tornado that keeps itself going) is it right to think of them as generating themselves as individuals?

In the case of simulated flock behavior (Reynolds 1987), which presumably generalizes to biological systems, it has been shown that swarm behavior can be obtained by combining three rules: collision avoidance, velocity matching, and flock centering. If moving agents follow these three rules, then (presumably with the addition from time-to-time of some deviance from within or outside the system) this seems to be sufficient for generating the patterns in flocking behaviour that strike us as seeming so complex as to be somehow a movement of a whole rather than of an aggregation of individuals (think for example of the murmuration of starlings). Of course such groups do not generate the individuals that compose them (in the timescale of the swarm), but that would only be a prerequisite for *autopoietic* organization and not for *autonomous* organisation which is what underpins the criterion of individuality. Likewise, there may not be a physical boundary of the system in place, but it does seem right to think of such swarms, or flocks, as nevertheless forming themselves as a system such that their systemhood is not heteronomous; it is not just thrust upon them by an observer but rather is a "real pattern" (see Dennett 1991). It is an open question as to whether such real patterns fulfill the criterion of individuality, after all if we are to allow for the possibility of artificial agents then we cannot stipulate, from the outset, that all individuality must result from a system creating its own boundary in autopoietic fashion.

Similarly, it does not seem entirely unintuitive to think that a flock's interaction with the environment might, for the most part, be generated more from within the flock itself. Whether or not it is right to think of a swarm as having any goal of its own or not, it's movement is not solely due to environmental perturbations, though these may play an important role in stimulating, modulating, or even dissipating, the

group behavior. Even such minimal rule following as is suggested by Reynolds (1987) appears to be sufficient for producing a collective system that does not merely hang together as it is passively shunted around the environment, but rather moves itself.

So, we cannot rule out that a swarm may satisfy the criterion of individuality, and it seems likely that it at least satisfies the criterion of interactional asymmetry. What then of normativity? Recall that the criterion of normativity stipulated that for a system to have genuine agency it is not enough that it has goals and acts so as to achieve those goals. The goals of a genuine agent arise from the needs of the system itself. This is where the normativity comes from; the system *should* act in such a way that brings it closer to achieving these goals, not because any external force or agency wills it or designs it to be so, but in order for it to continue its existence as that system. Can we see this kind of normativity instantiated in simple swarms or flocks?

It is not clear how it could be. The constituent interactions of such a system are such that it rather seems that either they are all instantiated by the components of the system (whether these are genuine or 'as if' agents) in which case they give rise to swarm behavior, or they are not, in which case swarm behavior does not emerge. This swarm binary does not leave any room for adaptivity, which – as we saw in the previous section – is what grounds normativity. If there are no grades of survival then the system has only one need; exactly the right internal and environmental conditions to survive. If it were going to regulate these conditions either internally or externally, as would be adaptive, then it would need to somehow alter its internal dynamics. But altering the internal dynamics – when the only dynamics are the three rules that together give rise to swarminess, would result in altering those rules, and therefore no swarminess arising. In other words, in this kind of biological multi-agent system there are emergent dynamics at the group level, which may even fulfill the enactive definition of an autonomous system (Thompson 2007; Froese and Di Paolo 2011), but not the stricter requirements of agency.

12.3.3 Multi-agent Agents: Towards a Genuine Collective Agency

There are however, multi-agent systems that present more plausible candidates for instantiating agency at both the individual and group levels. Consider the eusocial insects, such as some kinds of ants, bees and termites, whose interactions give rise to colonies and hives with highly complex emergent properties. In fact, these groups of agents are so impressively complex and coordinated that they can appear to us as strongly animalistic; so much so that they are often referred to as "superorganisms" (Sterelny and Griffiths 1999, Chap. 8) In contrast to simple swarms, in eusocial systems there is a clear division of labor. The individuals are so specialized in their morphology and behavior that they cannot survive in isolation; they depend on each other for their existence. In addition, there is a clear colony boundary in place. Not

every individual of the same species can join any colony, since colonies are individuated by means of chemical markers. Indeed, in contrast to simple swarms, individuals act in the interest of the whole even to their own detriment, rather than just for the sake of their own lives.

