Artificial Intelligence 173 (2009) 466–500

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Enactive artificial intelligence: Investigating the systemic organization of life and mind

Tom Froese [a,*], Tom Ziemke [b]

[a] Centre for Computational Neuroscience & Robotics (CCNR), Centre for Research in Cognitive Science (COGS), University of Sussex, Brighton, UK
[b] Informatics Research Centre, University of Skövde, Skövde, Sweden

ARTICLE INFO

Article history:
Received 3 August 2007
Received in revised form 4 December 2008
Accepted 12 December 2008
Available online 25 December 2008

Keywords:
Embodied
Situated
Enactive
Cognitive science
Agency
Autonomy
Intentionality
Design principles
Natural cognition
Modeling

ABSTRACT

The embodied and situated approach to artificial intelligence (AI) has matured and become a viable alternative to traditional computationalist approaches with respect to the practical goal of building artificial agents, which can behave in a robust and flexible manner under changing real-world conditions. Nevertheless, some concerns have recently been raised with regard to the sufficiency of current embodied AI for advancing our scientific understanding of intentional agency. While from an engineering or computer science perspective this limitation might not be relevant, it is of course highly relevant for AI researchers striving to build accurate models of natural cognition. We argue that the biological foundations of enactive cognitive science can provide the conceptual tools that are needed to diagnose more clearly the shortcomings of current embodied AI. In particular, taking an enactive perspective points to the need for AI to take seriously the organismic roots of autonomous agency and sense-making. We identify two necessary systemic requirements, namely constitutive autonomy and adaptivity, which lead us to introduce two design principles of enactive AI. It is argued that the development of such enactive AI poses a significant challenge to current methodologies. However, it also provides a promising way of eventually overcoming the current limitations of embodied AI, especially in terms of providing fuller models of natural embodied cognition. Finally, some practical implications and examples of the two design principles of enactive AI are also discussed.

1. Introduction – setting the scene

The field of artificial intelligence (AI) has undergone some important developments in the last two decades, as also discussed by Anderson [1,2] and Chrisley [25] in recent papers in this journal. What started out with Brooks' emphasis of embodiment and situatedness in behavior-based AI and robotics in the late 1980s (e.g. [21]) has continued to be further developed (e.g. [22,5,100]) and has considerably influenced the emergence of a variety of successful AI research programs such as, for example, evolutionary robotics (e.g. [57,96]), epigenetic and developmental robotics (e.g. [15,79]), and the dynamical systems approach to adaptive behavior and minimal cognition (e.g. [11,13]).

In other words, the embodied approach to AI[1] has matured and managed to establish itself as a viable methodology for synthesizing and understanding cognition (e.g. [100,103]). Furthermore, embodied AI is now widely considered to avoid or successfully address many of the fundamental problems encountered by traditional "Good Old-Fashioned AI" [62], i.e.

* Corresponding author.
E-mail addresses: t.froese@gmail.com (T. Froese), tom.ziemke@his.se (T. Ziemke).

[1] In the rest of the paper we will use the term 'embodied AI', but intend it in a broad sense to include all of the abovementioned research programs.

0004-3702/$ – see front matter
doi:10.1016/j.artint.2008.12.001

'classical' problems such as those pointed out in Searle's [108] famous "Chinese Room Argument", the notorious "frame problem" (e.g. [84,33]), Harnard's [59] formulation of the "symbol grounding problem", or even the extensive Heideggerian criticisms developed by Dreyfus [42,43,45]. Although there are of course significant differences between these criticisms, what they all generally agree on is that purely computational systems, as traditionally conceived by these authors, cannot account for the property of intentional agency. And without this property there is no sense in saying that these systems know what they are doing; they do not have any *understanding* of their situation [63]. Thus, to put it slightly differently, all these arguments are variations on the problem of how it is possible to design an artificial system in such a manner that relevant features of the world actually show up as significant from the perspective of that system itself, rather than only in the perspective of the human designer or observer.

Given that embodied AI systems typically have robotic bodies and, to a large extent, appear to interact meaningfully with the world through their sensors and motors, one might think that the above problems have either disappeared or at least become solvable. Indeed, it has been argued that some dynamical form of such embodied AI is all we need to explain how it is that systems can behave in ways that are adaptively sensitive to context-dependent relevance [139]. Nevertheless, there have been some warning signs that something crucial might still be amiss. In fact, for the researcher interested in the philosophy of AI and the above criticisms, this should not come as a surprise. While Harnard's [58] position is that of a robotic functionalism, and thus for him the robotic embodiment is a crucial part of the solution to the symbol grounding problem, this is not the case for Searle. Already Searle's [108] original formulation of the Chinese Room Argument was accompanied by what he called the "robot reply" – envisioning essentially what we call embodied AI today, i.e. computer programs controlling robots and thus interacting with the real world – but rejected that reply as not making any substantial difference to his argument. Let us shift attention though, from these 'classic' philosophical arguments to a quick overview of more recent discussions among practitioners of embodied AI, which will be elaborated in more detail in the following sections.

Already a decade ago Brooks [22] made the remark that, in spite of all the progress that the field of embodied AI has made since its inception in the late 1980s, it is certainly the case that actual biological systems behave in a considerably more robust, flexible, and generally more life-like manner than any artificial system produced so far. On the basis of this 'failure' of embodied AI to properly imitate even insect-level intelligence, he suggests that perhaps we have all missed some general truth about living systems. Moreover, even though some progress has certainly been made since Brooks' rather skeptical appraisal, the general worry that some crucial feature is still lacking in our models of living systems nevertheless remains (e.g. [23]).

This general worry about the inadequacy of current embodied AI for advancing our scientific understanding of natural cognition has been expressed in a variety of ways in the recent literature. Di Paolo [36], for example, has argued that, even though today's embodied robots are in many respects a significant improvement over traditional approaches, an analysis of the organismic mode of being reveals that "something fundamental is still missing" to solve the problem of meaning in AI. Similarly, one of us [143] has raised the question whether robots really are embodied in the first place, and has elsewhere argued [141] that embodied approaches have provided AI with physical grounding (e.g. [20]), but nevertheless have not managed to fully resolve the grounding problem. Furthermore, Moreno and Etxeberria [90] provide biological considerations which make them skeptical as to whether existing methodologies are sufficient for creating artificial systems with natural agency. Indeed, concerns have even been raised, by ourselves and others, about whether current embodied AI systems can be properly characterized as autonomous in the sense that living beings are (e.g. [110,146,147,52,61]). Finally, Heideggerian philosopher Dreyfus, whose early criticisms of AI (cf. above) have had a significant impact on the development of modern embodied AI (or "Heideggerian AI", as he calls it), has recently referred to these new approaches as a "failure" [46]. For example, he claims that embodied/Heideggerian AI still falls short of satisfactorily addressing the grounding problem because it cannot fully account for the constitution of a meaningful perspective for an agent.

Part of the problem, we believe, is that while the embodied approach has mostly focused on establishing itself as a viable alternative to the traditional computationalist paradigm [2], relatively little effort has been made to make connections to theories outside the field of AI, such as theoretical biology or phenomenological philosophy, in order to address issues of natural autonomy and embodiment of living systems [144]. However, as the above brief overview of recent discussions indicates, it appears that awareness is slowly growing in the field of embodied AI that something essential might still be lacking in current models in order to fulfill its own ambitions to avoid, solve or overcome the problems traditionally associated with computationalist AI,[2] and thereby provide better models of natural cognition.

We argue that it looks promising that an answer to the current problems might be gained by drawing some inspiration from recent developments in enactive cognitive science (e.g. [116–118,120,121,113,95]). The enactive paradigm originally emerged as a part of embodied cognitive science in the early 1990s with the publication of the book *The Embodied Mind* [127] that has strongly influenced a large number of embodied cognition theorists (e.g. [26]). More recent work in enactive cognitive science has more explicitly placed biological autonomy and lived subjectivity at the heart of enactive cognitive science (cf. [118,41]). Of particular interest in the current context is its incorporation of the organismic roots of autonomous agency and sense-making into its theoretical framework (e.g. [136,38]).

[2] We will use the term 'computationalist AI' to broadly denote any kind of AI which subscribes to the main tenets of the Representationalist or Computational Theory of Mind (cf. [60]), especially the metaphors 'Cognition Is Computation' and 'Perception Is Representation' (e.g. mostly GOFAI and symbolic AI, but also much sub-symbolic AI and some embodied approaches).

While the notion of 'enactive AI' has already been around since the inception of enactive cognitive science[3] (e.g. [127, 47,141]), how exactly the field of embodied AI relates to this further shift in the cognitive sciences is still in need of clarification [49]. Thus, while these recent theoretical developments might be of help with respect to the perceived limitations of the methodologies employed by current embodied AI, there is still a need to specify more precisely what actually constitutes such a fully enactive AI. The aim of this paper is to provide some initial steps toward the development of such an understanding.

The rest of the paper is structured as follows: Firstly, in Section 2 the embodied approach to AI is characterized and analyzed by means of a set of design principles developed by Pfeifer and colleagues (e.g. [103,101,100]), and some apparent problems of the embodied approach are discussed. Secondly, in Section 3, the biological foundations of the enactive approach to autonomous agency and sense-making are presented as a promising theoretical framework for enabling embodied AI practitioners to better understand and potentially address some of the perceived limitations of their approach. In particular, the historical roots of the enactive approach are briefly reviewed, and recent progress in the autopoietic tradition is discussed. In Section 4, some design principles for the development of fully enactive AI are derived from the theoretical framework outlined in Section 3, and some promising lines of recent experimental work which point in this direction are discussed. Section 5 then summarizes the arguments and presents some conclusions.

2. Embodied AI and beyond

The aim of this section is mainly twofold: (i) to briefly review some of the guiding design principles of the embodied approach to AI as developed by Pfeifer and others (e.g. [99–103]), and (ii) to discuss some of the concerns which have recently been raised regarding the limitations of the embodied AI approach.[4] This second part will unfold in three stages: (i) a critical analysis of Dreyfus' [46] arguments for the "failure" of embodied AI, (ii) a review of the reasons for holding that a closed sensorimotor loop is necessary but not sufficient to solve the problem of meaning in AI, and (iii) a defense of the claim that the grounding of meaning also requires autonomous agency, a property which cannot be derived from sensorimotor capacities alone. These considerations will set the stage for a brief introduction to the theoretical framework of enactive cognitive science.

2.1. Foundations of embodied AI

What is embodied AI? One helpful way to address this question is by means of a kind of field guide such as the one recently published in this journal by Anderson [1]. Another useful approach is to review the main design principles which are employed by the practitioners of embodied AI in order to engineer their autonomous robots. The latter is the approach adopted here because it will provide us with the background from which to propose some additional principles for the development of enactive AI later on in this paper (Section 4).

Fortunately, there has already been some effort within the field of embodied AI to make their design principles more explicit (e.g. [99]; see [103,101,100] for a more elaborate discussion). Here we will briefly recapitulate a recent overview of these principles by Pfeifer, Iida and Bongard [101]. It is worth emphasizing that Pfeifer's attempt at their explication has its beginnings in the early 1990s, and, more importantly, that they have been derived from over two decades of practical AI research since the 1980s [99]. The design principles are summarized in Table 1.

The design principles are divided into two subcategories, namely (i) the "design procedure principles", which are concerned with the general philosophy of the approach, and (ii) the "agent design principles", which deal more directly with the actual methodology of designing autonomous agents [102].

The first of the design procedure principles (P-1) makes it explicit that the use of the synthetic methodology by embodied AI should be primarily viewed as a scientific rather than as an engineering endeavor, while, of course, these two goals do not mutually exclude each other [100,56]. It is therefore important to realize that we are mostly concerned with the explanatory power that is afforded by the various AI approaches reviewed in this paper. In other words, the main question we want to address is how we should build AI systems such that they can help us to better understand natural phenomena of life and mind. Of course, since living beings have many properties that are also desirable for artificial systems and which are still lacking in current implementations [12], any advances in this respect are also of importance in terms of more practical considerations such as how to design more robust and flexible AI systems. Of course, it is certainly the case that the "understanding by building" principle has also been adopted by many practitioners within the traditional paradigm since the inception of AI in the 1950s, though it can be said that today's computationalist AI is generally more focused

[3] Varela, Thompson and Rosch [127] in fact referred to Brooks' work on subsumption architectures and behavior-based robotics (e.g. [20,21]) as an "example of what we are calling enactive AI" (p. 212) and a "fully enactive approach to AI" (p. 212). Nowadays, however, many researchers would probably not refer to this work as "fully enactive", due to the lack of constitutive autonomy, adaptivity and other reasons discussed in this paper.

[4] It might be worth noting that Pfeifer's principles here serve as representative for the principles and the state of the art of the embodied AI approach as formulated by one of the leading researchers (and his co-workers). Hence, the extensions required for enactive AI formulated in this paper should not be interpreted as criticisms of Pfeifer's principles (or other work) specifically, but rather as further developments of the general embodied approach to AI that they are taken to be representative for.

Table 1
Summary of the embodied AI design principles of autonomous agents (adapted from [101]). The first five (P-X) are "design procedure principles" and the remaining ones (A-X) are "agent design principles". See [100, pp. 357–358] for a recent summary of how these basic principles can be extended to include insights specifically related to the design of developmental systems, artificial evolution, and collective systems.

#	*Name*	*Description*
P-1	Synthetic methodology	Understanding by building
P-2	Emergence	Systems designed for emergence are more adaptive
P-3	Diversity-compliance	Trade-off between exploiting the givens and generating diversity solved in interesting ways
P-4	Time perspectives	Three perspectives required: 'here and now', ontogenetic, phylogenetic
P-5	Frame of reference	Three aspects must be distinguished: perspective, behavior vs. mechanisms, complexity
A-1	Three constituents	Ecological niche (environment), tasks, and agent must always be taken into account
A-2	Complete agent	Embodied, autonomous, self-sufficient, situated agents are of interest
A-3	Parallel, loosely coupled processes	Parallel, asynchronous, partly autonomous processes; largely coupled through interaction with the environment
A-4	Sensorimotor coordination	Sensorimotor behavior coordinated with respect to target; self-generated sensory stimulation
A-5	Cheap design	Exploitation of niche and interaction; parsimony
A-6	Redundancy	Partial overlap of functionality based on different physical processes
A-7	Ecological balance	Balance in complexity of sensory, motor, and neural systems: task distribution between morphology, materials, and control
A-8	Value	Driving forces; developmental mechanisms; self-organization

on engineering systems that do useful work (e.g. smart devices, military applications, search engines, etc.) rather than on systems that do scientific or philosophical explanatory work.

The emergence principle (P-2) is also shared by many computationalist AI systems, at least in the minimal sense that behavior always emerges out of the interactions of an agent with its environment. Nevertheless, as Pfeifer and Gómez [102] point out, emergence is a matter of degree and it is increased the further a designer's influence is removed from the actual behavior of the system. A combined dynamical and evolutionary robotics approach is a popular choice for embodied AI in this regard (e.g. [13,57,96]). Design procedure principle P-3 emphasizes awareness of the fact that there often is a trade-off between robustness and flexibility of behavior, a trade-off which can be encountered in a variety of domains. In Section 4 we will discuss some recent work which tries to address this problem in a novel manner.

P-4 highlights the important fact that organisms are temporally embedded in three timescales, namely (i) state-oriented (the immediate present), (ii) learning and developmental (ontogeny), and (iii) evolutionary change (phylogeny). Any complete explanation of an organism's behavior therefore must incorporate these three perspectives. The final design procedure principle (P-5) raises awareness of the different frames of reference which are involved in building and understanding autonomous systems. At least three points are worth emphasizing: (i) the need to distinguish between the external perspective of the observer or designer and the frame of reference of the system, (ii) as already stressed by P-2, behavior is a relational phenomenon which cannot be reduced either to the agent or its environment, and (iii) any behavior that appears quite clever to an external observer does not necessarily entail the existence of a similarly intelligent underlying mechanism. This last point especially sets apart embodied AI from the explicit modeling approach adopted by many proponents of computationalist AI and links it back to some research in connectionism as well as earlier work in the cybernetics tradition, such as that of Ashby [6,7].

The first of the agent design principles (A-1) underlines the important point that an autonomous system should never be designed in isolation. In particular, we need to consider three interrelated components of the overall system: (i) the target niche or environment, (ii) the target task and desired behavior, and (iii) the agent itself. Much can be gained from exploiting an agent's context during the engineering of appropriate task solving behavior. This already follows from the non-reducibility of an agent's behavior to internal mechanisms (P-5), but is further supported by the importance of embodiment and situatedness for real-world cognition (e.g. [21]). As a complement to design principle A-1, and in contrast to much work in traditional AI, principle A-2 holds that in order to better understand intelligence we need to study complete agents rather than sub-agential components alone. Of course, this is not to deny that designing isolated components can often be extremely useful for practical applications, but if we want to gain a better scientific understanding of intelligence then we need to investigate how adaptive behavior emerges out of the dynamics of brain-body-world systemic whole (e.g. [13,11]). As will become evident in the following sections, one of the central issues of this paper is to analyze exactly what defines a 'complete agent'.

Principle A-3 emphasizes that, in contrast to many computationalist AI systems, natural intelligent behavior is not the result of algorithmic processes being integrated by some sort of central controller. In terms of embodied AI, cognition is based on a large number of parallel, loosely coupled processes that run asynchronously, and which are coupled to the internal organization of an agent's sensorimotor loop. Indeed, design principle A-4 represents the claim that most cognition is best conceived of as appropriate sensorimotor coordination. The advantage of this kind of situatedness is that an agent is able to structure its own sensory input by effectively interacting with its environment. The problem of perceptual categorization, for example, is thus greatly simplified by making use of the real world in a non-computational manner.

We will quickly run through agent design principles A-5 to A-7 because they are mainly targeted at engineering challenges of designing physical robotic systems. The design principle of cheap robot design (A-5) also emphasizes the importance of taking an agent's context into account (cf. A-1), since it is possible to exploit the physics and constraints

of the target niche in order to build autonomous systems. The redundancy principle (A-6) holds that functionality of any subsystems should overlap to some extent in order to guarantee greater robustness. The principle of ecological balance (A-7) emphasizes two points, namely (i) that there should be a match in complexity of the sensory, motor and control systems, and (ii) that control will be easier if an agent's morphology and materials are appropriately selected with the target task in mind.

Finally, there is the value principle (A-8) which refers to designing the motivations of an agent. In particular, it is concerned with the implementation of 'online' learning in the sense of providing an agent with feedback with regard to its actions. As we will see later, this is one of the design principles which, at least in its current formulation, appears as most questionable from the point of view of enactive AI (cf. [41]).

As Pfeifer and Gómez [102] point out, this list of design principles is by no means complete and could, for example, be extended by a corresponding set of principles for designing evolutionary systems. Indeed, Pfeifer and Bongard [100, pp. 357–358] provide exactly such an extended list that include design principles for development, evolution and collective systems. Nevertheless, this overview should be sufficient for our current purpose of briefly outlining the basic ideas behind embodied AI, and for evaluating how they possibly differ from those of fully enactive AI. We will introduce some specifically enactive design principles later on in this paper (Section 4).

2.2. The "failure" of embodied AI?

Now that we have reviewed some of the main principles of embodied AI, we can ask: what is the current status of the field? As we have already mentioned in the introduction, this new approach to AI has been in many respects a great success. Indeed, the insights gained in this field have significantly contributed to the embodied turn in the cognitive sciences (e.g. [26]). It will thus come as a surprise to many that Dreyfus, a philosopher whose Heideggerian critique of computationalist AI has been an inspiration to many practitioners of embodied AI, has recently referred to current work in this field as a "failure" [46]. Moreover, he has argued that overcoming these difficulties would in fact require such "Heideggerian AI" to become even more Heideggerian. What does he mean by this? Analyzing the source of his concern will provide us with the starting point to motivate the development of the kind of enactive AI which we will advocate in this paper.