If we are to grant that swarms satisfy the conditions of individuality and interactional asymmetry then we should be willing to grant that superorganisms also do so. They are not only self-organising systems (and thus generate themselves in the manner of keeping themselves going as a result of the dynamics that the system itself generates) but they also literally generate new components to keep the system going; they have offspring which are incorporated in to the system to fulfill specific functions. And, while it may seem that swarms interact with the environment more than the environment acts on the swarm, in the case of superorganisms this clearly seems to be the case as they manipulate the environment around them to provide suitable living and breeding space for the group. Could these groups of organisms then also generate their own internal normativity, and thus satisfy the final enactive criterion of agency?

The colony as a whole does indeed have a variety of irreducible properties, such as levels of food supplies, external threat and internal temperature, which emerge from the interactions of the individuals and at the same time modulate the behavior of those individuals so as to keep the global properties within the colony's range of viability. For example, there is a clear inside-outside division of a colony, which is normatively enforced by specialized individuals at dedicated boundary points. If the colony is under attack by intruders, the individuals will sacrifice themselves in order to neutralize the threat to the colony, similar to the function of some white blood cells in our bodies.

It could therefore be argued that such a group of social insects does manage to satisfy the enactive criteria of agency, and that it qualifies as a collective agent in a strong sense of the term. Their colonies are composed of individual organisms whose normativity as individuals is subsumed under the normativity of the colony as a whole. Yet it must also be noted that the higher-level of agency is only behaviorally integrated and individuated, in contrast to forming a single material structure. This is especially evident in terms of the colony's movement. If an ant colony shifts to a new location this is achieved by means of all ants individually moving to that new location. In other words, movement is not something specifically realized at the level of the colony as a whole. We therefore have an intermediate example of integration and individuation between (unintegrated) simple swarms and materially integrated systems such as multicellular organisms.

Multicellular organisms instantiate a stronger form of higher-level individuation. Typically multicellular organisms might be thought to be multi-system agents; built as they are out of cells, and modularized components such as organs. However, on the enactive view of agency it is not obvious that multi-cellular organisms are indeed mere multi-system agents rather than multi-*agent* agents. After all, as outlined above, a solitary bacterium satisfies Barandiaran et al.'s conditions for minimal agency: identity, interactional asymmetry, and normativity. Might it therefore be the case that some of the single cells in our body also satisfy these conditions?

The cells in our bodies constitute themselves through self-producing (autopoietic) means, just as was described earlier in the case of bacteria, such that they form themselves as an identity independently of an observer's perspective. It may not be so obvious however that they easily satisfy the asymmetry and normativity conditions for agency. Recall that the asymmetry condition is that adaptive regulation of the system must be powered – in general – more from intrinsic processes than environmental ones. In the case of a bodily cell it seems that much of this regulation is constrained by the bodily environment. In extreme cases the body can even cause some of its cells to commit 'suicide' (a process known as apoptosis) – an action which goes against the most basic of biological values, namely self-preservation.

To some extent such subordination must of course be the case – if the cells were not constrained sufficiently, no physically integrated whole would be possible in the first place. It is also of course partially in their "interest" to remain so constrained. If their replication, development and behavior are not subordinate to the needs of the whole system then they themselves will not survive for as long as may otherwise have been possible. For example, cell replication run amok is not conducive to maintaining bodily homeostatic balance, as is illustrated by aggressive forms of cancer. Do such internal environmental constraints, however mean that the relation between the cell and the body is asymmetric, with the locus of agency predominantly on the side of the overarching system (i.e., the body rather than the cell) such that the body effectively confiscates the subordinate cell's agency? And does this confiscation also mean that the activity of the cell is subordinate to the normativity of the body as a system, i.e. to the goals of the whole system rather than acting according to its own intrinsic norms?

It is difficult to provide definite answers to the questions raised in this discussion, but some general trends can nevertheless be identified. While a swarm may be too little individuated at the group level, and not generate enough endogenous normativity, to count as a multi-agent *agent*, a multicellular organism may be too tightly integrated at the group level, and its endogenous normativity overly constraining on the cells that make it up, to count as a multi-*agent* agent. A colony of eusocial insects however, seems to be situated somewhere along the middle of this spectrum and may therefore provide us with the clearest example of a genuine kind of collective agency. Each of these forms of interaction has advantages and disadvantages regarding the relative capacities of the parts and the wholes. Future work could apply the enactive concept of agency to the social world of humans, where cultural principles of integration are the predominant factor. For example, Steiner and Stewart (2009) have highlighted that from the autonomous perspective of human individuals, social norms appear as heteronomous, that is, as externally determined by cultural traditions. Interestingly, human societies have addressed the potential instabilities of a purely behaviorally integrated group level of agency by forming social institutions that are independent of the people passing through them, and which in the modern world even have legal representation as individuals in their own right. The foundation of a country creates a new individual entity that goes beyond its founding members, that has its own domain of interactions (for example with other countries), and that has its own normativity. However, this still falls short