On the one hand, Dreyfus has a particular target in mind, namely the embodied AI philosopher Wheeler who has recently published a book on this topic in which he is unwilling to relinquish representationalism completely (cf. [138]).[5] However, this part of Dreyfus' argument is not that surprising since the rejection of symbolic representations was already at the core of his extensive critique of GOFAI (e.g. [45]), and similar concerns about representations are also shared by many embodied AI practitioners (e.g. [55,56]). However, on the other hand Dreyfus also takes issue with the field of embodied AI more generally. For him the "big remaining problem" is how to incorporate into current embodied AI an account of how we "directly pick up significance and improve our sensitivity to relevance" since this ability "depends on our responding to what is significant for *us*" given the current contextual background Dreyfus [46]. Thus, in spite of all the important contributions made by embodied AI, Dreyfus claims that the field still has not managed to properly address the problem of meaning in AI. Moreover, as long as there is no meaningful perspective from the point of view of the artificial agent, which would allow it to appropriately pick up relevance according to its situation in an autonomous manner, such a system cannot escape the notorious 'frame problem' as it is described by Dennett [33].

Why has the field's shift toward embodied artificial agents which are embedded in sensorimotor loops not been sufficient to account for a meaningful perspective as it is enjoyed by us and other living beings? The trouble for Dreyfus [46] is that if this significance is to be replicated artificially we seem to need "a model of our particular way of being embedded and embodied such that what we experience is significant for us in the particular way that it is. That is, we would have to include in our program a model of a body very much like ours". Furthermore, if we cannot design our models to be responsive to environmental significance in this manner then "the project of developing an embedded and embodied Heideggerian AI can't get off the ground". Accordingly, Dreyfus draws the skeptical conclusion that, even if we tried, since the appropriate "computer model would still have to be given a detailed description of our body and motivations like ours if things were to count as significant for it so that it could learn to act intelligently in our world", it follows that such models "haven't a chance of being realized in the real world".

What are we to make of these skeptical assessments? One possible initial response to Dreyfus would be to point out that most of current embodied AI does not actually aim to model human-level understanding. As such it does not require a full description of our human body in order to solve the grounding problem. However, this response is insufficient since the problem of the apparent lack of significance for embodied AI systems nevertheless remains; providing a sufficiently detailed model of a living body is still impossible even for the simplest of organisms. Hence, if such a complete model of a body is actually necessary to make progress on the issue of significance and grounded meaning, then we are forced to admit that Dreyfus is right to be unconvinced that embodied AI might do the trick.

However, it could also be argued that such a detailed modeling approach is not even desirable in the first place since it does not help us to understand why having a particular body allows things in the environment to show up as significant

[5] For another recent critique of Wheeler's (as well as Clark's and Rowland's) attempt to make space for the notion of representation within embodied cognitive science, see [53]. For Wheeler's response to the criticisms by Dreyfus [46], see [139].

for the agent possessing that body (cf. Searle's [108] robot reply). In other words, instead of blindly modeling the bodies of living beings in as much detail and complexity as possible, it would certainly be preferable to determine the *necessary conditions* for the constitution of an individual agent with a meaningful perspective on the world. Thus, another response to Dreyfus is to point out that the purpose of a model is not to replicate or instantiate a particular phenomenon, but to help explain it [39,91].

Accordingly, we can accept Dreyfus' rejection of the feasibility of detailed models of our human bodies, but nevertheless disagree with his conclusion that this necessarily implies a dead end for the project of embodied AI. Instead, we propose that what is needed is an understanding of the biological body which will enable us to design an embodied artificial agent that is at the same time (i) simple enough for us to actually construct and analyze, and (ii) still fulfills the required conditions for the constitution of a meaningful perspective for that agent. As we will see in Section 3, Dreyfus' own suggestions for these required conditions, namely some of the special systemic properties of the human nervous system such as self-organization and circular causality (e.g. [48]), are also constitutive of the overall biological organization of even the simplest organisms. Note that by conceptualizing Dreyfus' argument in this way we have transformed the seemingly insurmountable problem of meaning in embodied AI into something potentially more manageable, and thus made it conceivable that we can address its notorious symptoms, such as the problem of grounding meaning, in a systematic manner from the bottom-up.

In order to get an idea of the conditions which must be addressed by AI for an appropriate form of organismic embodiment, it is helpful to first diagnose more clearly the limitations which are currently faced by the methods of embodied AI. Accordingly, in the next subsection we will make a first pass at indicating why sensorimotor embodiment is a necessary but not a sufficient condition for the constitution of a meaningful perspective (Section 2.3), and then further deepen the discussion of this insufficiency by considering the topic of biological or natural agency (Section 2.4). This will then finally provide us with the theoretical background from which to motivate the development of an enactive approach to AI (Section 2.5).

2.3. The problem of meaning in embodied AI

Of the various difficulties which computationalist AI has to face whenever it attempts to extend its target domain beyond simplified 'toy worlds' in order to address context-sensitive real-world problems in a robust, timely and flexible manner, the frame problem is arguably the most widely discussed. From its beginnings as a formal AI problem [84] it has developed into a general philosophical concern with how it is possible for rational agents to deal with the complexity of the real world, as epitomized by a practically inexhaustible context, in a meaningful way (e.g. [33,46,139]).

In response to this problem, most embodied AI practitioners have accepted Dreyfus' [45] argument that this problem is largely derived from computationalist AI's focus on abstract reasoning and its reliance on internal representations for explicit world modeling. Thus, a practical solution to the frame problem is to embody and situate the artificial agents such that they can use the 'world as its own best model' [21].[6] This is usually accomplished by designing appropriate *closed sensorimotor loops*, an approach which emphasizes the fact that the effects of an agent's actuators, via the external environment, impact on the agent's sensors and, via the internal controller, again impact the actuators [27]. Moreover, the difficulty of designing and fine-tuning an agent's internal dynamics of these sensorimotor loops is nowadays often relegated to evolutionary algorithms, thereby making it unnecessary for the engineer to explicitly establish which correlations are relevant to the agent's situation (e.g. [13]). This methodological shift toward situatedness, dynamical systems and artificial evolution has significantly contributed to the establishment of the field of embodied AI and still continues to be the generative method of choice for many practitioners (cf. [57]).

The focus on the organization of sensorimotor situatedness has several important advantages. The crucial point is that it enables an artificial agent to dynamically structure its own sensory inputs through its ongoing interaction with the environment [103, p. 377]. Such a situated agent does not encounter the frame problem, more generally conceived, because of its tight sensorimotor coupling with the world. It never has to refer to an internal representation of the world that would always quickly get out of date as its current situation and the world around it continually changed. Furthermore, it has been claimed that for such an agent "the symbol grounding problem is really not an issue – anything the agent does will be grounded in its sensory-motor coordination" [99]. In other words, from the perspective of embodied AI it seems that the problem of grounding meaning has been practically resolved by generating artificial agents that are embedded in their environment through their sensorimotor capabilities [27].

Accordingly, it seems fair to say that embodied AI has made progress on the classical problems associated with computationalist AI, and that it has developed methods which can generate better models of natural cognition. However, how has this change in methodology actually resolved the traditional problem of grounding meaning in AI? Can we claim that an artificial agent's embeddedness in a sensorimotor loop is sufficient for grounding a meaningful perspective *for* that agent? This would be a considerable trivialization, particularly in light of the complexity involved in the constitution of such situatedness in biological agents [90]. Following Di Paolo [36], one of the arguments of this paper is that the existence of a closed sensorimotor feedback loop is a *necessary* but *not sufficient* condition for the attribution of an intrinsically meaningful

[6] An outstanding issue with this approach, which will not be discussed further here, is that so far there have been no convincing demonstrations that this methodology can be successfully scaled to also solve those 'higher-level' cognitive tasks which have been the focus of more traditional AI (see, for example, [75]). Here we are only concerned whether it can resolve the problem of meaning in AI as such.

perspective for the agent such that it can be said to engage in purposeful behavior. It follows that, generally speaking, the 'world' of an embodied AI system is "quite devoid of significance in that there is no sense other than the figurative in which we can say that a robot *cares* about what it is doing" [36]. Or, in the words of Nagel [93], we could say that there is nothing 'what it is like to be' such a system. While this assessment is probably shared by the majority of embodied AI practitioners, the widespread use of this kind of figurative language often invites confusion.

As a point in case, consider Franklin's [47, p. 233] use of teleological terms when he invites us to "think of an autonomous agent as a creature that senses its environment and acts on it so as to further its own agenda", and then continues by claiming that "any such agent, be it a human or a thermostat, has a single, overriding concern – what to do next". Thanks to the evidence of our own lived experience we can confirm that we are indeed autonomous agents in this sense, and it could be argued that the continuity provided by an evolutionary perspective enables us to attribute a similar concerned experience to other living beings [136]. But do we really want to commit ourselves to attributing such a perspective of concern to a thermostat?

As will be argued in this section and the next, there are strong philosophical arguments against holding such a position. In particular, it is important to emphasize that the existence of what could be described by an external observer as 'goal-directed' behavior does not necessarily entail that the system under study itself has those goals, that is, they could be *extrinsic* (i.e. externally imposed) rather than *intrinsic* (i.e. internally generated) [71, pp. 108–134]. In the case of the thermostat its 'agenda' is clearly externally imposed by the human designer and, in spite of being embedded in a negative feedback loop, it is therefore reasonable to assume that any talk of its 'concern about what to do next' should be judged as purely metaphorical. It follows that for any system of this sensorimotor type we can say that its 'goals' are not its own (cf. [61]).

It is also worth emphasizing in this context that adding extra inputs to the dynamical controllers of embodied AI systems and labeling them "motivational units" (e.g. [98]), does not entail that these are actually motivations *for* the robotic system itself. Thus, in contrast to the indications of the value design principle (A-8) for embodied AI, we agree with Di Paolo, Rohde and De Jaegher [41] that the problem of meaning cannot be resolved by the addition of an explicitly designed 'value' system, even when it can generate a signal to modulate the behavior of the artificial agent (e.g. [101]). It might seem that such a functional approach avoids Dreyfus' earlier Heideggerian critique of symbolic AI, which was based on the claim that "facts and rules are, by themselves, meaningless. To capture what Heidegger calls significance or involvement, they must be *assigned relevance*. But the predicates that must be added to define relevance are just more meaningless facts" [44, p. 118]. However, even though most embodied AI does not implement 'value' systems in terms of such facts and rules, it does not escape Heidegger's general criticism:

> The context of assignments or references, which, as significance, is constitutive for worldliness, can be taken formally in the sense of a system of relations. [But the] phenomenal content of those 'relations' and 'relata' – the 'in-order-to', the 'for-the-sake-of', and the 'with which' of an involvement – is such that they resist any sort of mathematical functionalization.
> [64, pp. 121–122]

To illustrate this point we can consider Parisi's [98] example of a robot which is provided with two inputs that are supposed to encode its motivational state in terms of hunger and thirst. While it is clear that these inputs play a *functional* role in generating the overall behavior of the robot, any description of this behavior as resulting from, for example, the robot's *desire* to drink in order to avoid being thirsty must be deemed as purely metaphorical at best and misleading at worst. From the perspective of Heidegger's critique, there is no essential difference between encoding significance in terms of explicit facts and rules or as input functions; both are forms of representation whose meaning is only attributed by an external observer. Thus, while embodied AI has been able to demonstrate that it is indeed possible to at least partly model the *function* of significance as a system of relations in this manner, it has not succeeded in designing AI systems with an intrinsic perspective from which those relations are actually encountered as significant. The shift of focus toward sensorimotor loops was an important step in the right direction since it resulted in more robust and flexible systems, but it nevertheless did not fully solve the problem of meaning in AI [36,141,147].

Instead, the problem has reemerged in the form of how to give the artificial system a perspective on what is relevant to its current situation such that it can structure its sensorimotor relationship with its environment appropriately. The essential practical implication for the field of AI is that instead of attempting the impossible task of explicitly capturing relations of significance in our models, we need to design systems which manage to satisfy the appropriate necessary conditions such that these relations are able to emerge spontaneously for that system. We are thus faced with the problem of determining what kind of embodiment is necessary so that we can reasonably say that there is such a concern *for* the artificial agent just like there is for us and other living beings. What kind of body is required so that we can say that the agent's goals are genuinely its own? This question cannot be answered by current embodied AI. Accordingly, the shift from computationalist AI to embodied AI seems to have coincided with a shift from the symbol grounding problem to a "body grounding problem" [141,147].

2.4. The problem of agency in embodied AI

We have argued that the addition of an explicit 'value' system to an artificial agent only amounts to an increase in the complexity of the transfer function between its sensors and motors. Accordingly, we do not consider such an agent to be essentially different from other embodied AI systems with simpler sensorimotor loops for the purpose of this discussion. We have also indicated that we do not consider such sensorimotor systems as a sufficient basis for us to be able to speak of the constitution of an agent's own meaningful perspective on the world. Since we nevertheless accept that humans and other living beings do enjoy such a perspective, we need to consider more carefully what exactly it is that is lacking in these artificial systems.

In order to answer this question it is helpful to first consider it in the context of recent developments in the cognitive sciences, in particular in relation to the sensorimotor approach to perception (e.g. [97,95]). The idea behind this approach can be summarized by the slogan that 'perceiving is a way of acting'; or more precisely, "*what we perceive* is determined by *what we do* (or what we know how to do)" [95, p. 1]. In other words, it is claimed that perception is a skillful mode of exploration of the environment which draws on an implicit understanding of sensorimotor regularities, that is, perception is constituted by a kind of bodily know-how [97]. In general, the sensorimotor account emphasizes the importance of *action* in perception. The capacity for action is not only needed in order to make use of sensorimotor skills, it is also a necessary condition for the acquisition of such skills since "only through *self*-movement can one *test* and so *learn* the relevant patterns of sensorimotor dependence" [95, p. 13]. Accordingly, for perception to be constituted it is not sufficient for a system to simply undergo an interaction with its environment, since the exercise of a skill requires an intention and an agent (not necessarily a 'homunculus') that does the intending. In other words, the dynamic sensorimotor approach needs a notion of selfhood or *agency* which is the locus of intentional action in the world [117].

However, can sensorimotor loops by themselves provide the conceptual means of distinguishing between the intentional action of an autonomous agent and mere accidental movement? If such sensorimotor loops alone are not sufficient to account for the existence of an intentional agent, then we have identified a serious limitation of the methodologies employed by the vast majority of embodied AI. Thus, in order to determine whether this is indeed the case, it is useful to ask whether a system that only consists of a simple negative feedback loop could perhaps be conceived of as an intentional agent in its own right. This removes any misleading terminology and complexity from the problem, and the correspondence between the feedback loop in a closed-loop controller and the sensorimotor feedback provided by the environment is also explicitly acknowledged in embodied AI (e.g. [27]).

The grounding of discourse about teleological behavior in terms of negative feedback loops is part of a long tradition which can be traced back at least as far as the publication of a seminal paper by Rosenblueth, Wiener and Bigelow [105] on the topic during the early cybernetics era. And, indeed, it is thanks to this tradition that we can now recognize that some form of feedback is a necessary condition for the constitution of purposeful behavior [36]. However, whether such a basic sensorimotor feedback account of behavior can supply the crucial distinction between *intrinsic* and *extrinsic* teleology, i.e. whether the behavior is meaningful *for* the system or is only attributed *to* the system metaphorically, is not clear.

As Jonas [71, p. 117] notes in his critique of cybernetics, this depends on "whether effector and receptor equipment – that is motility and perception alone – is sufficient to make up motivated animal behavior". Thus, referring to the simple example of a target-seeking torpedo, Jonas [71, p. 118] rephrases the question of agency as "whether the mechanism *is* a 'whole', having an identity or selfness that can be said to be the bearer of purpose, the subject of action, and the maker of decisions". Similar to Dreyfus' [46] assessment of embodied AI, Jonas concludes that if the system in question essentially consists of the two elements (motility and perception) which are somehow coupled together, as they are in artificial sensorimotor loops, it follows that "sentience and motility alone are not enough for purposive action" [71, p. 120]. Why? Because in order for them to constitute intrinsically purposive action there must be interposed between them a center of "concern". What is meant by this?

We have already seen that the inclusion of a 'value' system in the internal link of the sensorimotor loop is not sufficient for this task because it does not escape the Heideggerian critique. What, then, is the essential difference between a target-seeking torpedo and a living organism when the activity of both systems can be described as 'goal-directed' in terms of sensorimotor feedback loops? Jonas observes:

> A feedback mechanism may be going, or may be at rest: in either state the machine exists. The organism has to keep going, because to be going is its very existence – which is revocable – and, threatened with extinction, it is concerned in existing.
> [71, p. 126]

In other words, an artificial system consisting only of sensors and motors that are coupled together in some manner, and which for reasons of design does not have to continually bring forth its own existence under precarious conditions, cannot be said to be an individual subject in its own right in the same way that a living organism can. Accordingly, we might describe the essential difference between an artificial and a living system in terms of their mode of being: whereas the former exists in a mode that could be described as *being by being*, namely a kind of being which can give rise to forms of doing but not necessarily so for its being, the latter *only* exists in a mode that can be defined as *being by doing*. A living system not only *can* actively engage in behavior, it necessarily *must* engage in certain self-constituting operations in order

to even exist at all. The significance of this ontological distinction, i.e. a difference in being (cf. Heidegger's [64] notion of Being or "Sein"), will be further unpacked in the following sections.

For now it is important to realize that, even though Jonas' existential critique might sound too philosophical to be of any practical use, it actually brings us right back to our original problem of what is currently lacking in the field of embodied AI. Similar to the negative feedback mechanisms investigated by Rosenblueth, Wiener and Bigelow [105], according to this distinction embodied AI systems "cannot be rightly seen as centers of concern, or put simply as *subjects*, the way that animals can. [...] Such robots can never by truly autonomous. In other words the presence of a closed sensorimotor loop *does not* fully solve the problem of meaning in AI" [36]. However, at this point we are finally in a position to ask the right kind of question which could potentially resolve this fundamental problem: through which mechanism are living organisms able to enjoy their peculiar mode of existence?

Jonas puts his finger on metabolism as the source of all intrinsic value and proposes that "to an entity that carries on its existence by way of constant regenerative activity we impute *concern*. The minimum concern is to be, i.e. to carry on being" [72]. Conversely, this leads to the rather trivial conclusion that the current AI systems, which just carry on being no matter whether they are doing anything or not, have nothing to be concerned about. In contrast, metabolic systems must continually reassert their existence from moment to moment in an ongoing effort of self-generation that is never guaranteed to succeed.[7]

This grounding of intrinsic meaning in the precarious mode of metabolic existence, namely in the form of 'being by doing', might be a rather disappointing conclusion for the field of embodied AI. While it avoids Dreyfus' [46] claim that a "detailed description" of our human bodies is necessary to avoid the "failure" of this field, it is still rather impractical – if not impossible – to design artificial agents that are fully metabolizing. Nevertheless, leaving the question of whether metabolism is the only way to realize this particular mode of existence aside for now, it is interesting to note that from this perspective it appears that the problem of meaning and intentionality has been fundamentally misunderstood in computationalist AI: it is not a problem of *knowledge* but rather of *being*. In this respect the development of embodied AI has already made an important contribution toward a potential resolution of this problem, namely by demonstrating that it is an essential aspect of living being to be tightly embedded in a world through ongoing sensorimotor interaction.

Nevertheless, in order to make further progress in this direction we need a theoretical framework that enables us to gain a better understanding of the essential features of that peculiar mode of existence which we call living being.

2.5. From embodied AI to enactive AI

The preceding considerations have given support to Dreyfus' [46] claim that there is something crucial lacking in current embodied AI such that we cannot attribute a human-like perspective to these systems. Moreover, by drawing on the arguments of Jonas [72] this lack has been generalized as a fundamental distinction between (current) artificial and living systems in terms of their mode of existence. This distinction has provided us with the initial step toward a new theoretical framework from which it will be possible to respond to Brooks' challenge:

> [P]erhaps we have all missed some organizing principle of biological systems, or some general truth about them. Perhaps there is a way of looking at biological systems which will illuminate an inherent necessity in some aspect of the interactions of their parts that is completely missing from our artificial systems. [...] I am suggesting that perhaps at this point we simply do not *get it*, and that there is some fundamental change necessary in our thinking.
> ([22]; cf. [23])

One of the reasons for this problematic situation is that current work in embodied AI does not in itself constitute an internally unified theoretical framework with clearly posed problems and methodologies (cf. [99], [100, p. 62]). Since the field still lacks a firm foundation it can perhaps be better characterized as an amalgamation of several research approaches in AI which are externally united in their opposition to the orthodox mainstream. To be sure, this is understandable from a historical point of view since "the fight over embodied cognition in the 1990s was less about forging philosophically sound foundations for a new kind of cognitive science than it was about creating institutional space to allow such work to occur" [2]. However, the subsequent establishment of embodied AI as an important research program also means that it is time for the field to move beyond mere opposition to computationalist AI. Though there has been some excellent work done to incorporate these different approaches into a more unified framework of embodied AI (e.g. [103,100]) and cognitive science (e.g. [26,138]), such attempts have not been without their problems (cf. [46,2]). What is needed is a more coherent theoretical foundation:

> The current flourishing of embodied and situated approaches to AI, cognitive science and robotics has shown that the arguments from that period [i.e. the 1990s] were indeed convincing to many, but time and reflection has in fact cast

[7] The claim that our meaningful perspective is ultimately grounded in our precarious mode of being as metabolic entities does not, of course, strictly follow from some logical necessity of the argument. As such, it does not really solve the 'hard problem' [24] of why there is something 'it is like to be' [93] in the first place. Nevertheless, this mode of being does appear to have the right kind of existential characteristics, and we therefore suggest that the claim's validity should be judged in terms of the theoretical coherence it affords (see, for example, Section 3).

doubt on whether they were right. This is precisely the situation that most calls out for philosophical reflection. [2]

Fortunately, the conceptual framework provided by the development of the enactive approach in the cognitive sciences (e.g. [127,118]) might be exactly what is required in order to move the field of embodied AI into its next phase [49]. Indeed, there are already promising signs that this relationship will be a fruitful one. For example, while Di Paolo [36] also acknowledges that current embodied AI still lacks the property of intentional agency, he is nevertheless more optimistic than Dreyfus' [46] appraisal of the field. He suggests that "we have not yet seen the ending of this story, but that all the elements are in place at this moment for moving on to the next chapter. As before, practical concerns will be strong driving motivations for the development of the necessary ideas" [36]. Accordingly, we propose that this potential for moving forward should be directed toward the development of an enactive AI, namely an AI based on the principles of enactive cognitive science.

To put it briefly, enactive cognitive science has captured Jonas' [72] bio-philosophical insights in systemic terms by conceiving of metabolism as a particular physiochemical instantiation of a more general organizational principle, namely that of an *autonomous organization* [122,81]. Furthermore, this theory has recently been extended in a principled manner such that it can also account for the constitution of worldly significance through an understanding of cognition as *sense-making* [136,38]. The biological foundation of enactive cognitive science thus has the potential to help us address the problems currently faced by embodied AI from the bottom up, namely by starting from a systemic understanding of life as such. Accordingly, the contribution this paper makes can be firmly placed in the tradition of thinking which sees a strong continuity between life and mind (e.g. [81,111,112,137,136,36,116,118]).

However, before we are in a position to more precisely determine what a shift from embodied AI to enactive AI entails in terms of actually designing artificial systems (cf. Section 4), we must first briefly familiarize ourselves more generally with the theoretical foundations of the enactive approach. In particular, a few words about the label 'enactive' are in order so as to avoid any potential confusion.

2.6. Enactive cognitive science

The enactive approach to cognitive science was first introduced in 1991 with the publication of *The Embodied Mind* by Varela, Thompson and Rosch. This book brought many radical ideas to bear on cognitive science research, and drew inspiration from a wide variety of different sources such as Heidegger's [64] existentialism, Merleau-Ponty's [88] phenomenology of the body, as well as Buddhist psychology. One of the most popular ideas put forward by this book is an 'enactive' account of perception, namely the idea that perceptual experiences are not events that are internal to our heads, but are rather something which we *enact* or bring forth through our active engagement and sensorimotor exploration of our environment. A similar idea was later developed in an influential paper by O'Regan and Noë [97], in which the authors also argued that perception is an exploratory activity, and in particular that vision is a mode of exploration of the environment that is mediated by knowledge of sensorimotor contingencies. This much discussed paper was followed in 2004 by the publication of Noë's widely read book *Action in Perception*, which was essentially based on his work with O'Regan, but which also introduced the term 'enactive' in order to describe their sensorimotor contingency approach to perception.

These more recent developments have had the unfortunate effect that the notion of 'enaction' has for many researchers become almost exclusively associated with Noë's and O'Regan's work on the sensorimotor approach to perception, and often their work has been explicitly criticized under this label (e.g. [104,129]). This identification between the two approaches is problematic for the tradition of enactive cognitive science founded by Varela and colleagues because, while the sensorimotor contingency theory is in many ways compatible with this tradition, it is nevertheless lacking an appropriate foundation in lived phenomenology and especially in the biology of autonomous agency [117]. This has the consequence that, for example, O'Regan and Noë's sensorimotor account of perceptual awareness and experience is open to the criticisms presented in Section 2.4 of this paper.

Accordingly, we will use the term enactive cognitive science mainly to refer to the tradition started by Varela and colleagues. The biological foundation of this tradition, which can be traced back to Varela's early work with Maturana in the 1970s (e.g. [128,81,82]), was admittedly largely absent from *The Embodied Mind*. However, this aspect has recently been more explicitly developed (e.g. [136,116,38,120,41]). It is especially prominent in *Mind in Life*, a recent book by Thompson [118] that was originally destined to be the follow up to *The Embodied Mind* before Varela's untimely death. With this disclaimer in place we can now summarize the two main theoretical strands of contemporary enactive cognitive science in Table 2.

Table 2
Summary of the theoretical foundations of enactive cognitive science. The main focus is the study of subjectivity in both its lived and living dimensions using the methods of phenomenological philosophy and systems biology, respectively. Both of these methods can provide important insights for the field of AI.

#	Methodology	Phenomenon	Critical target in AI
ECS-1	phenomenological philosophy	lived subjectivity	computationalist AI
ECS-2	systems biology	living subjectivity	embodied AI

Enactive cognitive science has made use of phenomenological insights right from its inception (e.g. [127]), and phenomenology continues to be the main philosophical foundation of the enactive approach today (e.g. [118]). The investigation of lived experience through the methods of phenomenology is thus at the core of one of the main theoretical strands of enactive cognitive science (ECS-1).[8] Since Dreyfus' influential critique of (e.g. [42,43], [44, p. 118]) is also based on phenomenological insights, it is largely compatible with the enactive approach. Thus, it can be said that phenomenological philosophy is in many ways responsible for the shift toward embodied AI.

Indeed, the field has already started to incorporate many important phenomenological insights, for example the role of embodiment (e.g. [88]) as well as temporality and worldly situatedness (e.g. [64]). Nevertheless, something is still amiss in current embodied AI and, considering the arguments of this section, we are most likely to find out exactly what that lack is by paying closer attention to the essential aspects of biological embodiment. However, while phenomenology has shown itself to be a powerful method of criticizing overly intellectualist approaches to the study of mind, especially by exposing their guiding premises as based on a naïve understanding of conscious experience[9], it is less suitable for providing a detailed critical analysis of the field of embodied AI, though the work of Jonas [71,72] provides a helpful start.

In the next section we will elaborate this critical analysis of organismic embodiment by drawing on the second main theoretical strand of enactive cognitive science (ECS-2), namely a systems biological approach to intentional agency.

3. Biological foundations of enactive AI

In brief, a systemic approach to the biology of intentional agency lies at the very heart of the enactive approach to cognitive science (e.g. [118]). It is based on an account of constitutive autonomy and sense-making, which is essentially a synthesis drawn from a long tradition of philosophical biology and more recent developments in complex systems theory (e.g. [136]).

Accordingly, this section first highlights some important insights of the continental tradition of philosophical biology, and then unfolds the enactive account of intentional agency in three stages: (i) it outlines the central tenets and developments of the autopoietic tradition in theoretical biology leading up to the claim that constitutive autonomy is a necessary condition for intrinsic teleology [136], (ii) it argues that the additional systemic requirement of adaptivity is also necessary for sense-making and therefore for the constitution of a world of significance for the agent [38], and finally (iii) it evaluates the possibility that constitutive autonomy might not be a necessary requirement for sense-making, which, if true, would require less drastic changes in current methodologies of embodied AI in order to shift the field toward a fully enactive AI.

3.1. A view from philosophical biology

In Section 2 it was argued that both organisms and artifacts can be described as 'goal-directed', but that while (current) artifacts can only be characterized in this way because of their involvement in a purposeful context that is external to them (extrinsic teleology), organisms appear to have the peculiar capacity to enact and follow their own goals (intrinsic teleology). It is worth emphasizing here that, in contrast to enactive cognitive science, mainstream biology generally does not make any distinction between these two cases (cf. Appendix B). Fortunately, however, there is an alternative tradition in biology which attempts to explain the purposive being of organisms in a naturalistic but non-reductive manner. A brief look at some of the history of this alternative tradition will help us to better understand what it is about the systemic organization of living beings that enables us to attribute to them purposeful behavior which is motivated by goals that are genuinely their own.

Here we will highlight three different important influences on the enactive account of intentional agency: (i) Kant's notion of *natural purpose* as a necessary and sufficient condition for intrinsic teleology, (ii) von Uexküll's view of the organism as a living subject related to a corresponding *Umwelt* of significance, and (iii) Jonas' existential notion of *needful freedom* which is a part of his philosophy of biological individuality.

3.1.1. Kant and the notion of 'natural purpose'

It was Kant who first made the connection between the intrinsic teleology of organisms and a modern understanding of self-organization [136]. Kant refers to a living being as a *natural purpose*, a notion which he defines as follows: "a thing exists as a natural purpose *if it is* (though in a double sense) *both cause and effect of itself*" [73, §64]. In order to illustrate this notion Kant provides the example of a tree. A tree is a natural purpose in three distinct ways: (i) through reproduction the tree is both cause and effect of its *species*, (ii) through metabolism it also produces itself as an *individual*, and finally (iii) a *part* of the tree is also self-producing in as much as there is a mutual dependence between the preservation of one part and that of the others.

[8] We will not explicitly engage with the phenomenological foundations of enactive cognitive science any further in this paper (but see [118]). For a general introduction to phenomenology, see [119] as well as [54]; for a brief analysis of phenomenology's relationship to the ongoing paradigm shift in AI, see [49].

[9] This is not to say that enactive cognitive science simply *replaces* the more traditional cognitivist conception of mind and cognition. See Appendix X for a brief analysis of the two paradigms.

Furthermore, Kant outlines two criteria that must be met for something to be considered as a natural purpose: (i) every part exists for the sake of the others and the whole, and (ii) the parts combine into the unity of a whole because they are reciprocally cause and effect of their form (cf. [118, p. 134]). Note that while both artifacts and organisms fulfill criterion (i), generally only living beings also fulfill criterion (ii). However, only when both criteria (i) and (ii) are met can something be considered as a natural purpose:

> In such a product of nature every part, as existing through all the other parts, is also thought as existing for the sake of the others and that of the whole, i.e. as a tool (organ); [...] an organ bringing forth the other parts (and hence everyone bringing forth another) [...]; and only then and because of this such a product as an *organized* and *self-organizing* being can be called a *natural purpose*.
> [73, §65]

Moreover, Kant claims that, because of this self-organizing reciprocal causality, it follows that all relations of cause and effect in the system are also at the same time relations of means and purpose. More importantly, this reciprocal causality entails that a natural purpose then, as an interrelated totality of means and goals, is strictly intrinsic to the organism [136]. Kant's philosophy thus provides the beginning of a theory of the organization of the living which attempts to capture the observation that organisms appear to generate their own goals. This organization is characterized by a special kind of self-organization that is better described as a form of *self-production*: a living system is both cause and effect of itself.

Still, for Kant it was inconceivable that this peculiar organization of the living could be understood without having recourse to the idea of some kind of teleological causality as an observer-dependent regulatory concept, which led him to suggest that natural purposes are not purely naturalistically explicable [136]. However, recent advances in complex systems theory have started to provide the necessary conceptual tools for a scientific framework. In particular, the notions of circular causation and nonlinear emergence are promising candidates for this job. They allow us to capture the dynamics of a self-perpetuating whole that self-organizes out of a network of local processes while subsuming those very processes so that they no longer have a merely local and independent identity (cf. [118, p. 138]).

3.1.2. Von Uexküll and the notion of 'Umwelt'

After this consideration of Kant's influence on the biological foundations of enactive cognitive science, let us now briefly introduce the biologist von Uexküll [132–134]. Von Uexküll considered it the task of biology to expand the result of Kant's philosophical research by investigating the role of the living body in shaping the relationship between subjects and their worlds. The influence of von Uexküll's work on embodied AI has already been widely acknowledged in the literature (e.g. [109,148,142,80]), and it has also extensively informed the biological basis of enactive cognitive science.

In the present context it is interesting to note that von Uexküll considered the self-constituting autonomy of the living as the essential difference between artificial mechanisms and organisms. He observed that it allows organisms, unlike machines (at least in von Uexküll's time), for example, to repair themselves when they are damaged (cf. [142,148]). This conception of the autonomy of the living is also closely related to what von Uexküll [134] described as the "principal difference between the construction of a mechanism and a living organism", namely that "the organs of living beings have an innate meaning-quality, in contrast to the parts of machine; therefore they can only develop centrifugally". Accordingly, similar to Kant's notion of natural purpose, von Uexküll's notion of *centrifugal development* also emphasizes the intrinsic goal-directedness of organisms:

> Every machine, a pocket watch for example, is always constructed centripetally. In other words, the individual parts of the watch, such as its hands, springs, wheels, and cogs, must always be produced first, so that they may be added to a common centerpiece. In contrast, the construction of an animal, for example, a triton, always starts centrifugally from a single cell, which first develops into a gastrula, and then into more and more new organ buds. In both cases, the transformation underlies a plan: the 'watch-plan' proceeds centripetally and the 'triton-plan' centrifugally. Two completely opposite principles govern the joining of the parts of the two objects.
> [134, p. 40]

Von Uexküll therefore rejected purely mechanistic/behavioristic descriptions of living organisms because they overlooked, according to him, the centrifugal organization which integrates the organism's components into a purposeful whole.

Von Uexküll's starting point, namely that animals are subjects in their own right, also forms the basis for the concept he is most famous for, that of the *Umwelt*. While this concept is often used in embodied AI to refer to sensorimotor embeddedness, it is important to realize that for von Uexküll it denotes a world of significance precisely because it was grounded in the subjectivity of the organism. The *Umwelt* cannot be divorced from the internal organization of the organism; it is both generated by it and causally connected to its ongoing preservation [36]. In this manner von Uexküll extends Kant's work on intrinsic teleology by considering how an animal's relation to its environment constitutes a meaningful perspective for that animal:

> [W]e who still hold that our sense organs serve our perceptions, and our motor organs our actions, see in animals as well not only the mechanical structure, but also the operator, who is built into their organs as we are into our bodies.

> We no longer regard animals as mere machines, but as subjects whose essential activity consists of perceiving and acting. We thus unlock the gates that lead to other realms, for all that a subject perceives becomes his perceptual world and all that he does, his effector world. Perceptual and effector worlds together form a closed unit, the *Umwelt*.
> [133, p. 6]

In spite of von Uexküll's rather old-fashioned language, we can identify and elaborate three important steps in his argument that lead him to posit an *Umwelt* for all animals:

(i) if we choose to accept as evidence our own lived experience of a world that we perceive as a meaningful basis for intentional action, then it is possible to reject the claim that our being is exhausted by its external mechanical structure, and
(ii) if we choose to accept the evidence for (i) as well as our own lived experience of a world in which we perceive other animals as intentional agents in their own right, then it is also possible to reject the claim that another animal's being is exhausted by its mechanical structure, and
(iii) if we accept the evidence for (i) and (ii), it becomes reasonable to assume that what can be scientifically described as an animal's sensorimotor behavior constitutes for the animal its own lived world, or *Umwelt*.

We suggest that it is von Uexküll's appeal to the evidence of our own *lived experience*, which directly reveals us and other living beings as embodied subjects, that forms the basis for his research. This conception of living beings as embodied subjects is the crucial premise for his biological investigations, especially as it motivates his distinction of the centrifugal organization of living bodies and his claim that their sensorimotor interaction constitutes for them a world of meaning. This starting point in lived experience is very much in line with enactive cognitive science's insistence on the importance of phenomenology for our understanding of life and mind (cf. [118]). Interestingly, von Uexküll's conception of the organism as a living subject and his notion of the *Umwelt* were incorporated by Heidegger into his philosophical analysis of the different existential situations that are characteristic of material, living and human being [65]. Heidegger's existential account of the living mode of being still deserves closer study, especially in relation to the biological foundations of enactive cognitive science, but this endeavor is beyond the scope if the current paper.

3.1.3. Jonas and the notion of 'needful freedom'

Let us now turn to one of the most important philosophical influences on the enactive account of intentional agency, namely the writings of the bio-philosopher Hans Jonas [71,72]. In some of his early work Jonas, who was a student of Heidegger, is especially concerned with the peculiar mode of identity that is characteristic of living beings. Whereas philosophy has traditionally viewed the individual as either something that is differentiated in space and time or separated by its substantial essence, Jonas claims that:

> Only those entities are individuals whose being is their own doing [...]. Entities, therefore, which in their being are exposed to the alternative of not-being as potentially imminent [...]; whose *identity* over time is thus, not the inert one of a permanent substratum, but the self-created one of continuous performance.
> [72]

Just like Kant and von Uexküll before him, Jonas thus points to an essential difference between living and non-living beings, since only the identity of the former can be said to be intrinsically generated by that being for its own being.

How does the peculiar mode of being which is achieved in metabolism relate to the problem of meaning in AI? Jonas proposes that the process of ongoing metabolic self-construction has the necessary existential characteristics such that we can speak of the constitution of a meaningful perspective by that process for that process. More precisely, we can identify three aspects: (i) the ongoing metabolic self-generation of a distinct 'self' by which a living being separates itself from non-living matter, (ii) the precarious existence of this 'self' which is continually faced by material and energetic requirements, and (iii) the establishment of a basic *normativity* in relation to whether events are good or bad for the continuation of this living being (cf. [136]).

In this manner we can link Jonas's account of the precarious self-production of an identity to the constitution of a world of significance for that identity. Indeed, for Jonas this is not a contingent relationship: metabolic individuals only achieve their being in answer to the constant possibility of not-being, namely of becoming something 'other'. They are entities "whose *difference* from the *other*, from the rest of things, is not adventitious and indifferent to them, but a dynamic attribute of their being, in that the tension of this difference is the very medium of each one's maintaining itself in its selfhood by standing off the other and communing with it at the same time" [72]. Thus, there is an inherent and insurmountable tension in the living mode of being since an organism needs to separate itself from non-living matter in order to preserve its identity, but at the same time it is also dependent on that matter as the background from which it can distinguish its identity.

Jonas coined the phrase *needful freedom* to denote this peculiar relation of the organic form to matter (e.g. [72, p. 80]). This relation is best expressed through the fact that, while the existential form of an organism is independent of any *particular* configuration of matter through which it passes in virtue of its metabolism, it is nevertheless dependent on

the continual perpetuation of this ongoing flow of material configurations. If this flow ceases and the organic form fully coincides with its current material configuration then it is no longer living.

Another way to illustrate the notion of needful freedom is to contrast living being with the being of one of our current robots. In that case, the system's identity is determined and fully exhausted by its current material configuration. Of course, in engineering terms this is certainly useful because it enables the designer to impose the system's identity by arranging matter in a suitable manner. However, this also necessarily entails that the robot's identity is defined by external circumstances rather than being intrinsically generated. And since the systemic identity is not continually generated by the system itself, it also does not depend on anything other than itself to continue this substantial mode of being – even if the robot eventually runs out of energy it still remains to be the same system (cf. [61]). Again, this is clearly useful from an engineering perspective since we do not want our artifacts to fall apart when they run out of power, but it also means that a robot's relation to its environment cannot be characterized in terms of needful freedom.