of the traditional metaphor of society being an organism in its own right, and that is fortunate for us. As Di Paolo (2009b) has pointed out, we should be wary of trying to push the ideal of living within a super-organism too far. Moving human society along the collective agency spectrum by increasing the powers of the institutional agent as a whole, i.e. by shifting it from a multi-agent system toward a multi-agent agent, implies a significant reduction of our personal liberties.

12.4 Collective Agency Is Not an All-or-Nothing Concept

The temptation is to see agency in black and white terms. Either a system fulfills the conditions for minimal agency or it does not. Indeed by talking of necessary and sufficient conditions for agency it might seem that we are implicitly propagating this kind of black-and-white thinking. While it may be right to think that a system that never satisfies these conditions is not agentive, a system may not have to always satisfy these conditions in order to be an agent. It seems right to think of ourselves as agents (in this minimal sense) even when we allow ourselves to go with the environmental or psychological "flow", that is, we do not lose our agentive nature just because – for a period of time – our interaction with the environment is either non-asymmetric or asymmetrically powered by the environment. Consider even the example of minimal agency being instantiated in a bacterium: it does not unceasingly act one-sidedly in its environment but rather oscillates between actively moving "in search" of higher sugar gradients (or away from noxious substances) and being passively modulated by the environment in which it currently finds itself (see also Di Paolo and Iizuka 2008). Agents can also *actively* become what according to Barandiaran et al.'s definition would seem to be non-agents for a period of time and temporarily let the environment control them, as a lizard does when it plays dead until the cat gets bored of playing with it. Similarly, agents can allow external social norms to take precedence over their intrinsic norms, and then later return to following their own norms.

Our inclination towards a binary concept of agency may arise from the tendency to view systems at just one time-slice or in abstraction. When viewed statically either the system satisfies the conditions or it does not. But this does not give us an accurate description of the system. Agentive systems are dynamic; they exist through time not only in the trivial sense that all things that exist do so in time, but rather it is part of the very concept of 'action' that it takes place over a period of time. If we define agency as the regulation of interaction according to norms then we are dealing with an extended process by definition. This of course raises the question of what time period is appropriate to take into account if we want to assess whether a system is agentive. From our everyday point of view, those systems which unfold in the timescales that we are used to observing actions in will be the most likely to be attributed agenthood by us. That is, actions that unfold in seconds, minutes, days, or perhaps months. Actions that unfold in very tiny time scales, or over many years, decades or centuries do not intuitively seem agentive to us. We do

not naturally think of plants and trees for example as 'acting'. And yet, when we view films of plants growing in fast-forward this intuition begins to be a little undermined and they seem quite animalistic; they can move, they can climb, they can strangle other plants, they can shoot seeds, etc. To be sure, these actions may be more limited than those of animals, but the notion of interactional asymmetry does not have to be an all-or-nothing concept, either. It seems plausible that there are different grades of asymmetry. We must therefore be careful to not pre-judge agency on the basis of what seems to us to be the relevant time-scale. This is once more just a call for looking for an autonomous perspective rather than attributing agency on the basis of an observer's external considerations, except this time in reverse: rather than a false tendency to attribute agency to systems *that do not in fact have* intrinsic agency, our intuitions falsely guide us away from attributing agency to systems *that in fact do have* intrinsic agency.

Relatedly, the appearance of a conflict between sub-agents and the superordinate agent may arise as a result of viewing both at the same 'level' rather than acknowledging each in terms of their own ecological niche. The body taken as a whole must have interactional asymmetry with its *Umwelt*, and perhaps in so far as one of its cells ever becomes that *Umwelt* – in cases of cancer perhaps – the body must interact asymmetrically with that cell. But in normal functioning when we take the perspective of the cell, it is not interacting with 'the body' at all. Rather its *Umwelt* happens to be *in* the body. This body of course presents constraints on the cell's possibilities for action but surely this is not a case of the environment directly and irrevocably dictating the behavior of the cell, but rather just the presentation of a more constrained environment for that cell. From the cell's point of view, i.e. once we zoom in to consider the cell's activity in its own tiny *Umwelt*, we may no longer be inclined to think that it does not interact asymmetrically with its environment and therefore it is no longer unintuitive to think that it might be a genuine agent that partially constitutes a super-ordinate "collective" agent.

This is an important point and bears elaborating upon a little more. Our considerations of agency in this context reveal that agency in each system may be more, or less, visible at different levels of analysis. This means that it may not be the case that a genuine collective agency strikes us – from a single level of analysis, and at a single time period – as agentive both at the level of the group/collective and at the level of the components that make up that group. Add to this that, as we have argued, agency is a spectrum varying along dimensions rather than an all-or-nothing concept. We can therefore see that even if we define collective agency as a multi-agent agent in contrast to either multi-agent systems or a multi-system agents, we nevertheless are faced with a variety of kinds of agency in collective agents: agency varies in each case along the dimensions of individuality, interactional asymmetry and normativity both at the level of the individuals composing the group, and of the group itself. Depending on what time-slice you use to analyse the system, agency may be more or less visible. But nevertheless if a system is to be a genuinely collective system, according to our definition and as can be seen by the examples that we have presented, the variations along the dimensions of agency at each level mutually enable and constrain those at the other. Although what we consider to be the most

compelling examples of what might be a genuine collective agency are to be found in nature since the enactive concept of agency is not fully realized by current robotic 'agents', artificial examples are nevertheless not excluded by definition. Furthermore, cultural institutions that act as individuals in the social domain, and may even interact with their component individuals in, for example a court of law, may potentially fit the criteria.

12.5 Collective Agency Versus Collective Subjectivity

If we are to consider the possibility of genuine collective agency, rather than using 'agency' as a mere metaphor when it comes to the collective level, then we must also open ourselves up to considering whether such a thing as collective subjectivity may also exist. Even if we were to approach the topic from a non-enactive viewpoint this would seem warranted because in our paradigm case of agents – humans – the concept of agency is tightly interwoven with the interrelated concepts of mindedness, subjectivity, intentionality, and consciousness. The burden of proof should perhaps then be on showing that agency, or the processes upon which agency depends, does *not* entail subjectivity rather than the reverse. But of course the idea of collective subjectivity is a much harder pill to swallow than even that of collective agency. Let us then separate the question of subjectivity from that of consciousness, and address the question of whether a genuine collective agency on the enactive account might imply collective subjectivity even if not a collective consciousness. According to the enactive approach an organism's subjectivity is it's lived perspective, or in other words, it's meaningful point of view on the world (e.g., Weber and Varela 2002; Thompson 2007). Is there something it is like to be an ant colony or even a country like Germany? Do they have a subjective perspective and *Umwelt* of their own? Common intuition would deny this, but on what basis?

Let us consider the case of dyadic human interaction. Each of the individual participants is certainly a subject with a lived perspective, but what about the dyad as such? After many years of methodological individualism in cognitive science, which holds that the proper level of analysis is the individual (or even just their brain), there are now many paradigms emphasizing a deeper dynamic interconnectivity between people (Kyselo and Tschacher 2014; De Jaegher and Di Paolo 2007; Oullier and Kelso 2009; Riley et al. 2011). To give an example from our own work, Froese et al. (2013) used the minimal cognition approach developed by Beer (1996) and others to show that interacting robots become different kinds of systems while interacting. In this case each robot was only equipped with one artificial 'neuron', but via the interaction process they managed to extend each other's capacities such that the neurons exhibited oscillations and chaos – properties that are in principle impossible for a 1D continuous system, as realized by each robot's single neuron. For oscillations a minimum of two dimensions are necessary, while chaos requires a minimum of three. The fact that these properties were observed in each robot's neural activity therefore implies that the dimensionality of their artificial 'brains'

became mutually extended via their embodied social interaction. The two brains and their bodies and their interaction via the environment formed one system, thereby making it impossible to reduce a robot's neural activity to its neural system. What this minimal robotic model shows is that, in principle at least, nothing stands in the way of an extended body realizing a socially extended mind (Froese and Fuchs 2012).