3.1.4. Philosophical insights for embodied AI

We are now in a better position to understand why the field of AI has had significant difficulties in even diagnosing the nature of the problem that is preventing it from designing more lifelike machines (cf. [22,23]). It appears that the field's strict adherence to the synthetic methodology, namely "understanding by building" (design principle P-1), is making it constitutively blind to some of the essential aspects of the living organization. Life is not something which can simply be built from the 'outside' like a robot; it needs to self-generate under certain conditions.

The writings of Kant, von Uexküll, and Jonas thus provide the field of embodied AI with valuable insights with regard to the organismic foundation of intrinsic goal generation, worldly significance, and biological identity, respectively. Most importantly, their different points of view converge on the claim that, in contrast to non-living matter, the being of an organism is intrinsically generated by the ongoing activity of that organism. Jonas then elaborates this common observation into the philosophical claim that having a meaningful perspective of concern for the world is an essential property of those systems which are continuously threatened by the possibility of non-being, and which prevent this fatal event by directing their own activity toward on ongoing self-realization. In other words, a world of significance, or *Umwelt*, is encountered only by those systems *whose being is their own doing*. This ontological difference between living and non-living beings poses a significant challenge to the synthetic methodology.

Admittedly, these considerations are couched in rather vague formulations and as such are unlikely to be of direct help for most researchers working in embodied AI. Fortunately, there have also been corresponding attempts in the field of theoretical biology to give a more precise systemic definition of the living identity.

3.2. The enactive approach to intentional agency

We are now in a position to unpack the biological foundations of enactive cognitive science in more detail. We will begin by introducing the systemic concepts of autopoiesis, organizational closure, and constitutive autonomy (Section 3.2.1). This is followed by a closer consideration of the notion of sense-making and its necessary dependence on both adaptivity (3.2.2) and constitutive autonomy (3.2.3). The latter subsection illustrates the systemic requirements for sense-making in relation to AI.

3.2.1. Constitutive autonomy is necessary for intrinsic teleology

The notion of *autopoiesis* as the minimal organization of the living first originated in the work of the Chilean biologists Maturana and Varela in the 1970s (e.g. [128,81]; see [82] for a popular introduction). While the concept was developed in the context of theoretical biology, it was right from its inception also associated with computer simulations [128] long before the term "artificial life" was first introduced in the late 1980s [77]. Nowadays the concept of autopoiesis continues to have a significant impact on the field of artificial life in both the computational and chemical domain (see [86] and [78], respectively, for overviews of these two kinds of approaches). Moreover, there have been recent efforts of more tightly integrating the notion of autopoiesis into the overall framework of enactive cognitive science (e.g. [136,117,118,38,85,28]).

During the time after the notion of autopoiesis was first coined in 1971[10] its exact definition has slowly evolved in the works of both Maturana and Varela (cf. [118, pp. 99–101], [18]). For the purposes of this article we will use a definition that has been used extensively by Varela in a series of publications throughout the 1990s (e.g. [123,124,126]), but which has also been used as the definition of choice in more recent work (e.g. [136,38,36,52]). According to this definition "an autopoietic system – the minimal living organization – is one that continuously produces the components that specify it, while at the same time realizing it (the system) as a concrete unity in space and time, which makes the network of production of components possible" [123]:

> More precisely defined: an autopoietic system is organized (defined as a unity) as a network of processes of production (synthesis and destruction) of components such that these components:

[10] See [125] for an account of the historical circumstances under which the notion of autopoiesis was first conceived and developed.

1. continuously regenerate and realize the network that produces them, and
2. constitute the system as a distinguishable unity in the domain in which they exist.

[123]

In addition to the two explicit criteria for autopoiesis we can add another important point, namely that the self-constitution of an identity entails the constitution of a relational domain between the system and its environment. The shape of this domain is not pre-given but rather co-determined by the organization of the system, as it is produced by that system, and its environment. Accordingly, any system which fulfills the criteria for autopoiesis also generates its own domain of possible interactions in the same movement that gives rise to its emergent identity [118, p. 44].

From the point of view that current embodied AI fails to capture intentional agency, it is interesting to note that the autopoietic tradition has been explicitly characterized as a "biology of intentionality" [124]. In other words, for enactive cognitive science the phenomenon of autopoiesis not only captures the basic mode of identity of the living, but is moreover at the root at how living beings enact their world of significance. Thus, the notion of autopoiesis in many ways reflects Kant's, von Uexküll's, and Jonas' intuitions regarding the organization of the living, but with the added advantage that it formalizes them in a systemic, *operational* manner.[11]

The paradigmatic example of an autopoietic system is a single living cell [128], an example which is useful to illustrate the circularity that is inherent in metabolic self-production. In the case of the cell this circularity is expressed in the co-dependency between the (boundary) semi-permeable membrane and the (internal) metabolic network. The metabolic network constructs itself as well as the membrane, and thereby distinguishes itself as a unified system from the (external) environment. In turn, the membrane makes the metabolism possible by preventing the network from fatally diffusing into the environment.

While there are cases in the literature where multi-cellular organisms are also classed as autopoietic systems in their own right, this is an issue that is far from trivial and still remains controversial (cf. [118, pp. 105–107]). Nevertheless, we intuitively want to say that such organisms also meet the requirements for *autonomy* in that a multi-cellular organism is "distinctly different from an autopoietic minimal entity in its mode of identity, but similar in that it demarcates an autonomous entity from its environment" [123]. Indeed, in the late 1970s Varela became dissatisfied with the way that the concept of autopoiesis was starting to be applied loosely to other systems, with its use even extended to non-material systems such as social institutions. He complained that such characterizations "confuse autopoiesis with autonomy" [122, p. 55]. Nevertheless, there was still a need to make the explanatory power offered by the systemic approach to autonomy available for use in other contexts than the molecular domain. Thus, while autopoiesis is a form of autonomy in the biochemical domain, "to qualify as autonomy, however, a system does not have to be autopoietic in the strict sense (a self-producing bounded molecular system)" [118, p. 44].

Accordingly, Varela put forward the notion of *organizational closure*[12] by taking "the lessons offered by the autonomy of living systems and convert them into an operational characterization of *autonomy in general*, living or otherwise" [122, p. 55]:

> We shall say that autonomous systems are organizationally closed. That is, their organization is characterized by processes such that
> 1. the processes are related as a network, so that they recursively depend on each other in the generation and realization of the processes themselves, and
> 2. they constitute the system as a unity recognizable in the space (domain) in which the processes exist.
>
> [122, p. 55]

This definition of autonomy applies to multi-cellular organisms [82, pp. 88–89], but moreover to a whole host of other systems such as the immune system, the nervous system, and even to social systems [123]. Since it does not specify the particular domain of the autonomous system, it is also to some extent more amenable to the sciences of the artificial [52]. Maturana and Varela [82] introduced the following ideograms to denote systems which are characterized by organizational closure (Fig. 1).

We will refer to the autonomy entailed by organizational closure as *constitutive autonomy* in order to demarcate it from the concept's more general usage (cf. [52]). For a more detailed description of how the notions of emergence through self-organization, constitutive autonomy, and autopoiesis relate to each other, see Appendix C.

In summary, when we are referring to an autonomous system we denote a system composed of several processes that actively generate and sustain the systemic identity under precarious conditions (e.g. [39]). The precariousness of the identity

[11] The term 'operational' denotes that the autopoietic definition of life can be used to distinguish living from non-living entities on the basis of a concrete instance and without recourse to wider contextual (e.g. functional, historical) considerations. Autopoiesis can be considered as a response to the question of how we can determine whether a system is a living being or not on the basis of what kind of system it is rather than on how it behaves or where it came from. As such it can be contrasted with functional (e.g. [94]) or historical (e.g. [89]) approaches to intentional agency.

[12] In recent literature the term *organizational* closure is often used more or less interchangeably with the notion of *operational* closure. However, the latter seems better suited to describe any system which has been distinguished in a certain epistemological manner by an external observer, namely so as *not* to view the system under study as characterized by inputs/outputs, but rather as a self-contained system which is parametrically coupled to its environment. On this view, an *organizationally* closed system is a special kind of system, namely one which is characterized by some form of self-production when it is appropriately distinguished by an external observer as *operationally* closed.

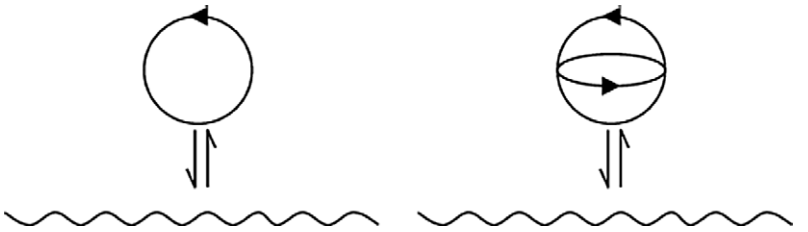

Fig. 1. Maturana and Varela's ideograms for autonomous systems, namely those systems which can be characterized by organizational closure. The ideogram on the left depicts a basic autonomous system: the closed arrow circle indicates the system with organizational closure, the rippled line its environment, and the bidirectional half-arrows the ongoing structural coupling between the two. The ideogram on the right extends this basic picture by introducing another organizational closure within the autonomous system, which could be the nervous system, for example.

is explicitly mentioned in order to emphasize that the system's identity is actively constituted by the system under conditions which tend toward its disintegration, and which is therefore constantly under threat of ceasing to exist. Accordingly, this working definition of constitutive autonomy captures the essential insights of both the situation of the organism as described in the philosophical biology tradition, as well as the operational definitions provided by the autopoietic tradition. Both of these traditions converge on the claim that it is this self-constitution of an identity, an identity that could at each moment become something different or disappear altogether, which grounds our understanding of intrinsic teleology [136]. These considerations allow us to state the first core claim of enactive cognitive science in the form of a systemic requirement (SR-1): *constitutive autonomy is necessary for intrinsic teleology*.

3.2.2. Adaptivity is necessary for sense-making

In contrast to the robots of embodied AI whose identity and domain of interactions is externally defined, constitutively autonomous systems bring forth their own identity and domain of interactions, and thereby constitute their own 'problems to be solved' according to their particular affordances for action. Such autonomous systems and "their worlds of meaning stand in relation to each other through *mutual specification* or *co-determination*" [124]. In other words, there is a mutual dependence between the intentional agent (which must exist in some world) and its world (which can only be encountered by such an agent): there is a fundamental circularity at the core of intentionality [85].

Furthermore, what an autonomous system does, due to its precarious mode of identity, is to treat the perturbations it encounters from a perspective of significance which is not intrinsic to the encounters themselves. In other words, the meaning of an encounter is not determined by that encounter. Instead it is evaluated in relation to the ongoing maintenance of the self-constituted identity, and thereby acquires a meaning which is relative to the current situation of the agent and its needs [123]. This process of meaning generation in relation to the perspective of the agent is what is meant by the notion of *sense-making* [136]. Translating this concept into von Uexküll's terms we could say that sense-making is the ongoing process of active constitution of an *Umwelt* for the organism. It is important to note that the significance which is continuously brought forth by the endogenous activity of the autonomous agent is what makes the world, as it appears from the perspective of that agent, distinct from the physical environment of the autonomous system, as it is distinguished by an external observer [126]. Sense-making is the enaction of a meaningful world *for* the autonomous agent.

Similar to Jonas' notion of 'needful freedom', the enactive account of constitutive autonomy and sense-making entails that meaning is not to be found in the elements belonging to the environment or in the internal dynamics of the agent. Instead, meaning is an aspect of the *relational* domain established between the two [41]. It depends on the specific mode of co-determination that each autonomous system realizes with its environment, and accordingly different modes of structural coupling will give rise to different meanings (Colombetti, in press). However, it is important to note that the claim that meaning is grounded in such relations does not entail that meaning can be *reduced* to those relational phenomena. There is an asymmetry underlying the relational domain of an autonomous system since the very existence of that domain is continuously enacted by the endogenous activity of that system. In contrast to most embodied AI, where the relational domain exists no matter what the system is or does, the relational domain of a living system is not pre-given. It follows from this that any model that only captures the relational dynamics on their own, as is the case with most work on sensorimotor situatedness, will only be able to capture the functional aspects of the behavior but not its intrinsic meaning. We will return to this problem in Section 3.2.3.

In order for these considerations to be of more specific use for the development of better models of natural cognition, we need to unpack the notion of sense-making in more detail. Essentially, it requires that the perturbations which an autonomous agent encounters through its ongoing interactions must somehow acquire a valence that is related to the agent's viability. Varela [124] has argued that "the source of this world-making is always the breakdowns in autopoiesis". However, the concept of autopoiesis (or constitutive autonomy more generally) by itself allows no gradation – either a system belongs to the class of such systems or it does not. The self-constitution of an identity can thus provide us only with the most basic kind of norm, namely that all events are good for that identity as long as they do not destroy it (and the latter events do not carry any significance because there will be no more identity to which they could even be related). On this basis alone there is no room for accounting for the different shades of meaning which are constitutive of an organism's *Umwelt*. Furthermore, the operational definitions of autopoiesis and constitutive autonomy neither require

Table 3
Summary of the enactive approach to intentional agency, which includes at least two necessary conditions: (i) constitutive autonomy is necessary for intrinsic teleology, and (ii) adaptivity is necessary for sense-making.

#	Systemic requirement	Entailment	Normativity
SR-1	constitutive autonomy	intrinsic teleology	uniform
SR-2	adaptivity	sense-making	graded

that such a system can actively compensate for deleterious internal or external events, nor address the possibility that it can spontaneously improve its current situation. What is missing from these definitions? How can we extend the meaningful perspective that is engendered by constitutive autonomy into a wider context of relevance?

Di Paolo [38] has recently proposed a possible resolution of this problem. He starts from the observation that minimal autopoietic systems have a certain kind of tolerance or *robustness*: "they can sustain a certain range of perturbations as well as a certain range of internal structural changes before they lose their autopoiesis", where "these ranges are defined by the organization and current state of the system". We can then define these ranges of non-fatal events as an autonomous system's *viability set*, which is "assumed to be of finite measure, bounded, and possibly time-varying" [38]. However, in order for an autopoietic system to actively improve its current situation, it must (i) be capable of determining how the ongoing structural changes are shaping its trajectory within its viability set, and (ii) have the capacity to regulate the conditions of this trajectory appropriately. These two criteria are provided by the property of *adaptivity*, for which Di Paolo [38] provides the following definition:

> A system's capacity, in some circumstances, to regulate its states and its relation to the environment with the result that, if the states are sufficiently close to the boundary of viability,
> 1. Tendencies are distinguished and acted upon depending on whether the states will approach or recede from the boundary and, as a consequence,
> 2. Tendencies of the first kind are moved closer to or transformed into tendencies of the second and so future states are prevented from reaching the boundary with an outward velocity.
>
> [38]

Similar to the case of robustness, the notion of adaptivity[13] implies tolerance of a range of internal and external perturbations. However, in this context it entails a special kind of context-sensitive tolerance which involves both "actively monitoring perturbations and compensating for their tendencies" [38]. The explicit requirement of active monitoring is crucial for two reasons: (i) it allows the system to distinguish between positive and negative tendencies, and (ii) it ensures that the system can measure the type and severity of a tendency according to a change in the regulative resources required.

It is important to note that the capacity for (i) does not contradict the organizational closure of the autonomous system because of (ii). In other words, the system does not have any special epistemic access to an independent (non-relational) environment, and it therefore does not violate the relational nature of constitutive autonomy, but this is not a problem since it only needs to monitor internal effort. Furthermore, it is worth emphasizing that the capacity for (ii) already implies the need for suitable compensation. In the context of sense-making we can therefore say that both elements, i.e. self-monitoring and appropriate regulation, are necessary to be able to speak of different kinds of meaning from the perspective of the organism. Thus, "if autopoiesis in the present analysis suffices for generating a natural purpose, adaptivity reflects the organism's capability – necessary for sense-making – of evaluating the needs and expanding the means towards that purpose" [38].

While it is likely that some form of adaptivity as defined here was assumed to be implicit in the definition of autopoiesis as constitutive of sense-making (e.g. [136]), it is nevertheless useful to turn this assumption into an explicit claim. This allows us to state the second core claim of enactive cognitive science in the form of another systemic requirement (SR-2): *adaptivity is necessary for sense-making*.

3.2.3. Constitutive autonomy is necessary for sense-making

We can now summarize the enactive approach to intentional agency in Table 3. We have argued that the systemic requirements of constitutive autonomy and adaptivity are necessary for intrinsic teleology and sense-making, respectively. They are also necessary for intentional agency in the sense that having a purposeful and meaningful perspective on a world is at least partly constitutive of what it means to be such an agent. However, we are not making the stronger claim that constitutive autonomy and adaptivity are also *sufficient* conditions for intentional agency. In fact, we expect that more systemic requirements will be added to this list as the enactive approach begins to address a wider range of phenomena. Some promising lines of research in this regard are the development of an enactive approach to emotion theory [28], to goals and goal-directedness [85] and to social cognition [32]. All of these developments are consistent and continuous with the notions of constitutive autonomy and sense-making as they have been presented in this paper.

[13] Note that this form of adaptivity, as a special kind of self-regulatory mechanism, must be clearly distinguished from the more general notion of 'adaptedness'. This latter sense is usually used to indicate all viable behavior that has evolutionary origins and contributes to reproductive success.

When it comes to the practical challenge of how to go about implementing the two systemic requirements in the form of artificial systems it might be tempting to initially avoid SR-1, which we will call the 'hard problem' of enactive AI, and first focus our attempts to address SR-2. However, is it possible to design systems with adaptivity as the basis for sense-making independently of constitutive autonomy?

While an affirmative answer to this question might sound desirable for AI research, unfortunately things are not as simple. This is best illustrated by an analysis of the relationship between autopoiesis and cognition as it has been presented by Bourgine and Stewart [18] in a paper which bases its insights on a mathematical model of autopoiesis. Whereas traditionally it was held that 'autopoiesis = life = cognition' (e.g. [81,111,112]), Bourgine and Stewart propose that "a system is cognitive if and only if type A interactions serve to trigger type B interactions in a specific way, so as to satisfy a viability constraint", where "type A interactions can be termed 'sensations', and type B interactions can be termed 'actions' ".[14] They note that their notion of "viability constraint" has been deliberately left vague so that their definition of cognition can by "metaphorical extension" also be usefully applied to non-living systems.

As a hypothetical example they describe a robot that navigates on the surface of a table by satisfying the constraints of neither remaining immobile nor falling of the edge. Since this robot is cognitive by definition when it satisfies the imposed viability constraint, but certainly not autopoietic, Bourgine and Stewart [18] claim that autopoiesis is not a necessary condition for cognition. Furthermore, they provide a mathematical model of a simple chemical system, which they maintain is autopoietic but for which it is nevertheless impossible to speak of "action" and "sensation" in any meaningful manner. Accordingly, they also make the second claim that autopoiesis is not a sufficient condition for cognition.

While this last claim might sound at least vaguely analogous to Di Paolo's [38] argument that minimal autopoiesis is insufficient to account for sense-making, there are some important differences. It is worth noting that, as side effect of not further restricting what is to count as a "viability constraint", Bourgine and Stewart's [18] definition of cognition is different from Di Paolo's [38] notion of sense-making for two important reasons: (i) the viability constraint can be externally defined (as illustrated by the example of the robot), and (ii) even if the viability constraint was intrinsic to the cognitive system, there is no requirement for that system to measure and actively regulate its performance with regard to satisfying that constraint.

To illustrate the consequences of (i) we can imagine defining an additional arbitrary constraint for the hypothetical navigating robot, namely that it must also always stay on only one side of the table. Accordingly, we would have to treat it as cognitive as long as it happens to stay on that side, but as non-cognitive as soon as it moves to the other side of the table. Clearly, whether the robot stays on one side or the other does not make any difference to the system itself but only to the experimenter who is imposing the viability criteria (whether these be externalized into the robot or not). Thus, the only overlap between their definition of cognition and Di Paolo's [38] view of sense-making is that both require the capacity for some form of sensorimotor interaction which, as we have seen, is not sufficient for grounding meaning.