One might think that this kind of socially extended mind suggests that it is at least a legitimate question to ask whether it is possible to also share each other's subjective perspectives, i.e. to give rise to a genuinely second-person perspective with its own social actions. It is important not to conflate these second-person interactions with joint action as conceived under the title of the "we-mode" (Galotti and Frith 2013). The idea behind Galotti and Frith's we-mode is that when we engage in joint actions we do not need to represent our individual goal, and the goals of the others with whom we act (indeed this may in some circumstances be counter-conducive). Rather we enter a particular "mode" in which we represent the task as one that we – as a couple or a group – are doing together so that the representation that guides each of us is a "we-representation" rather than an "I-representation". But this is not to say that the representation is "shared" between individuals in any interesting way. The we-representation is still a representation inside the individual and although each member of the group that is acting jointly must have a we-representation to be cooperating successfully in a joint action, and presumably these we-representations must have broadly similar contents, nevertheless it is certainly not "the same" representation that is shared by the actors. Galotti and Frith's "we-mode" therefore does not yield any interesting notion of shared or collective agency that is relevant to the question of whether the individuals do truly come together to form some kind of supra-individual agentive system with its own unique subjective perspective.

A possibly genuine form of shared intentionality and lived perspective is suggested by a psychological study conducted by Froese et al. (2014). Making use of a minimalistic virtual reality setup first proposed by Auvray and colleagues (2009), which has come to be known as the 'perceptual crossing' paradigm (Auvray and Rohde 2012), they explored the conditions under which people come to be aware of the presence of another person on the basis of differences in interaction dynamics alone. The interaction between a pair of spatially separated participants was mediated by a human-computer interface that reduced the scope of their interaction to moving an avatar in an invisible linear virtual space and receiving tactile feedback while overlapping with a virtual object. Participants were instructed to try to help each other to locate each other in the virtual space while avoiding distractor objects. One of these objects was static while the other was an exact, but unresponsive, instantaneous playback of the partner's avatar movements. They had 15 trials in which they were asked to click when they felt that they had succeeded in finding the other, and after each trial they were asked to rate the clarity of the experience of their partner if they had clicked. No feedback about the correctness of a click was provided during the experiment. Despite the reduced scope for social interaction and the presence of distractor objects, most pairs of participants were successful at solving the task in a majority of trials.

Two aspects of this outcome are of particular relevance to our discussion here. Firstly, it seems that success was largely a cooperative achievement. It was more common that both participants clicked correctly on the other than for one participant to succeed alone, and the delay between such jointly successful clicks was often less than a few seconds. In other words, although participants could not be directly aware of each other's clicks, the recognition of the other participant was highly synchronized between the participants, to the extent that we could interpret this as evidence of genuinely shared intentionality, i.e., mutual recognition of each other. Secondly, there was an experiential difference between two kinds of correct clicks, namely clicks occurring in jointly successful trials and clicks occurring in individually successful trials. Participants were much more likely to rate their experience of the other as being most clear after jointly successful clicks. The clarity of social awareness therefore had less to do with the objective presence of the other, and more with whether that awareness was mutually shared by both participants. In other words, it is suggested that the co-regulation of mutual embodied interaction also gave rise to a shared lived perspective, i.e., a second-person perspective.

This kind of collective, *second*-person subjectivity already goes beyond the traditional constraints of internalist-individualist cognitive science (Froese and Fuchs 2012). But does this study enable us to talk about the interactive constitution of a new subject with its own *first*-person perspective? This does not seem to be the case because, although there is a deep integration of minds, the participants do not lose their individual point of view on their shared situation during this process. We are left with the idea of a genuine second-person perspective, but not a collectively constituted first-person perspective. It is doubtful that it is any different for social interaction processes involving more participants, such as a football team or a nation state: under certain conditions of mutual interaction a multi-person perspective may be formed (the famous 'mob mentality'), but this is not sufficient for attributing a collective first-person perspective. The concepts of collective agency and collective subjectivity, where the latter refers to a collective agent with its own unique, unified and meaningful point of view on the world, may therefore not always coincide. The former seems necessary but not sufficient for the latter.