It will also be interesting from the perspective of AI to draw out the consequences of the second difference (ii). Bourgine and Stewart's [18] claim that a given interaction between a system and its environment "will not be cognitive unless the consequences for the internal state of the system are employed to trigger specific actions that promote the viability of the system". How do we know what constitutes an *action*? They define actions as "those interactions that have consequences for the state of the (proximal) environment, or that modify the relation of the system to its environment". However, this criterion is trivially met by all systems which are structurally coupled to their environment since any kind of interaction (whether originating from the system or the environment) changes the relation of the system to its environment at some level of description. Thus, while their definition enables the movement of the hypothetical navigating robot to be classed as an action, it has also has the undesirable effect of making it impossible to distinguish whether it is the system or the environment that is the 'agent' giving rise to this action.

In order to remove this ambiguity we can follow [38] in drawing "the important distinction between structural coupling and the *regulation* of structural coupling" since only the latter "fully deserves the name of *behavior* because such regulation is *done* by the organism [...] as opposed to simply being *undergone* by it". This regulative capacity is captured by the notion of adaptivity. Moreover, this view entails that "cognition requires a natural center of *activity* on the world as well as a natural *perspective* on it" [38]. We have already seen that it is the principle of constitutive autonomy which introduces this required asymmetry: an autonomous system brings forth the relational domain that forms the basis for adaptive regulation by constituting its own identity, which is the reference point for its domain of possible interactions. It is essentially the lack of this asymmetry in Bourgine and Stewart's conception of cognition, which has made their proposal problematic. This allows us to combine the previous two core claims of enactivism as follows: *constitutive autonomy and adaptivity are both necessary for sense-making*.

This line of reasoning is further supported by recent work on chemical autopoiesis by Bitbol and Luisi [16]. While they broadly agree with Bourgine and Stewart's [18] definition of cognition, provided that extended homeostasis is considered to be a special variety of sensorimotor action, they nevertheless reject its proposed radical dissociation from autopoiesis. Thus, while Maturana and Varela's original assertion was that autopoiesis is strictly equivalent to cognition, Bitbol and Luisi [16] weaken this claim slightly by holding that minimal cognition requires both (i) the self-constitution of an identity

[14] They further clarify their position by stating that "it is only analytically that we can separate sensory inputs and actions; since the sensory inputs guide the actions, but the actions have consequences for subsequent sensory inputs, the two together form a dynamic loop. Cognition, in the present perspective, amounts to the emergent characteristics of this dynamical system" [18].

(constitutive autonomy), and (ii) dynamical interaction with the environment. Since they maintain that minimal autopoiesis can provide the foundation for (i) but does not necessarily entail (ii), it follows that their position, like Di Paolo's [38], falls between the extreme positions of radical identity (e.g. [81,111,112]) and radical dissociation (e.g. [18]).

This conclusion further supports the case already developed in Section 2, namely that sensorimotor interaction alone is not sufficient to ground intrinsic meaning and goal ownership. For a more extensive argument that constitutive autonomy is a necessary condition for natural agency, adaptivity and 'higher-level' cognition in terms of biological considerations, see [9].

3.3. From life to mind: the enactive framework

This has been a long section dealing mainly with issues that belong to theoretical biology so it is appropriate to briefly relate its main conclusion, namely that relational phenomena such as cognition, behavior, and sense-making cannot be decoupled from the operations of a constitutively autonomous and adaptive system without rendering them intrinsically meaningless, to enactive cognitive science as a whole. How does this systems biological foundation inform the enactive approach's views on cognition and the mind?

One of the most important consequences of the conclusion we have developed in this section is that it strongly underlines the deep continuity between life and mind, and it is this continuity which forms the very core of the theoretical foundation of enactive cognitive science. In order to better illustrate this link between the systemic approach to biology, as it has been presented in this section of the paper, and the enactive approach as a cognitive science research program we have adapted Thompson's [116] five steps from life to mind for the present context:

1. Life = constitutive autonomy + adaptivity
2. Constitutive autonomy entails emergence of an identity
3. Emergence of an adaptive identity entails emergence of a world
4. Emergence of adaptive identity and world = sense-making
5. Sense-making = cognition

This section has provided only the basic theoretical elements for an understanding of these five steps, and we cannot explicate them in any more detail here. We would only like to emphasize that, as Thompson points out, these steps amount to an explicit hypothesis about the natural roots of intentionality. In other words, they form the basis of the claim that the 'aboutness' of our cognition is not due to some presumed representational content that is matched to an independent external reality (by some designer or evolution), but is rather related to the significance that is continually enacted by the precarious activity of the organism during its ongoing encounters with the environment. Here we thus have the beginnings of how the enactive approach might go about incorporating its two main foundations, namely phenomenological philosophy (ECS-1) and systems biology (ECS-2), into one coherent theoretical framework (cf. [118]).

It is also worth noting that the field of AI has not yet properly responded to this shift of focus toward phenomenology and biology, which has had the effect that the insights generated by the models of this field have become less relevant in terms of pushing the conceptual boundaries of the most recent developments in cognitive science [49]. It will be interesting to see how the field will cope with this changing conceptual landscape. On the side of phenomenology it is likely that technology will continue to play an important role, but more in terms of artificial systems that involve a human subject, for example in the form of human–computer or enactive interfaces (e.g. [51]).

While this might seem like an undesirable conclusion from the perspective of proper AI and robotics, all is not lost: the field's biggest opportunity to make a significant contribution is to use its synthetic methodology to better explicate the biological foundations of intentional agency, especially because the organization of actual living systems are still prohibitively complex for our understanding. Nevertheless, it is also clear that the shift toward enactive cognitive science presents a unique challenge to the most fundamental design principle of embodied AI (P-1), namely the idea that we can do 'philosophy of mind using a screwdriver' [56] and thus gain 'understanding by building' [103]. In the next section we provide at least the beginnings of an enactive approach to AI, by discussing how this formidable challenge can be addressed in a pragmatic manner. The general aim is to put the development of enactive AI on a more concrete foundation.

Accordingly, we will propose some essential design principles for enactive AI, which are based on the philosophical and biological considerations that have been developed in the previous sections. Some relevant AI work will be reviewed in terms of its successes and failures to implement these principles.

4. Design principles for enactive AI

In the previous section we have outlined the biological foundations of the enactive account of intentional agency as grounded in the systemic requirements of constitutive autonomy and adaptivity. In this section we will make use of that theoretical background by proposing two design principles which should be taken into account by those researchers interested in developing fully enactive AI (Sections 4.1 and 4.2). This is followed by a general discussion of the scope and aims of enactive AI as we envision it in this paper (Section 4.3). We then introduce a set of schematics which categorizes current work in AI in order to focus the discussion on the relevant issues, and some relevant examples will be analyzed in terms of how well they satisfy the enactive AI design principles (Section 4.4).

4.1. Toward enactive AI (i): constitutive autonomy

Before we introduce the design principles for enactive AI we should make it clear that they are not meant to replace embodied AI, but mainly complement and refine the important insights that have already been generated by the embodied approach. In essence, it is complementing the latter's focus on engineering the emergence of robust and flexible real-world behavior with a concern for the emergence of an autonomous identity which can constitute and regulate that behavior. Another way of phrasing this change in perspective is to view embodied AI as attempting to ground meaning in sensorimotor interaction, and enactive AI as attempting to ground that sensorimotor interaction itself in autonomous agency. We will indicate how the new design principles affect those of embodied AI as we go along.

Let us recall the most essential difference between living and non-living systems, namely that the former is characterized by an ontological mode of 'being by doing'. In contrast, the mode of the latter can be compared with Jonas [72] description of a particle as something which is individuated in space and time, but only insofar as it is "simply and inertly what it is, immediately identical with itself, without the need to maintain that self-identity as an act of its existence. Its enduring is mere remaining, not re-assertion of being". This static form of identity can also be attributed to the systems developed by current AI. Whether a robot is switched on and engaged in its operations or switched off, it remains to be the same system in both cases. In contrast, when a living being ceases in its operations, it also ceases to be a living being. It follows from this that "we cannot truly model what is proper to the living organization unless we leave room in our descriptions for the fact that this system could, at any given time, become something quite different" [36]. From these considerations we can derive the first and most basic design principle of enactive AI:

Enactive AI design principle 1 (EAI-1): the system must be capable of generating its own systemic identity at some level of description.

Note that this design principle could also be interpreted as an enactive approach to the value design principle of embodied AI (A-8). In contrast to the idea that AI systems require the addition of a 'value system' to provide them with motivations (e.g. [99]), the enactive approach to AI holds that design principle EAI-1 is *necessary* if the problem of intrinsic motivation is to be fully resolved. Only systems with a self-generated identity can be said to genuinely own and enact their own goals [61]. Furthermore, it is only when dealing with a system which meets design principle EAI-1 that we are justified in shifting our language from talking about the frame-of-reference *of* the system to saying that there is a perspective *for* the system (cf. embodied AI design principle P-5). This perspective for the autonomous system, understood as the enaction of a world that presents meaningful affordances according to the intrinsic goals of that system, can be put into context and extended by additional design principles, but must necessarily be grounded in the systemic requirement of constitutive autonomy (SR-1).

Design principle EAI-1 is qualified by the condition "at some level of description" because it is always possible to describe the same observed phenomenon in different ways. Depending on what kind of distinctions we make when studying an organism, for example, we will end up with different kinds of systems of relations, some of which will not be constitutively autonomous. In practice what kind of systemic organization is distinguished will depend on the level of abstraction used, the kind of components and relations identified and determined as noteworthy, as well as the purposes and general context of the activity. The reason this ambiguity is explicitly included in the design principle is that it is always possible that a system which looks non-autonomous from one perspective might actually turn out to be autonomous from another. We will illustrate this shift in perspective with some examples later.

One of the most important implications of design principle EAI-1 is the need to shift the design process from directly engineering an agent toward designing the appropriate conditions of emergence for an agent to self-constitute (e.g. [39]). We can call this the problem of *engineering second-order emergence*. Whereas embodied AI is faced with the challenge of designing an agent such that, when it is coupled with its environment, it gives rise to the desired behavior, here the target behavioral phenomenon is one step further removed. We need to start thinking about how to design an environment such that it gives rise to an agent which, when it is coupled with its environment, gives rise to the desired kind of behavior.

It is likely that such an investigation into the enabling conditions will require a more serious consideration of the kind of active dynamics which are intrinsic to the chemical components involved in natural metabolism [90]. Moreover, it is not possible to simply artificially *evolve* a non-autonomous system into an autonomous agent as "the Darwinian theory of evolution by natural selection does not provide any account of organization at the level of biological individuals. On the contrary, the theory must presuppose biologically organized individuals that reproduce" [118, p. 131]. Similarly, evolutionary algorithms presuppose that there are entities with identities to evaluate and select, and these need to be at least roughly specified by an external designer. While the dynamical/evolutionary approaches to embodied AI thus made advances in limiting the influence of the external designer on the artificial agent to be studied in order to reduce the impact of implicit presuppositions on the results, enactive AI further radicalizes this idea by demanding that the designer has no *direct* influence at all. Accordingly, since accommodating design principle EAI-1 appears to rule out the two most popular AI methodologies for engineering autonomous systems, namely (i) explicit design, and (ii) evolutionary robotics, we can call it the *hard problem* of enactive AI. We will discuss some examples of how to potentially resolve this problem later.

As a final observation, it is important to realize that there will be practical advantages of engaging in second-order engineering of emergence to solve the hard problem of enactive AI, especially in terms of robustness. It is well known that

systems which are characterized by homeostasis are less brittle because they keep certain essential variables within their regions of viability. However, since only these essential variables are buffered by the homeostatic mechanism, this leaves other parts of the system vulnerable to destructive perturbations, most importantly the homeostatic mechanism itself. In this respect it is interesting to note that autopoiesis is sometimes characterized as a kind of second-order homeostasis, that is, a homeostatic system which has its own organization as the essential variable (e.g. [81, p. 79]). Accordingly, in such systems the robustness afforded by homeostasis is applicable to the systemic identity as a whole, including the very homeostatic mechanism which makes this robustness possible in the first place. In other words, successful implementation of design principle EAI-1 has the potential to afford the system with what we will call *holistic robustness*.

This holistic robustness is enactive AI's answer to what Wheeler [139] has called the "intra-context frame problem": how can a system achieve appropriate, flexible and fluid action within a context? Design principle EAI-1 provides a strong basis for such behavior because the system, due to its constitutive autonomy, enacts a domain of interactions that is relevant for the system, its *Umwelt*, in relation to a viability constraint that is also brought forth by its own endogenous activity. Such a system never encounters meaningless facts about its situation which then need to be assigned context-dependent relevance; it brings forth its own context of significance in the same movement that generates its precarious identity.

Nevertheless, as we have argued in Section 3.2.2, the basic normativity afforded by constitutive autonomy needs to be put into a wider sense-making perspective, or otherwise the system's flexibility will be severely limited. This limitation brings us to the second enactive AI design principle (EAI-2).

4.2. Toward enactive AI (ii): adaptivity

With the first enactive AI design principle (EAI-1) we have motivated those projects in AI which are interested in generating models of systems about which we can say that their perspective and the goals which they enact are genuinely their own. As discussed at length in Section 3, the systemic requirement of constitutive autonomy (SR-1), on which this design principle is based, lies at the very core of the biological foundation of enactive cognitive science. However, in that section we have also argued for the need of extending this core principle with the systemic requirement of adaptivity (SR-2) in order to properly account for the sense-making which enacts an agent's *Umwelt*. This leads us to propose a second design principle for enactive AI:

> **Enactive AI design principle 2 (EAI-2):** the system must have the capacity to actively regulate its ongoing sensorimotor interaction in relation to a viability constraint.

The notion of viability constraint has been left explicitly vague such that design principle EAI-2 can be interpreted in two essentially different ways. On the one hand, if the viability constraint is given as externally imposed, whether explicitly by an external designer or implicitly through an evolutionary algorithm, then it provides the basis for studying the dynamics of adaptivity in isolation. In this manner EAI-2 can be used to investigate the role of certain dynamical phenomena, for example homeostasis and ultrastability, in organizing, maintaining, and regulating a system's closed sensorimotor loops [41]. On the other hand, if the viability constraint is interpreted as intrinsically related to the identity of the system, then this principle builds on EAI-1 and makes the hard problem of enactive AI even harder. Nevertheless, there is a clear motivation for addressing both EAI-1 and EAI-2 at the same time, as both are necessary for sense-making (cf. Section 3.2.3). Moreover, an artificial system that implements both would constitute a significant step toward better models of natural agency, a property which not only requires constitutive autonomy but also adaptive self-maintenance [90].

It is worth emphasizing again that this focus on adaptive constitutive autonomy should not be misunderstood as a mere reversal of a completely reactive system into its opposite extreme of absolute internal determination. Natural agents are not only thermodynamically open systems; they are also always open to interaction with their environment on which they depend for their existence [9]. As such their behavior, conceived as the coherent dynamics emerging from a brain-body-world systemic whole, is always caused by a multiplicity of internal and external factors. Thus, it might be more useful to view design principle EAI-2 as providing the artificial system with the capacity to change between different modes of dynamical engagement with its environment; for example, from committed ongoing coping to open susceptibility to external demands [39].

In practical terms we can say that design principle EAI-2 represents enactive AI's response to what Wheeler [139] has called the "inter-context frame problem": how can a system achieve appropriate, flexible and fluid action in worlds in which adaptation to new contexts is open-ended and in which the number of potential contexts is indeterminate? Indeed, AI systems which satisfy the systemic requirement of adaptivity (SR-2) are well equipped to deal with new contexts that are brought on through drastic changes in internal and/or external conditions. They have the general capacity to autonomously regulate their sensorimotor interactions so as to regain overall stability. However, since such systems do not make use of any internal representation or explicit model of their domain during adaptation, they seem to be vulnerable to the problems raised by Kirsh [75] long ago:

> [H]ow can a system whose response to failure is to try a simpler behavior achieve this innovative resilience? The reason this is a hard problem is that the response the system has to make varies across situations. The same behavioral failure

can require different fixes. That means that at a superficial level there is not enough information to determine an action. [75]

The first steps toward a theoretical foundation which can address these concerns were already proposed by Ashby [6] during the cybernetics era. His work on the origin of adaptive behavior in terms of random step functions can be seen as an early proof of concept that such behavior, which might look compellingly intelligent to an outside observer, does not necessarily require equally intelligent internal mechanisms [36]. The systemic requirement of adaptivity (SR-2) is another step in this direction.

It might be helpful at this point to briefly consider the manner in which current work in computationalist AI attempts to solve this kind of 'inter-context frame problem' by designing solutions that exhibit both robustness and innovative resilience. This is often accomplished by implementing another layer of cognition which controls the necessary changes to the system, namely a kind of cognition about cognition, or "metacognition" [29]. Two popular approaches in this regard are the design of "introspective multistrategy learning" systems (e.g. [31,30]) and cognitive architectures that incorporate a "metacognitive loop" [4,3]. Essentially, these metacognitive strategies are implemented by equipping systems with some extra computational modules which provide the capacity to: (i) monitor for perturbations, (ii) assess the perturbations, and (iii) adjust the system appropriately. It has been demonstrated that the addition of such metacognitive components can improve the system's perturbation tolerance under some conditions (e.g. [3]), though such positive results are not necessarily always the case [29].

This work on metacognition as a means to increase perturbation tolerance might appear to be an example of design principle EAI-2 because the systemic requirement of adaptivity (SR-2), on which that principle is based, also involves both "actively monitoring perturbations and compensating for their tendencies" [38]. Indeed, both are attempts to increase the robustness and behavioral flexibility of AI systems that are faced by dynamic and uncertain conditions (both internally and externally), namely through the use of dedicated regulatory mechanisms. However, there is still an essential methodological difference between the computationalist and the enactive approach to this problem that can be seen as characteristic of their respective design philosophies as a whole. The former approach tries to resolve the problem of brittleness by means of some additional top-down, symbolic control mechanisms, such as designing an oversight module that implements a metacognitive loop [4]. In this manner some robustness and flexibility is gained by deferring the problem of brittleness to the meta-domain. The latter approach, on the other hand, proposes to implement the required adaptivity by means of additional bottom-up, dynamic regulatory mechanisms, such as evolving artificial neurons that are characterized by homeostasis [35]. How we can design such adaptive systems will become clearer when we examine an example of such 'organismically-inspired robotics' in Section 4.4.

4.3. What is 'enactive AI'?

Before we start discussing some concrete examples of AI systems to determine in what manner they succeed or fail to satisfy design principles EAI-1 and EAI-2, a few words about the label 'enactive AI' are in order, so as to minimize potential confusion. Strictly speaking, it would be consistent with our arguments of Section 3 to say that only systems which somehow satisfy design principle EAI-1 should be classified as examples of enactive AI. It is essentially the systemic requirement of constitutive autonomy (SR-1) which distinguishes much of enactive cognitive science from the rest of the embodied cognitive sciences. However, things are not this simple; there is also a broader notion of enactive AI which overlaps to some extent with what we have been calling embodied AI. Of course, this is not very surprising considering that enactive cognitive science incorporates many insights of the dynamical and embodied approach and vice versa (cf. Section 2.5), but it is still worth making explicit.

Thus, there are evidently AI systems in the form of robotics and simulation models which do not satisfy EAI-1, but which are nevertheless still informative for the enactive approach. Some pertinent examples are investigations into the dynamics of minimal cognition (e.g. [13,57]), social interaction (e.g. [50]), vision (e.g. [114]), and adaptivity (e.g. [35]). It is in this sense that we can also understand why Varela, Thompson and Rosch [127] refer to Brooks' early work on behavior-based robotics (e.g. [20,21]) as an "example of what we are calling enactive AI" (p. 212). These robots are a concrete demonstration that a non-representational conception of cognition, which views the 'world as its own best model', is a viable position at least in some contexts. They are a powerful proof of concept that coherent behavior can emerge out of the distributed dynamics of a 'brain-body-world' systemic whole. In addition, there is also ongoing work in cognitive robotics and cognitive systems architectures that is strongly informed by enactive cognitive science (e.g. [107,131,92]).