But if we allow the possibility of collective agency, why are we more skeptical about the possibility of collective subjectivity? Partly this has to do with the fact that, according to the enactive approach, subjectivity has more requirements than just agency. We have discussed the relationship between agency and subjectivity at length elsewhere (Stapleton and Froese ms.), so we will be brief here. Essentially, subjectivity depends on a specialized subsystem for the monitoring and regulating of internal processes, which we interoceptively experience as valence or emotion. In animals this system is realized as a special neural system that is spread throughout the entire body. It would be difficult to realize an operationally equivalent system on the basis of social interaction without some kind of physical integration that allows for the structuring of the necessary processes of monitoring and regulating in a stable manner. In other words, one important reason why we, as individual human beings, are collective subjects in addition to being collective agents is that our collective agency is realized as one physical

living body. Nevertheless, on closer inspection it may well turn out that we are dealing with another spectrum of possibilities, and that we should avoid becoming trapped in a way of thinking that assigns an absolute categorical difference between first- and second-person perspectives. One or the other type of perspective may become more or less prevalent depending on conditions. For an extreme example, we can consider that reports of how others are experienced by people with schizophrenia demonstrate how that phenomenology can vary from complete autistic isolation (Stanghellini and Ballerini 2004) to normal co-presence (during periods of relative well-being) and to complete fusion with others and self-dissolution (Lysaker et al. 2005).

12.6 Conclusion

Whether collective agency is a coherent concept depends on the theory of agency that we choose to adopt. We have argued that the enactive theory of agency developed by Barandiaran et al. (2009) provides a principled way of grounding agency in systems to which we already attribute it: biological organisms. The instantiation of the necessary and jointly sufficient conditions of individuality, interactional asymmetry, and normativity give rise to a system that is ontologically demarcated, endogenously active, and generates its own needs; its agency does not depend on the viewpoint of an observer nor does it exist only relative to a particular explanatory project. Enactivism, and therefore also the enactive theory of agency, is however grounded in biological embodiment which might lead one to be skeptical as to whether artificial systems or collectives of individuals could instantiate genuine agency. To explore this issue we contrasted the concept of collective agency with the ideas of multi-agent systems and multi-system agents, and argued that a genuine collective agency would instantiate agency at both the collective level and at the level of the component parts. We argued that although swarms present impressively complex behavior at the level of the collective system, this collective nevertheless fails to instantiate genuine agency because it does not generate its own normativity. We then considered the case of eusocial insect colonies, sometimes termed 'superorganisms' and multicellular systems like ourselves. Eusocial colonies, unlike swarms, may be seen to not only instantiate individuality and interactional asymmetry but also to generate endogenous normativity. While this would bring us to what might be considered to be a paradigm case of collective agency, we questioned whether the behavioural (rather than material) integration and individuation was quite strong enough to instantiate the individuation required for genuine agenthood at the collective level. In contrast, we questioned whether the material integration of cells in multicellular organisms such as ourselves might actually be *too tight*, and the normativity generated by the collective *too strong* to allow for genuine agency of the components of the collective. We proposed that agency cannot be judged at a single time-slice and that we should therefore understand agency as a spectrum that varies along dimensions of individuality, interactional asymmetry, and normativity

rather than as an all-or-nothing concept in which necessary and sufficient conditions either are – or are not – instantiated. On such an understanding agency is not necessarily lost when, for example, interactional asymmetry is temporarily reversed or absent. Furthermore, it can help explain how agency may be being instantiated even if it may not be clearly visible at both the level of the collective and the component at the same time as both the agency of the components, and the agency of the collective are spectra, and both will differ individually not only along the dimensions of individuality, interactional asymmetry, and normativity but also through time.

We highlighted that our own collective agency, based on our existence as multicellular organisms, coincides with collective subjectivity, that is, we have a lived perspective of concern, as do the individual organisms that make up our bodies (albeit in a much more minimal form than ourselves). But subjectivity (even when understood minimally as having a meaningful point of view on the world) coinciding with agency as it does in multicellular organisms such as ourselves, seems to be the exception rather than the rule. We argued that while it may be possible to genuinely share one's individual lived perspectives with other agents by forming a second- or multi-person perspective, to fully satisfy the additional operational conditions of first-person subjectivity at the level of the collective system as a whole, collective agents may have to be more than merely collective; they must be materially integrated into one living body.