All of these various research directions can be considered as examples of enactive AI in a broad sense, and they can certainly be useful in further explicating the theoretical foundations of enactive cognitive science. Nevertheless, in the following discussion we will highlight those pieces of work which can be interpreted as addressing the design principles EAI-1 and EAI-2 in the most direct manner. It is our hope that such a strict distinction has the benefit of drawing attention to existing work that explores the ways in which we might tackle the hard problem of enactive AI, as well as bring to light those areas of research which still need to be more fully explored, especially in terms of the challenge of engineering second-order emergence.

Another important point that needs clarification here is that we make use of the term 'AI' to denote any kind of scientific or philosophical investigation that is based on a synthetic methodology. This is nothing other than a commitment to the foundational design principle of 'understanding by building' (P-1), which often takes the form of designing simulation mod-

els and/or robots but is not limited to these approaches. We explicitly include the field of artificial life in our discussion, no matter whether the particular approach is based on computational models or actual/simplified chemistry. This is important because it might turn out that there are material and energetic requirements that are just as crucial for natural agency as the systemic requirements proposed by enactive cognitive science (cf. [106]). We will not discuss research in 'wet' artificial life in any detail here, because this approach deals with actual instantiations rather than models. Nevertheless, there is a good chance that advances in this area will provide us with a better understanding of how to address the hard problem of enactive AI in modeling terms.

This distinction between instantiations and models leads us to a final consideration, namely about the status of AI models. In enactive cognitive science it is still a contentious issue whether a computational (mathematical or simulation) *model* of a constitutively autonomous system can at the same time also *be* such a system (cf. [118, pp. 143–145]). In other words, the question is whether a particular example of 'weak' enactive AI would also entail a 'strong' enactive AI counterpart. Though opinions are divided on this matter, Varela [126] specifically speaks of his simulation work as a kind of proof of concept, and that is one important way in which we view the role of modeling here. In addition to the possibility of using models to evaluate the sufficiency or necessity of theories when it comes to accounting for data or other observations, a lot of modeling work can also be considered as simply exploratory in nature [91]. On this view, simulation models are a type of 'opaque' thought experiment in which the consequences necessarily follow from the premises provided by the experimenter, but which do so in a non-obvious manner, and as such must be revealed through an additional effort of systematic enquiry [40].

In summary, we are not interested in the 'strong AI' position of replicating natural phenomena; instead we treat simulation models as useful tools that can help us to better understand complex theoretical positions. Including work in robotics and other artificial systems in this consideration, we can thus say that we use the notion 'enactive AI' to denote any approach to creating artificial systems which stands in a mutually informative relationship to enactive cognitive science.

4.4. A typography of AI systems

We will now introduce a simple typography of AI systems in order to give some structure to our survey of current work in terms of the design principles of enactive AI. We have argued that the most crucial distinction between simple embodied and enactive cognitive science is to be found in an ontological difference of the systems under consideration: whereas the former can be characterized by a mode of existence that is essentially 'being by being', the latter are constitutively autonomous in the sense that their mode of existence is 'being by doing'. We will refer to these two kinds of systems as Type I and Type II, respectively. Further subcategories of these two types will be introduced as well.

4.4.1. The typography of Type I systems

The overall typography of Type I systems is illustrated in Fig. 2.

How do current AI systems fall into these three different categories and how do they relate to the enactive approach? Let us begin with a consideration of Type I systems. This ontological class is characterized by a mode of 'being by being' such that they essentially exist in a non-precarious manner. In other words, their existence is one of indifferent permanence. We define Type Ia systems as follows:

Type Ia: non-precarious systems with minimal forms of interaction

Note that the way these systems are represented in Fig. 2, namely as exhibiting no interaction, should only be understood as indicating an ideal minimal point that is never realized in any actual work. Of course, all AI systems are open to some form of interaction, no matter how minimal that interaction turns out to be in practice. The kind of systems which fall into this category are therefore those for which interaction is not an essential component of their operational activity, and/or which do most of their work in an 'offline' mode. Accordingly, much of the early work in computationalist AI can be classed as Type Ia systems. Expert systems are a good example of this approach because they explicitly attempt to minimize the

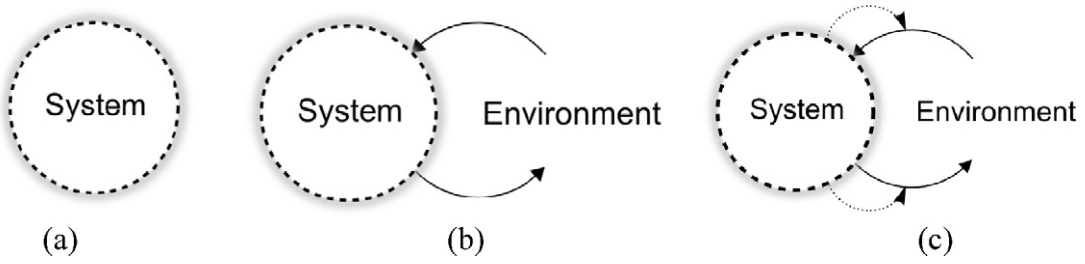

Fig. 2. The typography of Type I systems: their essential mode of existence is one of 'being by being', represented by the closed circle. They can be further categorized as follows: (a) systems with minimal forms of interaction, (b) systems which are embedded in sensorimotor interactions with their environment, and (c) systems which have the further capacity of adaptively regulating their sensorimotor interaction according to a viability constraint.

required interaction with the end user. More generally, we include any input/output device which only exhibits a linear interaction with its environment. This is the case with many examples in traditional connectionism [27], but by and large also includes any AI approach which focuses on designing independent modules that instantiate a simple input/output mapping. Type Ia systems have certainly proved useful for a variety of practical applications but, from the point of view of enactive cognitive science, they do not contribute much to our understanding of natural cognitive phenomena and as such they will not be further considered here.

Type Ib: non-precarious systems which are embedded in sensorimotor interactions

The move toward Type Ib systems generally coincides with a transition from computationalist AI to embodied AI for which an emphasis on sensorimotor embeddedness is a defining feature. Notable precursors to this work are W. Grey Walter's [135] mechanical 'tortoises', which displayed a variety of interesting behaviors grounded in sensorimotor interactions. The first mobile GOFAI robots, such as Stanford's Shakey and early research in behavior-based robotics (cf. [5]) also fits into this category to the extent that appropriate sensorimotor interaction has become a more important design consideration. However, work on Type Ib systems really took off with Brooks' emphasis on embodiment and situatedness in robotics in the late 1980s (e.g. [21]), an approach which has continued to be further developed into embodied AI (e.g. [100]). This category thus encompasses a variety of recent AI research programs including, for example, evolutionary robotics (e.g. [57,96]), epigenetic and developmental robotics (e.g. [15,79]), and the dynamical approach to cognition (e.g. [11,13]).

From the point of view of enactive cognitive science, Type Ib systems have the potential to inform our understanding of natural cognition and as such we could consider them to constitute an enactive approach to AI in a broad sense. This is even more the case when an explicit effort is made to reduce the designer's impact on the behavior and organization of the system under study, such as is done in most evolutionary and dynamical approaches. Moreover, as proposed by Beer [14], a dynamical perspective on behavior and cognition follows directly from an autopoietic perspective on life when two key abstractions are made: (i) we focus on an agent's *behavioral dynamics*, and (ii) we abstract the set of destructive perturbations that an agent can undergo as a *viability constraint* on its behavioral dynamics.

Since these abstractions basically underline the relevance of the dynamical approach to AI for an enactive theory of cognition grounded in constitutive autonomy, it is worth spelling out Beer's argument in more detail. He begins with the observation that a natural agent's normal behavior takes place within a highly structured subset of its total domain of interaction. This makes it possible to capture the behavioral dynamics while ignoring other structural details which may not be directly relevant.[15] Moreover, since it is only meaningful to study an agent's behavior while it is living, we can largely take an agent's ongoing constitutive autonomy for granted in our models. Finally, the possibility of undergoing a lethal interaction is represented as a viability constraint on the agent's behavior such that if any actions are ever taken that carry the agent into this terminal state, no further behavior is possible [14].

It follows from these considerations that research with AI systems of Type Ib has the potential to develop a mutually informing relationship with some of the theoretical foundations of enactive cognitive science. However, since the insights generated by this research are equally informative for other cognitive science approaches, which might also be embodied, embedded and dynamical but not grounded in constitutive autonomy (e.g. [138,26]), this is strictly speaking still not fully enactive AI according to the criteria we outlined previously. Indeed, due to its necessary abstractions such AI research cannot aid our understanding of how natural cognition arises through the self-constituted activity of biological systems, and as such it is unable to address the criticisms which have recently been leveled against the work of embodied AI by Dreyfus and others (cf. Section 2).

Type Ic: non-precarious systems which can adaptively regulate their sensorimotor interaction

Of course, by definition, none of the AI systems falling into the Type I categories fulfill the constitutive autonomy design principle (EAI-1). Nevertheless, there has been some recent work on Type 1c systems that manages to address a 'weak' version of design principle EAI-2, namely by introducing adaptive regulation in relation to externally imposed (rather than internally generated) viability constraints.

Most of current embodied AI has focused on the problem of how the internal organization of a system gives rise to its external behavior during structural coupling with its environment [98]. However, the development of enactive AI that satisfies design principle EAI-2 requires that the inward arm of the causal loop, namely how external or internal events can influence the organization of the system, must also be included in the system design [36]. As Vernon and Furlong [130] point out, one of the main challenges facing the development of this kind of enactive AI is the "need for new physical platforms that offer a rich repertoire of perception-action couplings and a morphology that can be altered as a consequence of the system's own dynamics". Only when this possibility of internal re-organization is explicitly included in the design process of artificial systems is there a potential for the system to satisfy the systemic requirement of adaptivity (SR-2) in the face of internal or external perturbations. The implementation of design principle EAI-2 thereby provides the system with a

[15] It is also possible to defend the abstraction of the behavioral domain as a domain of dynamics in biological terms, namely by arguing that the nervous system in animals is hierarchically decoupled from the enabling metabolic processes (cf. [8]).

possibility of 'becoming' rather just inert being, as suggested by Ziemke [141]. Accordingly, Type Ic systems thus have the potential to represent a significant advance toward modeling the organismic mode of existence.

In general, it should be noted that all the models which will now be discussed as specific examples of Type Ic systems have been generated by some variation of an evolutionary robotics and dynamical approach to AI. While this is not a necessary requirement for this category, it does confer certain advantages, for example in terms of the explanatory power of the evolved solutions, as discussed in the case of Type Ib systems. Moreover, the increased complexity of these models, already a problem for most of the advanced cases of Type Ib systems, makes it prohibitively difficult to design them by hand. In those cases where the generative mechanism deviates in a relevant manner from this standard approach, we will highlight any important aspects of the model agents which have been explicitly introduced as extensions by an external designer.

Let us first discuss two instances of Type Ib systems in which a special component was introduced into the dynamical controller in order to enable the system to regulate its sensorimotor interaction. In these cases the additional component was evolved to mediate between the internal dynamics of the system and its sensorimotor interface. As an initial step we can consider the work of Macinnes and Di Paolo [80]. They draw inspiration from the biology of von Uexküll [133] in order to implement a simple pole balancing robot model that illustrates the advantages of making the sensorimotor interface adaptable by artificial evolution. Their work demonstrated that adding one or more genetically determined 'offsets' to the angle sensor can significantly increase the evolvability of the solutions. However, even though this model nicely illustrates the advantage of increasing sensorimotor adaptability, it nevertheless falls short of being a proper example of sensorimotor regulation or adaptivity. Since the sensory 'offsets' are genetically determined by artificial evolution, it follows that they are not adjustable by the internal dynamics of the system and thus cannot be used to modulate the system's sensorimotor interaction in relation to environmental conditions. Additional work would be required to show that this additional controller structure gives rise to internal dynamics that could be described as regulatory.

Iizuka and Ikegami [66] developed a model that more explicitly focuses on the regulation of sensorimotor interaction. In this case a simulated mobile robot was evolved to actively discriminate between light sources of different frequencies by engaging in approaching or avoiding behavior. They introduced a special 'gate' node into the artificial neural network controller which determined the coupling between the internal dynamics and sensory input.[16] In this manner the internal dynamics of the system were capable of switching sensory perturbations on or off and, at the same time, this regulation of sensorimotor interaction affected the internal dynamics. They showed that the system made use of this circular relationship in order to establish the behavioral coherence that was required to perform successfully during the ongoing discrimination task. Here we have the beginnings of what we might call a form of self-determined sensorimotor regulation because the activity of the 'gate' node was dependent on internal dynamics as well as on external conditions (cf. [145]).

While these two examples explicitly introduced elements into the controller which mediate between the system's internal dynamics and its sensory perturbation, this is not a necessary requirement for sensorimotor regulation. Recent work by Izquierdo and Harvey [70], for example, has demonstrated that a model mobile robot controlled by a simple continuous-time recurrent neural network (cf. [11]) with fixed weights and *without* synaptic plasticity mechanisms can be evolved to change its sensorimotor behavior according to environmental circumstances. The task of the simulated agent is to find 'food' at a location that is placed either up or down a temperature gradient. The best evolved solution can not only learn to associate the appropriate temperature gradient with the target location, thereby improving its foraging behavior on successive trials, but even subsequently modify this association when environmental conditions change. An analysis of the dynamic controller revealed that the system's ability to regulate its sensorimotor correlations in this manner had largely to do with dynamics on multiple timescales. In particular, the model neuron with the largest time constant appeared to have evolved to keep track of the type of environment in which the agent is currently located.

The work by Iizuka and Ikegami [66] and Izquierdo and Harvey [70] can be considered as examples which begin to illustrate design principle EAI-2. They are artificial systems with the capacity to regulate their ongoing sensorimotor interaction in relation to (external) viability constraints. However, they are still relatively limited because the simulated agents were explicitly selected for their ability to regulate their sensorimotor interaction during artificial evolution. In other words, during the testing phase they did not have to engage in adaptive regulation in the face of extreme sensorimotor perturbations that the solutions had not previously been encountered in their evolutionary history.

However, that such adaptation to more radical sensorimotor distortions is possible in biological systems has been demonstrated extensively for a variety of organisms. A particularly famous example is the eventual adaptation of human subjects to the wearing of visual distortion goggles (e.g. [76]). What makes these cases especially interesting is that they are difficult to explain in terms of specific evolutionary pressures, and as such they are probably the result of more general principles of organismic operation [35]. Taylor [115] proposed a promising framework for interpreting this kind of sensorimotor adaptation by incorporating insights from the field of cybernetics in the form of Ashby's [7] work on homeostasis and ultrastability in living systems. According to this view, eventual adaptation comes about because the stability of different internal systems of the organism are challenged by the sensorimotor disruption, and these systems regulate the internal operations of the system until it adjusts to the disruption in such a way that stability is regained. Such radical sensorimotor adaptation is

[16] For a similar modulatory mechanism see also [149].

therefore a good example of the systemic requirement of adaptivity (SR-2), and provides the right kind of inspiration for addressing design principle EAI-2.

Di Paolo and colleagues have implemented a variety of simulation models that have been inspired by this combination of ultrastability and radical sensorimotor adaptation, an approach which has been called 'organismically-inspired robotics' [36]. The initial model by Di Paolo [35] consisted of simple simulated mobile robot which was evolved to perform phototaxis on a series of light sources that were randomly placed in the environment. A light source would appear, the robot is expected to approach it, and eventually the light disappears to be replaced by another one somewhere in the distance, etc. until the trial ends. What makes this model interesting is that a homeostatic mechanism was introduced into the dynamical controller by evolving it such that on average each neuron will maintain a firing rate that is neither too high nor too low. Whenever a neuron fires outside this range of (external) viability, a genetically specified rule of synaptic change is activated that adjusts the weights of the incoming connections. The robot is thus evolved both to perform phototaxis as well as to keep the neuronal firing rates in their regions of stability.

The evolved solutions were then tested according to whether they were able to adapt to visual inversion, which was induced by swapping the left and right light sensors. It was found that some robots exhibited eventual adaptation and regained their ability to perform phototaxis. In these cases the internal adaptive regulation of sensorimotor interaction occurs because the visual inversion initially drives the robot to avoid the light, which pushes the firing rates of the neurons outside of their viability range. Consequently, this will result in plastic changes to the connection weights until their viability is regained. And finally, because the evolutionary algorithm has happened to intertwine internal stability with the conditions also favoring behavioral stability, the robot successfully resumes its phototaxis. This regulatory mechanism is very flexible because the same kind of adaptation was also observed when inducing a variety of other sensorimotor disruptions that the robot had never encountered before.

The use of this kind of homeostatic neurons provides a novel approach to modeling adaptivity which has been used to investigate a variety of topics, such as the formation of habits in terms of identity generation [36], the minimal dynamics of behavioral preference [67], preservative reaching (the A-not-B error) in young infants [140], and the autonomous alternation between behavioral modes of ongoing coping and open susceptibility [39]. These examples of modeling homeostatic mechanisms and adaptivity in artificial systems are highly compatible with design principle EAI-2.

Nevertheless, it also needs to be emphasized that in all of these cases the viability constraint is still externally imposed on the system. As a consequence the link between internal stability and behavioral competence is largely arbitrary: it is always possible that a dynamical controller evolves for which the neurons can regain viable firing rates independently of appropriate changes in the agent's overall behavior. This is a problem of the approach which can be improved upon at least to some extent (e.g. [68]), but which can never be fully resolved. Like all Type I systems, the internal organization of these robot models does not satisfy the systemic requirement of constitutive autonomy (SR-1), which is necessary for the generation of intrinsic viability constraints.

However, this certainly does not mean that these models cannot be informative for the study of autonomous behavior. On the contrary, similar to the arguments for the dynamical approach to AI presented earlier, there are biological considerations which can justify the necessary abstraction of the adaptive regulatory mechanism from its metabolic underpinnings to some extent. Barandiaran and Moreno [9] argue that there must be a relative decoupling between the dynamics of the regulatory subsystem and those of its basic constitutively autonomous organization. This is because adaptivity, as the regulatory capacity to distinguish and compensate deleterious tendencies, must be dynamically differentiated from what it distinguishes and acts upon. In other words, it requires a dedicated mechanism that is not directly involved in metabolic self-production.

In summary, we have argued that Type Ia systems, while often useful as technological artifacts, cannot generally contribute to an understanding of the kind of phenomena that are investigated by enactive cognitive science. In the case of Type Ib systems we highlighted the relevance of the evolutionary robotics and dynamical approach to AI for its potential to explore the dynamics of minimal cognition. Finally, we argued that Type Ic systems have the potential to deepen our understanding of the systemic requirement of adaptivity (SR-2), at least in its 'weak' form. In particular, the organismically-inspired robotics approach currently represents the most promising route to incorporate design principle EAI-2 into our AI systems.

While none of the Type I systems satisfy design principle EAI-1 it should again be emphasized that the enactive approach, which started with Maturana and Varela's [81] theory of the autonomous organization of the living, is in some cases an important inspiration for the development of these systems (e.g. [36,14,11,107,92]). However, while the insights generated by such research with Type I systems can thus be informative for enactive cognitive science, they are also of similar (if not more) interest to the more general embodied and dynamical approaches to cognitive science (e.g. [138]). This general applicability of the models is not a problem from the enactive point of view, but only as long as the various abstractions, which are necessary to isolate the cognitive domain in dynamical form, are not ignored. In those cases where authors are compelled by their models to reduce living being to some kind of sensorimotor embodiment, they become targets of the philosophical criticisms developed in Section 2 of this paper. The generative methods used to engineer Type I systems can only generate essentially partial models of natural agency.

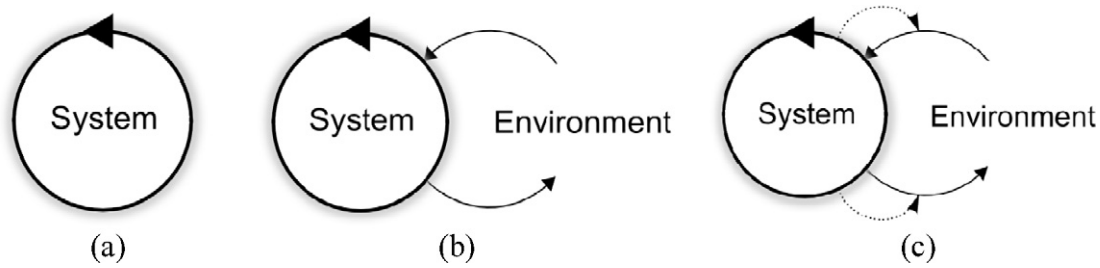

Fig. 3. The typography of Type II systems: their essential mode of existence is one of 'being by doing', represented by the closed arrow circle. They can be further categorized as follows: (a) systems with minimal forms of interaction, (b) systems which are embedded in sensorimotor interactions with their environment, and (c) systems which have the further capacity of adaptively regulating their sensorimotor interaction according to a viability constraint. Note that in all of these cases the domain of interactions, regulatory behavior and viability constraint is brought forth by the activity of the system itself.