References

Auvray, Malika, and Marieke Rohde. 2012. Perceptual crossing: The simplest online paradigm. *Frontiers in Human Neuroscience* 6(181). doi:10.3389/fnhum.2012.00181.

Auvray, Malika, Charles Lenay, and John Stewart. 2009. Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology* 27: 32–47. doi:10.1016/j.newideapsych.2007.12.002.

Barandiaran, Xabier, and Alvaro Moreno. 2008. Adaptivity: From metabolism to behavior. *Adaptive Behavior* 16: 325–344. doi:10.1177/1059712308093868.

Barandiaran, Xabier, Ezequiel Di Paolo, and Marieke Rohde. 2009. Defining agency: Individuality, normativity, asymmetry and spatio-temporality in action. *Adaptive Behaviour*17(5): 367–386.

Beer, R.D. 1996. Toward the evolution of dynamical neural networks for minimally cognitive behavior. In *From animals to animats 4: Proceedings of the fourth international conference on simulation of adaptive behavior*, ed. P. Maes, M. Mataric, J. Meyer, J. Pollack, and S. Wilson, 421–429. Cambridge, MA: MIT Press.

de Haan, Sanneke, Erik Rietveld, Martin Stokhof, and Damiaan Denys. 2013. The phenomenology of deep brain stimulation-induced changes in OCD: An enactive affordance-based model. *Frontiers in Human Neuroscience* 7(653). doi:10.3389/fnhum.2013.00653.

De Jaegher, Hanne, and Ezequiel Di Paolo. 2007. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences* 6: 485–507.

De Jaegher, Hanne, Ezequiel Di Paolo, and Shaun Gallagher. 2010. Can social interaction constitute social cognition? *Trends in Cognitive Sciences* 14: 441–447. doi:10.1016/j.tics.2010.06.009.

Dennett, Daniel C. 1991. Real patterns. *The Journal of Philosophy* 88: 27–51.

Di Paolo, Ezequiel. 2003. Organismically-inspired robotics: Homeostatic adaptation and natural teleology beyond the closed sensorimotor loop. In *Dynamical systems approach to embodiment and sociality*, ed. K. Murase and T. Asakura, 19–42. Adelaide: Advanced Knowledge International.

Di Paolo, Ezequiel. 2005. Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences* 4: 429–452. doi:10.1007/s11097-005-9002-y.

Di Paolo, Ezequiel. 2009a. Extended life. *Topoi* 28: 9–21. doi:10.1007/s11245-008-9042-3.

Di Paolo, Ezequiel A. 2009b. Chapter 3 Overcoming autopoiesis: An enactive detour on the way from life to society. In *Advanced series in management*, vol. 6, ed. Rodrigo Magalhães and Ron Sanchez, 43–68. Bingley: Emerald.

Di Paolo, Ezequiel A., and Iizuka Hiroyuki. 2008. How (not) to model autonomous behaviour. *Biosystems* 91: 409–423. doi:10.1016/j.biosystems.2007.05.016.

Di Paolo, Ezequiel, and Evan Thompson. 2014. The enactive approach. In *The routledge handbook of embodied cognition*, ed. Shapiro Lawrence. New York: Routledge Press.

Egbert, Matthew D., Xabier E. Barandiaran, and Ezequiel A. Di Paolo. 2012. Behavioral metabolution: The adaptive and evolutionary potential of metabolism-based chemotaxis. *Artificial Life* 18(1): 1–25.

Froese, Tom. 2010. From cybernetics to second-order cybernetics: A comparative analysis of their central ideas. *Constructivist Foundations* 5: 75–85.

Froese, Tom, and Ezequiel Di Paolo. 2008. Can evolutionary robotics generate simulation models of autopoiesis? In *Cognitive science research paper*, vol. 598. Brighton, University of Sussex.

Froese, Tom, and Ezequiel Di Paolo. 2011. The enactive approach: Theoretical sketches from cell to society. *Pragmatics and Cognition* 19: 1–36.

Froese, Tom, and Thomas Fuchs. 2012. The extended body: A case study in the neurophenomenology of social interaction. *Phenomenology and the Cognitive Sciences* 11: 205–235. doi:10.1007/s11097-012-9254-2.

Froese, Tom, and Tom Ziemke. 2009. Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence* 173: 466–500. doi:10.1016/j.artint.2008.12.001.

Froese, Tom, Nathaniel Virgo, and Eduardo Izquierdo. 2007. Autonomy: A review and a reappraisal. In F. Almeida e Costa, L. M. Rocha, E. Costa, I. Harvey & A. Coutinho (Eds.), *Advances in Artificial Life: 9th European Conference, ECAL 2007* (pp. 455–464).

Froese, Tom, Carlos Gershenson, and David A. Rosenblueth. 2013. The dynamically extended mind – A minimal modeling case study. In *2013 IEEE Congress on Evolutionary Computation* (pp. 1419–1426), IEEE Press.

Froese, Tom, Iizuka Hiroyuki, and Takashi Ikegami. 2014. Embodied social interaction constitutes social cognition in pairs of humans: A minimalist virtual reality experiment. *Scientific Reports* 4(3672). doi:10.1038/srep03672.

Gallagher, Shaun, and Dan Zahavi. 2008. *The phenomenological mind: An introduction to philosophy of mind and cognitive science*. New York: Routledge.

Gallotti, Mattia, and Chris D. Frith. 2013. Social cognition in the we-mode. *Trends in Cognitive Sciences* 17: 160–165. doi:10.1016/j.tics.2013.02.002.

Hutto, Daniel D., and Erik Myin. 2013. *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.

Kyselo, Miriam, and Wolfgang Tschacher. 2014. An enactive and dynamical systems theory account of dyadic relationships. *Frontiers in Psychology* 5(452). doi:10.3389/fpsyg.2014.00452.

Lysaker, Paul Henry, Jason K. Johannesen, and John Timothy Lysaker. 2005. Schizophrenia and the experience of intersubjectivity as threat. *Phenomenology and the Cognitive Sciences* 4: 335–352.

Maturana, Humberto R., and Francisco J. Varela. 1987. *The tree of knowledge: The biological roots of human understanding*. Boston: New Science Library/Shambhala Publications.

Morse, Anthony F., Carlos Herrera, Robert Clowes, Alberto Montebelli, and Tom Ziemke. 2011. The role of robotic modelling in cognitive science. *New Ideas in Psychology* 29(3): 312–324. doi:10.1016/j.newideapsych.2011.02.001.

Noë, Alva. 2004. *Action in perception*. Cambridge, MA: MIT Press.

Oullier, Olivier, and J.A. Scott Kelso. 2009. Social coordination from the perspective of coordination dynamics. In *Encyclopedia of complexity and systems sciences*, ed. Robert A. Meyers, 8198–8212. Berlin: Springer.

Protevi, John. 2009. *Political affect*. Minneapolis: University of Minnesota Press.
Reddy, Vasudevi. 2008. *How infants know minds*, Cambridge, MA: Harvard University Press.
Reynolds, Craig W. 1987. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on computer graphics and interactive techniques*, 25–34. New York: ACM. doi: 10.1145/37401.37406.
Riley, Michael A., Michael Richardson, Kevin Shockley, and Verónica C. Ramenzoni. 2011. Interpersonal synergies. *Movement Science and Sport Psychology* 2: 38. doi:10.3389/fpsyg.2011.00038.
Stanghellini, Giovanni, and Massimo Ballerini. 2004. Autism: Disembodied existence. *Philosophy, Psychiatry, & Psychology* 11: 259–268. doi:10.1353/ppp.2004.0069.
Stapleton, Mog, and Tom Froese. ms. The enactive philosophy of embodiment: From biological foundations of agency to the phenomenology of subjectivity. Manuscript submitted for publication.
Steiner, Pierre, and John Stewart. 2009. From autonomy to heteronomy (and back): The enaction of social life. *Phenomenology and the Cognitive Sciences* 8: 527–550. doi:10.1007/s11097-009-9139-1.
Sterelny, Kim, and Paul Griffiths. 1999. *Sex and death: An introduction to philosophy of biology*. Chicago: University of Chicago Press.
Thompson, Evan. 2007. *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.
Thompson, Evan. 2011. Reply to commentaries. *Journal of Consciousness Studies* 18: 5–6.
Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 1991. *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.
Weber, Andreas, and Francisco J. Varela. 2002. Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences* 1(2): 97–125.
Wheeler, Michael. 2011. Mind in life or life in mind? Making sense of deep continuity. *Journal of Consciousness Studies* 18: 148–168.