4.4.2. The typography of Type II systems

We will now consider some examples of AI systems that manage to satisfy the systemic requirement of constitutive autonomy (SR-1). This area has received much less attention in the AI research community compared to the development of Type I systems, but this will have to change if we want produce better models of natural agency. The subcategories of Type II systems are illustrated in Fig. 3.

This ontological class is characterized by a mode of 'being by doing' such that they essentially exist in a precarious manner. In other words, their existence is one of concerned autonomy. We define Type IIa systems as follows:

Type IIa: self-constituting systems with minimal forms of interaction

Before we start the discussion of some concrete examples of Type IIa systems, it should again be noted that the way these systems are represented in Fig. 3a should only be understood as an ideal reference point. In actuality there are no constitutively autonomous systems that do not also necessarily engage in at least some form of minimal interaction [16]. This interactive component is the basis for Jonas' [71] existential notion of the 'needful freedom' which characterizes all living beings, and is essential for their precarious situation. It has recently also been the topic of more detailed biological considerations by Barandiaran and Moreno [9]. They criticize the notion of constitutive autonomy that has been employed in the biological tradition started by Maturana and Varela [81] for not placing enough emphasis on the importance of interaction. Instead, they argue that any minimal autonomous organization must always be characterized by two kinds of constitutive processes, namely both constructive and interactive processes. We agree with these considerations. However, it is still useful to distinguish between systems of Type IIa and Type IIb because it is possible to design models of minimal constitutive autonomy whose interactive aspects are simplified to a bare minimum.

As a starting point for this discussion of Type IIa systems it will be helpful to carefully examine an example of what might be considered embodied AI's most promising approach in this regard, namely the design of energetically autonomous systems that appear to have some sort of 'metabolism'. Melhuish and colleagues [87] report on a robot called EcoBot-II, which is designed to power itself solely by converting unrefined biomass into useful energy using on-board microbial fuel cells with oxygen cathodes. Let us imagine for a moment that this robot could successfully gather enough biomass such that its fuel cells would provide it with sufficient energy to support its ongoing foraging activity. Is such a robot metabolic in the sense that living beings are? Or are its operations perhaps essentially the same as those of a 'normal' robot that can recharge its energy levels by finding an electrical socket in the wall? In this normal case the lack of energy is clearly not a viability constraint that is intrinsic to the system itself; without battery power the robot simply rests its operations until someone recharges it.

The EcoBot-II, on the other hand, does appear to be in a slightly more precarious situation due to the precariousness of the living microbes on which it depends for its energy supply. Nonetheless, it is still the case that the viability constraint to maintain a stable energy level by finding appropriate amounts of biomass is something which is imposed on the system externally. In actual metabolic systems such a constraint on the conditions for its existence is necessarily constituted by the activity of the system itself in continually bringing forth its own metabolic organization. As such we agree with Melhuish and colleagues who suggest that the EcoBot-II robot is an exemplar of a microbe-robot 'symbiosis', rather than an instance of metabolic self-production. However, before we move on to examine a model which explicitly attempts to satisfy the systemic requirement for constitutive autonomy (SR-1), there is one important general lesson which can be learned from the EcoBot-II project.

In terms of their long-term goal for this work Melhuish and colleagues propose two conditions that can be seen as the defining properties of energetically autonomous robots: (i) the robot must recharge energy only through interaction with its environment, and (ii) it must achieve tasks which require, in total, more energy than is provided at the start of the mission. Interestingly, this long-term goal leads them to a consideration of Kaufmann's [74] definition of an autonomous agent as a self-producing system that is able to perform at least one thermodynamic work cycle. It is of course debatable to what extent these two definitions overlap, and Melhuish and colleagues are quick to exclude the "burden of self-reproduction" from

their target: of course, the robot itself would not be self-producing. Nevertheless, they might have intuited an important aspect of such an energetically autonomous robot's operations which we can more fully uncover through a slight shift in perspective. After all, there is indeed a circular relationship between stable foraging behavior and stable energy production. More precisely, both of these levels of processes require each other for their existence: without foraging there cannot be any microbes, and without microbes there cannot be any foraging. The two processes thereby form a coherent unity which is necessarily precarious since its existence depends not only on internal energy stability but also on the encounter of certain favorable environmental conditions which can only be successfully managed through appropriate interaction.

Thus, if Melhuish and colleagues achieved their goal of implementing such a properly energetically autonomous robot, it would not be the robot which is constitutively autonomous, but rather the particular form of the dynamics which it instantiates by being appropriately coupled to its environment. On this view, the robot-environment systemic whole could be considered as providing an appropriate dynamic substrate for the emergence of an autonomous identity in the form of a self-constituting network of processes under precarious conditions. From an engineering perspective this would be a desirable outcome because the existence of this self-constituted identity requires robust and flexible operation of the robot in its task domain. Indeed, by viewing the network of processes as an autonomous system it becomes possible to design the robot-environment systemic whole with this goal in mind. For example, the processes of energy production and foraging behavior could be coupled in a manner that allows for richer dynamics to emerge, i.e. by enabling the status of the microbes to somehow affect the robot's controller and vice versa. As such, we could consider this to be an example of the problem of engineering second-order emergence.

This little thought experiment illustrates that we must become more aware of how the distinctions we make as external observers carve up the phenomena which we want to study. Even in the case of other Type I systems we might find that a similar change in perspective has the potential to reveal how autonomous identities can emerge at the level of the combined internal and sensorimotor dynamics [39]. Similarly, it might be possible that a system with constitutive autonomy could emerge in the domain of behavioral dynamics between several Type I systems, for example as a result of co-dependent social interaction (e.g. [50]). Finally, if this change in perspective indeed turns out to be feasible with regard to such AI systems, then it could also provide us with a theoretical background from which to gain a novel understanding of the behavior of actual living systems, too. Perhaps there exist other means of autonomous identity generation which have so far eluded us because we have not yet distinguished them appropriately.

Unfortunately, however, at the moment we neither have the tools nor the theory to determine the necessary conditions for an autonomous identity to constitute itself in the relational dynamics of these systems. This is because we do not even yet know how to identify the existence of an autonomous system in such an abstract dynamical domain [52]. At least in the case of the molecular domain some rough guidelines have already been developed for detecting the presence of autopoiesis (e.g. [118, p. 103], [128]), but searching for such a self-constituting identity in an abstract dynamical domain is significantly more complicated. First of all, there is no intuitive manner of distinguishing a particular subsystem of the dynamical system under study as a potential autonomous system, whereas in the case of autopoiesis the existence of a semi-permeable boundary is at least a helpful indicator. Moreover, we are faced by fundamental methodological problems: what, for example, constitutes a *process* in a domain of dynamics when abstracted away from actual thermodynamics? However, the fact that these questions are being raised indicates that the modeling approach is helpful in forcing us to be more precise in our theoretical formulations. It thus represents an important example of how a serious attempt at developing fully enactive AI is in the best interest of putting enactive cognitive science on more concrete foundations. Indeed, the possibility of constitutive autonomy in domains other than a chemical substrate is an area that deserves much closer study.

Let us now introduce some modeling work which specifically attempts to address the autonomous organization of the living. Since the study of this organization has been one of the main goals of the field of artificial life right from its beginning [19], it should come as no surprise that this field currently offers the most viable methodologies in this regard. However, for most artificial life researchers the need to find an appropriate way to address the systemic requirement of constitutive autonomy (SR-1), and the problem of engineering second-order emergence which it entails, has not been on top of the research agenda as the field has explored a wide range of topics. Nevertheless, there appears to be a resurgence of interest in the 'hard' problem of autonomy as evidenced by two recent special journal issues devoted to the topic [10,37]. We anticipate this renewed effort to further expand in the future as the enactive approach continues to develop and establish itself in the cognitive sciences, especially since it brings with it a greater need to understand the biological basis of natural agency.

In the field of artificial life there are at least four promising approaches to the problem of constitutive autonomy which are worth mentioning here: (i) the use of cellular automata in the tradition of computational autopoiesis (e.g. [128,86,69]), (ii) mathematical modeling of autopoiesis (e.g. [18]), (iii) artificial simulations of chemistry (e.g. [83]), and (iv) actual wet chemistry (e.g. [78,16]). It is worth emphasizing that all of these approaches focus on constitutive autonomy in the molecular domain. This makes sense because minimal metabolic systems are both ontologically and historically the basis for natural agency, and constitutive autonomy is currently best understood in terms of chemistry (e.g. [90,106]).

Since it is beyond the scope of this paper to review all of these different approaches in terms of how they address the design principle EAI-1, we will focus on one model which exemplifies a Type IIa system particularly well. Bourgine and Stewart [18] present a mathematical model of a 3D tessellation automaton which they consider to be a minimal example of autopoiesis (i.e. the systemic requirement of constitutive autonomy (SR-1) is satisfied in the chemical domain). In essence, the tessellation automaton consists of a spherical, semi-permeable membrane enclosing an internal volume and is charac-

terized by autocatalysis based on the interaction between four components A–D. The substrate A exists in the environment and freely diffuses across the membrane. The membrane is formed of C components which eventually disintegrate into D components at a particular rate-constant (i.e. $C \rightarrow D$), and the Ds diffuse into environment leaving a hole in the membrane. Inside the volume enclosed by the membrane there is the possibility of a formation of B components by a reaction between two A components (i.e. $A + A \rightarrow B$), which is catalyzed by the C components of the interior surface of the membrane. The intact membrane is impermeable to the B components which thus accumulate in the internal volume up to a particular upper limit on their concentration. If a B component collides with a hole in the membrane, it attaches to the surface and repairs the hole (fully if the hole was of size $1C$) by turning into a C component. If the hole is too big there is a chance that the B component can escape into the environment without attaching to the membrane.

Since these processes are (i) related as a network so that they recursively depend on each other in the generation and realization of the processes themselves, and (ii) constitute the system as a unity recognizable in the space (domain) in which the processes exist, we can consider the tessellation automaton to be a model that satisfies design principle EAI-1. While the automaton appears to be too simple to be of any practical use, it has intentionally been designed in this manner by Bourgine and Stewart in order to investigate several theoretical hypotheses, some of which we have discussed in Section 3.2.3. Finally, even though this tessellation automaton lacks adaptive mechanisms and only entails a highly restricted domain of interactions, it is still useful as a minimal model of life. The bacterium *Buchnera aphidicola*, for example, similarly obtains the necessary energetic and chemical input for self-maintenance from its medium with almost no change of environmental conditions, and thus has only highly degenerated interactive and adaptive capabilities [9].

Type IIb: self-constituting systems which are embedded in sensorimotor interactions

Type IIb systems satisfy design principle EAI-1 and build on this by constituting a broader domain of interactions. The most promising line of research in this regard can again be found in the field of artificial life, though there are not many actual examples.

Ikegami and Suzuki [69] describe an experiment in which the 2D cellular automata model of autopoiesis pioneered by Varela, Maturana and Uribe [128] is extended so as to enable the autopoietic unity to move due to the ongoing fluctuations of its boundary. The basic setup of the model is similar in style to the one by Bourgine and Stewart [18] so we will not repeat the details here, but focus on two crucial differences. In contrast to Bourgine and Stewart, who chose to use an abstract mathematical model to capture the cellular automata dynamics, Ikegami and Suzuki retain the detailed simulation approach to cellular automata as it was used in the original model by Varela, Maturana and Uribe. While the former approach is useful because it makes mathematical analyses more tractable and simplifies the transition to a 3D model, the latter approach has the advantage that it captures the exact position of each of the particles.

This spatiality of the simulation model enables Ikegami and Suzuki to introduce three additional rules to Varela, Maturana and Uribe's setup: (i) a single link element can become inserted into a neighboring membrane boundary if that boundary is composed of link elements in its Neumann neighborhood, (ii) a single link element can become removed from the membrane boundary if that boundary is composed of link elements that form a right angle (a link that is released in this manner decays into two substrate particles), and (iii) each of the bonded link elements can move randomly to empty neighboring spaces without breaking the membrane link. The likelihood of each of these rules being applied is determined by the given rates of insertion, removal, and movement, respectively.

The general result of these simple extensions is that the membrane boundary of the autopoietic cell begins to move. If such a cell is placed in an environment that contains homogeneous concentrations, then the cell as a whole will begin to show random Brownian motion. More importantly, Ikegami and Suzuki demonstrate that the dynamics of the membrane boundary are such that they also enable the autopoietic cell to follow an increasing chemical gradient of substrate particles. This chemotaxis can be accounted for by the fact that the membrane is more often broken and repaired (by linking the broken boundary edges with a link element from the inside of the cell) on that side of the cell which faces away from the substrate gradient. The speed of gradient following is also correlated with the structure of the autopoietic cell; circular forms are more efficient than rugged or spiky forms.

While such behavior might sound far removed from what is traditionally referred to as sensing and acting, it is worth noting that it nevertheless comes close to Bitbol and Luisi's [16] proposal that extended homeostasis should be considered as a special variety of sensorimotor action. Indeed, the capacity to switch between two modes of searching behavior according to environmental circumstances, namely random movement while there are no environmental cues and directed chemotaxis when a substrate gradient is present, can be described as an instance of minimal cognition. As such, we consider this autopoietic cell model to be a good illustration of how goal-directed behavior can emerge from the ongoing process of metabolic self-construction. Furthermore, since it was a change in the conditions of emergence which resulted in this qualitative change in behavior (in contrast to direct external design), this example is also a minimal model of intrinsic teleology, i.e. a simple investigation of how a system's goals can be genuinely its own. And finally, Ikegami and Suzuki's work provides us with another nice demonstration that certain behavior, which might be described as intelligent by an external observer, does not necessarily have to result from similarly intelligent internal mechanisms.

As a simulation of natural agency it is possible to compare the behavior of Ikegami and Suzuki's autopoietic cell model with that of the bacterium *E. coli*, which also engages in random tumbling and directed chemotaxis in order to locate environments with higher concentrations of metabolites. In the case of the *E. coli* bacterium, however, the searching behavior

results from dedicated internal mechanisms which correlate specific 'sensor' and 'effector' surfaces in an appropriate manner. This decoupling between the mechanisms that support interactive and adaptive tasks from the mechanisms directly involved in metabolic self-construction is an important pre-requisite to achieve increasingly complex forms of agency [9]. An interesting next step in the development of Type IIb systems would therefore be a simulation model which satisfies the conditions of emergence for an autopoietic cell that exhibits some internal functional differentiation. Indeed, this is likely to be a necessary step in generating more complex behavior in terms of the problem of engineer second-order emergence.

Type IIc: self-constituting systems which can adaptively regulate their sensorimotor interaction

For a system to be classed as Type IIc it would have to satisfy both of the enactive design principles EAI-1 and EAI-2. It needs to solve the 'hard' problem of constitutive autonomy, and some of the viability constraints that require adaptive regulation must be internally generated. In other words, a Type IIc system must bring forth its sensors, effectors, and their internal organizational link (some adaptive mechanism) on the basis of its self-constituting operations. So far, no one has been able to artificially generate such a system.

One possible starting point might be an attempt to extend Ikegami and Suzuki's autopoietic cell model. For example, in that model it was possible to show that variations in the removal rate led to variations in structural configurations which entailed variations in sensitivity to the gradient of substrate particles (i.e. higher removal rate → more circular membrane structure → faster gradient following). Would it be possible to somehow let this removal rate be modulated or constrained by the operations of the autopoietic cell itself? This would give it the capacity to regulate its gradient following behavior by adjusting the shape and size of its membrane structure. In what circumstances would such regulation be a requirement for adaptive behavior such that it could emerge on its own terms?

These questions make it clear that a different way of thinking is indeed required to move forward on the problem of second-order emergence. We need to research how to engineer the conditions of emergence for the self-constitution of an agent that exhibits the desired behavior. Moreover, we can see from these considerations that Brooks' [21] original proposal for AI, namely that we should first use the synthetic methodology to understand insect-level intelligence, might have already been too ambitious from the perspective of enactive cognitive science. In order to develop a better theory of the biological roots of intentional agency we first need to gain a better understanding of bacterium-level intelligence. Only by returning to the beginnings of life itself do we stand any chance of establishing a properly grounded theory of intentional agency and cognition.

In summary, it is certainly the case that the development of fully enactive AI poses a significant challenge to current synthetic methodologies. Although there are already a number of notable efforts in this direction, we expect that the field will have to engage in a lot more exploratory work before it is in a position to devise and exploit more rigorous procedures. It is our hope that the design principles for enactive AI and the typography of AI systems which we have presented here will provide a useful point of reference for this endeavor.

5. Conclusion

This paper has unfolded in three main stages going from embodied AI to enactive cognitive science to enactive AI. In Section 2 it was argued that the sensorimotor approach of embodied AI is necessary but insufficient for the modeling of important aspects of natural cognition, in particular the constitution of a meaningful perspective on a world affording purposeful action for an intentional agent. This led to the suggestion that the ongoing development of enactive cognitive science could provide an appropriate theoretical framework from which to address the perceived shortcomings of embodied AI. Accordingly, in Section 3 the biological foundations of enactive cognitive science were reviewed in more detail. The minimal organismic basis of intentional agency was summarized in the form of two systemic requirements, namely constitutive autonomy and adaptivity, which enabled us to properly ground our discourse on purposeful and goal-directed behavior (intrinsic teleology), as well as the enaction of meaningful perspective for an agent (sense-making).

Finally, in Section 4 these systemic requirements were used as the theoretical basis for deriving two design principles for the development of an enactive AI that can generate better models of natural agency. A classification of different types of AI systems was introduced, and some selected examples were analyzed in detail according to how well they manage to satisfy the design principles. This analysis also indicated certain areas which are in need of significant further development if the project of fully enactive AI is to become established as a successful methodology.

If successful, the insights generated by such a research program would be important in several respects: (i) in *practical* terms we could look forward to engineering artificial systems that can interact with their target niches as flexibly and robustly as living systems do with theirs [110], (ii) in *scientific* terms the field of AI could again resume its place at the very forefront of developments in cognitive science by explicating the systemic foundations of the enactive approach [49], (iii) in *philosophical* terms we could gain a better understanding of how our bodies matter to our autonomy and what makes our goals genuinely our own [61], and finally (iv) in *ethical* terms these insights into autonomy, sense-making, goals, purposes, agency, etc., are practically important because science affects the way in which we understand ourselves, and this includes the way in which we believe it is possible to behave [17].

Acknowledgements

Tom Froese wishes to thank Ezequiel Di Paolo and the participants of the Life and Mind seminars at the University of Sussex for their many helpful discussions. We would also like to give thanks for the kind support of everyone at the University of Skövde, where much of this paper was written, and to Joel Parthemore, Rob Lowe as well as three anonymous reviewers for providing detailed comments on an earlier draft of this paper. This work was financially supported in part by European FP6 cognitive-systems projects *euCognition* (www.eucognition.org), which funded Tom Froese's stay at the University of Skövde, and *ICEA* (www.iceaproject.eu).

Appendix A

Those readers, who have a theoretical background in cognitivism or, more generally, computationalism, might find our characterization of their preferred paradigm overly critical and negative. And, indeed, we are distancing ourselves from those traditions, but only in so far as their intellectualist interpretations of mind and cognition have become over-generalized. We believe that enactivism does not necessarily negate or replace the traditional mainstream of cognitive science and AI; instead it delimits its proper domain of applicability and at the same time provides that domain with a foundation. The burden is on enactivism to demonstrate that objective thought, and rationality more generally, emerges as an irreducible domain from the 'cognitive' operations of organismic life:

> When we widen our perspective to include such forms of cognitive behavior, symbolic computation might come to be regarded as only a narrow, highly specialized form of cognition. Although it might be possible to treat this specialized form as having a high degree of autonomy (by ignoring the larger system in which it is embedded), the study of cognition would nonetheless include systems consisting of many networks of cognitive processes, each perhaps with its own distinct cognitive domain. [...] An inclusive or mixed mode seems, therefore, a natural strategy to pursue.
> [127, p. 103]

Accordingly, important facts that have been uncovered by computationalist cognitive science will not simply be rejected. Instead of denying its accomplishments the aim is rather to understand them from within a new context that is more encompassing. In other words, the success of enactivism will partly be measured according to how well it manages to incorporate the existing insights of traditional cognitive science.

At least in philosophical terms this possibility exists in principle: the enactivist perspective involves an inherent self-relativization which necessarily opens the door to the objectivist epistemology of computationalism, though certainly stripped of its excess metaphysical baggage (cf. [111]). Indeed, this 'middle way' between relativism and objectivism has already been promoted since the very beginnings of the enactive approach (cf. [82]), and the establishment of a bridge from basic sense-making to 'higher-level' human cognition remains an active research goal (e.g. [41]).

Appendix B

Mainstream biology does not make any proper distinction between the apparently purposeful behavior of an artificial feedback system and that of an organism. Instead, it treats both of them only as instances of 'as-if' teleology, or "teleonomy" [94]. Similar to Dennett's [34] famous "intentional stance" in the cognitive sciences, on this view both living and non-living systems are only to be considered hypothetically 'goal-directed', that is, only as a useful explanatory short-hand that must eventually be reduced to underlying causal regularities. In the case of purposeful behavior this is usually done by appealing to evolutionary history, i.e. by stating, for example, that an animal acted in what appeared to be a 'purposeful' manner *because* it was selected to do so by natural selection (e.g. [89]).

It should be evident that this kind of 'as-if' approach to goal-directed behavior faces severe difficulties, if it can do so at all, in explaining what makes an agent's goals genuinely its own. As such, it cannot help us in resolving one of the main limitations of current embodied AI that we have identified in Section 2. In other words, the mainstream position in biology cannot help us to engineer artificial systems with the ability to generate their own goals rather than merely 'obeying' ones that have been imposed from the outside, for example through an artificial evolutionary algorithm. The teleonomical perspective, while useful in certain contexts, cannot supply us with the appropriate systemic distinctions.

Finally, when we reflectively apply the teleonomical perspective to our own situation as living beings, it follows that this stance reduces the reasons for our actions to external historical conditions. It thus ultimately becomes self-refuting because, if we no longer act for genuine reasons, it denies any possibility of the very rationality needed for us to accept its theoretical proposal (cf. [71, pp. 127–134]). Enactive cognitive science is at least open to the possibility of reconciliation between the role played by external conditions, including evolutionary history, and the evidence of our lived experience, especially the authentic nature of our acting for reasons (cf. [118]). Nevertheless, more work still needs to be done to properly link the biological and phenomenological evidence into one coherent framework.

Appendix C

The aim of this appendix is to clarify the relationship between (i) emergence through self-organization, (ii) autonomy through organizational closure, and (iii) autopoiesis through chemical self-production. One useful way to look at these concepts is in the form of class inclusion from emergence to autonomy to autopoiesis, which coincides with a movement from the most to the least inclusive class. In other words, (i), (ii) and (iii) can all be characterized by emergence through self-organization (cf. [137]). The concept of self-organization can be interpreted in many different ways, but in the autopoietic tradition it is noteworthy for two aspects: (a) local-to-global determination, such that the emergent process has its global identity constituted and constrained as a result of local interactions, and (b) global-to-local determination, whereby the global identity and its ongoing contextual interaction constrain the local interactions [41].

In the case of autonomy (ii), this kind of emergence is of a special kind, namely "dynamic co-emergence", such that the autonomous system is not only characterized by emergence through self-organization, but also by *self-production*: "the whole is constituted by the relations of the parts, and the parts are constituted by the relations they bear to one another in the whole" [118, p. 65]. Finally, autopoietic systems (iii) are also autonomous systems, since they are characterized by such dynamic co-emergence, but they are specific to the biochemical domain. Note that the notion of organizational closure goes further than the concept of "continuous reciprocal causation", as it is used by Wheeler [139] and others. Whereas the latter essentially refers to the relational phenomenon we have here called self-organization, the former denotes continuous reciprocal *generation*.

It is also worth emphasizing here that the notion of autonomy as it is used in the enactive approach is fundamentally different from how it is generally used in robotics and AI [148,147]. While the latter field is generally concerned with a kind of *behavioral* autonomy (e.g. robust and flexible environmental interactions), the former refers to *constitutive* autonomy, namely the self-constitution of an identity under precarious conditions [52]. Note that this does not mean that the enactive account of constitutive autonomy ignores behavioral aspects. On the contrary, as we have already indicated above, highlighting constitutive autonomy in fact entails behavioral autonomy since (i) constitutive autonomy is fundamentally a process of constitution of an identity, and (ii) this emergent identity gives, logically and mechanistically, the point of reference for a domain of interactions [126].

It is therefore an important question as to what extent the separation between the constitutive and behavioral domain of an 'autonomous' system, as generally practiced by current embodied AI, can be justified from the enactive point of view. We consider this issue more fully in Section 3.2.3 of the paper.

References

[1] M.L. Anderson, Embodied cognition: A field guide, Artificial Intelligence 149 (1) (2003) 91–130.
[2] M.L. Anderson, Strike while the iron is, Artificial Intelligence 170 (18) (2006) 1213–1217.
[3] M.L. Anderson, T. Oates, W. Chong, D. Perlis, The metacognitive loop I: Enhancing reinforcement learning with metacognitive monitoring and control for improved perturbation tolerance, Journal of Experimental and Theoretical Artificial Intelligence 18 (3) (2006) 387–411.
[4] M.L. Anderson, D.R. Perlis, Logic, self-awareness and self-improvement: The metacognitive loop and the problem of brittleness, Journal of Logic and Computation 15 (1) (2005) 21–40.
[5] R.C. Arkin, Behavior-Based Robotics, The MIT Press, Cambridge, MA, 1998.
[6] W.R. Ashby, The nervous system as physical machine: With special reference to the origin of adaptive behavior, Mind 56 (221) (1947) 44–59.
[7] W.R. Ashby, Design for a Brain: The Origin of Adaptive Behavior, second ed., Chapman and Hall, London, UK, 1960.
[8] X. Barandiaran, A. Moreno, On what makes certain dynamical systems cognitive: A minimally cognitive organization program, Adaptive Behavior 14 (2) (2006) 171–185.
[9] X. Barandiaran, A. Moreno, Adaptivity: From metabolism to behavior, Adaptive Behavior 16 (5) (2008) 325–344.
[10] X. Barandiaran, K. Ruiz-Mirazo, Modelling autonomy: Simulating the essence of life and cognition, BioSystems 91 (2) (2008) 295–304.
[11] R.D. Beer, A dynamical systems perspective on agent-environment interaction, Artificial Intelligence 72 (1–2) (1995) 173–215.
[12] R.D. Beer, The dynamics of adaptive behavior: A research program, Robotics and Autonomous Systems 20 (2–4) (1997) 257–289.
[13] R.D. Beer, The dynamics of active categorical perception in an evolved model agent, Adaptive Behavior 11 (4) (2003) 209–243.
[14] R.D. Beer, Autopoiesis and cognition in the game of life, Artificial Life 10 (3) (2004) 309–326.
[15] L. Berthouze, T. Ziemke, Epigenetic robotics – modelling cognitive development in robotic systems, Connection Science 15 (4) (2003) 147–150.
[16] M. Bitbol, P.L. Luisi, Autopoiesis with or without cognition: Defining life at its edge, Journal of the Royal Society Interface 1 (1) (2004) 99–107.
[17] M.A. Boden, Autonomy and Artificiality, in: M.A. Boden (Ed.), The Philosophy of Artificial Life, Oxford University Press, Oxford, UK, 1996, pp. 95–108.
[18] P. Bourgine, J. Stewart, Autopoiesis and cognition, Artificial Life 10 (3) (2004) 327–345.
[19] P. Bourgine, F.J. Varela, Introduction: Towards a practice of autonomous systems, in: F.J. Varela, P. Bourgine (Eds.), Toward a Practice of Autonomous Systems: Proc. of the 1st Euro. Conf. on Artificial Life, The MIT Press, Cambridge, MA, 1992, pp. 1–3.
[20] R.A. Brooks, Elephants don't play chess, Robotics and Autonomous Systems 6 (1–2) (1990) 1–16.
[21] R.A. Brooks, Intelligence without representation, Artificial Intelligence 47 (1–3) (1991) 139–160.
[22] R.A. Brooks, From earwigs to humans, Robotics and Autonomous Systems 20 (2–4) (1997) 291–304.
[23] R.A. Brooks, The relationship between matter and life, Nature 409 (6818) (2001) 409–411.
[24] D.J. Chalmers, The Conscious Mind: In Search of a Fundamental Theory, Oxford University Press, Oxford, UK, 1996.
[25] R. Chrisley, Embodied artificial intelligence, Artificial Intelligence 149 (1) (2003) 131–150.
[26] A. Clark, Being There: Putting Brain, Body, and World Together Again, The MIT Press, Cambridge, MA, 1997.
[27] D. Cliff, Computational neuroethology: A provisional manifesto, in: J.-A. Meyer, S.W. Wilson (Eds.), From Animals to Animats: Proc. of the 1st Int. Conf. on Simulation of Adaptive Behavior, The MIT Press, Cambridge, MA, 1991, pp. 29–39.
[28] G. Colombetti, Enaction, sense-making and emotion, in: J. Stewart, O. Gapenne, E.A. Di Paolo (Eds.), Enaction: Towards a New Paradigm for Cognitive Science, The MIT Press, Cambridge, MA, in press.
[29] M.T. Cox, Metacognition in computation: A selected research review, Artificial Intelligence 169 (2) (2005) 104–141.

[30] M.T. Cox, Perpetual self-aware cognitive agents, AI Magazine 28 (1) (2007) 32–45.
[31] M.T. Cox, A. Ram, Introspective multistrategy learning: On the construction of learning strategies, Artificial Intelligence 112 (1–2) (1999) 1–55.
[32] H. De Jaegher, E.A. Di Paolo, Participatory sense-making: An enactive approach to social cognition, Phenomenology and the Cognitive Sciences 6 (4) (2007) 485–507.
[33] D.C. Dennett, Cognitive wheels: The frame problem of AI, in: C. Hookway (Ed.), Minds, Machines, and Evolution: Philosophical Studies, Cambridge University Press, Cambridge, UK, 1984, pp. 129–151.
[34] D.C. Dennett, The Intentional Stance, The MIT Press, Cambridge, MA, 1987.
[35] E.A. Di Paolo, Homeostatic adaptation to inversion of the visual field and other sensorimotor disruptions, in: J.-A. Meyer, et al. (Eds.), From Animals to Animats 6: Proc. of the 6th Int. Conf. on Simulation of Adaptive Behavior, The MIT Press, Cambridge, MA, 2000, pp. 440–449.
[36] E.A. Di Paolo, Organismically-inspired robotics: Homeostatic adaptation and teleology beyond the closed sensorimotor loop, in: K. Murase, T. Asakura (Eds.), Dynamical Systems Approach to Embodiment and Sociality, Advanced Knowledge International, Adelaide, Australia, 2003, pp. 19–42.
[37] E.A. Di Paolo, Unbinding biological autonomy: Francisco Varela's contributions to artificial life, Artificial Life 10 (3) (2004) 231–233.
[38] E.A. Di Paolo, Autopoiesis, adaptivity, teleology, agency, Phenomenology and the Cognitive Sciences 4 (4) (2005) 429–452.
[39] E.A. Di Paolo, H. Iizuka, How (not) to model autonomous behavior, BioSystems 91 (2) (2008) 409–423.
[40] E.A. Di Paolo, J. Noble, S. Bullock, Simulation models as opaque thought experiments, in: M.A. Bedau, et al. (Eds.), Artificial Life VII: Proc. of the 7th Int. Conf. on Artificial Life, The MIT Press, Cambridge, MA, 2000, pp. 497–506.
[41] E.A. Di Paolo, M. Rohde, H. De Jaegher, Horizons for the enactive mind: Values, social interaction, and play, in: J. Stewart, O. Gapenne, E.A. Di Paolo (Eds.), Enaction: Towards a New Paradigm for Cognitive Science, The MIT Press, Cambridge, MA, in press.
[42] H.L. Dreyfus, What Computers Can't Do: A Critique of Artificial Reason, Harper and Row, New York, NY, 1972.
[43] H.L. Dreyfus, From micro-worlds to knowledge representation: AI at an impasse, in: J. Haugeland (Ed.), Mind Design: Philosophy, Psychology, Artificial Intelligence, The MIT Press, Cambridge, MA, 1981, pp. 161–204.
[44] H.L. Dreyfus, Being-in-the-World: A Commentary on Heidegger's Being and Time, Division 1, The MIT Press, Cambridge, MA, 1991.
[45] H.L. Dreyfus, What Computers Still Can't Do: A Critique of Artificial Reason, The MIT Press, Cambridge, MA, 1992.
[46] H.L. Dreyfus, Why Heideggerian AI failed and how fixing it would require making it more Heideggerian, Philosophical Psychology 20 (2) (2007) 247–268.
[47] S. Franklin, Artificial Minds, The MIT Press, Cambridge, MA, 1995.
[48] W.J. Freeman, How Brains Make up their Minds, Weidenfeld & Nicolson, London, UK, 1999.
[49] T. Froese, On the role of AI in the ongoing paradigm shift within the cognitive sciences, in: M. Lungarella, et al. (Eds.), 50 Years of AI, Springer-Verlag, Berlin, Germany, 2007, pp. 63–75.
[50] T. Froese, E.A. Di Paolo, Stability of coordination requires mutuality of interaction in a model of embodied agents, in: M. Asada, et al. (Eds.), From Animals to Animats 10: Proc. of the 10th Int. Conf. on Simulation of Adaptive Behavior, Springer-Verlag, Berlin, Germany, 2008, pp. 52–61.
[51] T. Froese, A. Spiers, Toward a phenomenological pragmatics of enactive perception, in: Enactive/07: Proc. of the 4th Int. Conf. on Enactive Interfaces, Association ACROE, Grenoble, France, 2007, pp. 105–108.
[52] T. Froese, N. Virgo, E. Izquierdo, Autonomy: A review and a reappraisal, in: F. Almeida e Costa, et al. (Eds.), Advances in Artificial Life: Proc. of the 9th Euro. Conf. on Artificial Life, Springer-Verlag, Berlin, Germany, 2007, pp. 455–464.
[53] S. Gallagher, Are minimal representations still representations?, International Journal of Philosophical Studies 16 (3) (2008) 351–369.
[54] S. Gallagher, D. Zahavi, The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science, Routledge, London, UK, 2008.
[55] I. Harvey, Untimed and misrepresented: Connectionism and the computer metaphor, Artificial Intelligence and Simulation of Behavior Quarterly 96 (1996) 20–27.
[56] I. Harvey, Robotics: Philosophy of mind using a screwdriver, in: T. Gomi (Ed.), Evolutionary Robotics: From Intelligent Robots to Artificial Life, vol. III, AAI Books, Ontario, Canada, 2000, pp. 207–230.
[57] I. Harvey, E.A. Di Paolo, R. Wood, M. Quinn, E.A. Tuci, Evolutionary robotics: A new scientific tool for studying cognition, Artificial Life 11 (1–2) (2005) 79–98.
[58] S. Harnard, Minds, machines and Searle, Journal of Theoretical and Experimental Artificial Intelligence 1 (1) (1989) 5–25.
[59] S. Harnard, The symbol grounding problem, Physica D 42 (1990) 335–346.
[60] R.M. Harnish, Minds, Brains, Computers: An Historical Introduction to the Foundations of Cognitive Science, Blackwell Publishers, Malden, MA, 2002.
[61] W.F.G. Haselager, Robotics, philosophy and the problems of autonomy, Pragmatics, Cognition 13 (3) (2005) 515–532.
[62] J. Haugeland, Artificial Intelligence: The Very Idea, The MIT Press, Cambridge, MA, 1985.
[63] J. Haugeland, What is mind design?, in: J. Haugeland (Ed.), Mind Design II: Philosophy, Psychology, Artificial Intelligence, The MIT Press, Cambridge, MA, 1997, pp. 1–28.
[64] M. Heidegger, Sein und Zeit (Being and Time), Blackwell Publishing Ltd., Oxford, UK, 1962, trans. by: J. Macquarrie, E. Robinson.
[65] M. Heidegger, Die Grundbegriffe der Metaphysik: Welt, Endlichkeit, Einsamkeit (The Fundamental Concepts of Metaphysics: World, Finitude, Solitude), Indiana University Press, Bloomington, IN, 1995, trans. by: W. McNeill, N. Walker. First edition 1929.
[66] H. Iizuka, T. Ikegami, Simulating autonomous coupling in discrimination of light frequencies, Connection Science 16 (4) (2004) 283–299.
[67] H. Iizuka, E.A. Di Paolo, Toward Spinozist robotics: Exploring the minimal dynamics of behavioral preference, Adaptive Behavior 15 (4) (2007) 359–376.
[68] H. Iizuka, E.A. Di Paolo, Extended homeostatic adaptation: Improving the link between internal and behavioural stability, in: M. Asada, et al. (Eds.), From Animals to Animats 10: Proc. of the 10th Int. Conf. on Simulation of Adaptive Behavior, Springer-Verlag, Berlin, Germany, 2008, pp. 1–11.
[69] T. Ikegami, K. Suzuki, From homeostatic to homeodynamic Self, BioSystems 91 (2) (2008) 388–400.
[70] E. Izquierdo, I. Harvey, The dynamics of associative learning in an evolved situated agent, in: F. Almeida e Costa, et al. (Eds.), Advances in Artificial Life: Proc. of the 9th Euro. Conf. on Artificial Life, Springer-Verlag, Berlin, Germany, 2007, pp. 365–374.
[71] H. Jonas, The Phenomenon of Life: Toward a Philosophical Biology, Northwestern University Press, Evanston, IL, 2001.
[72] H. Jonas, Biological foundations of individuality, International Philosophical Quarterly 8 (1968) 231–251.
[73] I. Kant, Kritik der Urteilskraft (Critique of Judgment), Hacket Publishing Company, Indianapolis, IN, 1987, trans. by: W.S. Pluhar. First edition 1790.
[74] S.A. Kauffman, Investigations, Oxford University Press, New York, NY, 2000.
[75] D. Kirsh, Today the earwig, tomorrow man?, Artificial Intelligence 47 (1–3) (1991) 161–184.
[76] I. Kohler, The formation and transformation of the perceptual world, Psychological Issues 3 (4) (1964) 1–173.
[77] C.G. Langton, Artificial life, in: C.G. Langton (Ed.), Artificial Life: Proceedings of an Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems, in: Santa Fe Institute Studies in the Sciences of Complexity, vol. 4, Addison-Wesley, Redwood City, CA, 1989, pp. 1–47.
[78] P.L. Luisi, Autopoiesis: A review and reappraisal, Naturwissenschaften 90 (2003) 49–59.
[79] M. Lungarella, G. Metta, R. Pfeifer, G. Sandini, Developmental robotics: A survey, Connection Science 15 (4) (2003) 151–190.
[80] I. Macinnes, E.A. Di Paolo, The advantages of evolving perceptual cues, Adaptive Behavior 14 (2) (2006) 147–156.
[81] H.R. Maturana, F.J. Varela, Autopoiesis and Cognition: The Realization of the Living, Kluwer Academic Publishers, Dordrecht, Holland, 1980.
[82] H.R. Maturana, F.J. Varela, The Tree of Knowledge: The Biological Roots of Human Understanding, Shambhala Publications, Boston, MA, 1987.
[83] F. Mavelli, K. Ruiz-Mirazo, Stochastic simulations of minimal self-reproducing cellular systems, Phil. Trans. R. Soc. B 362 (1486) (2007) 1789–1802.

[84] J. McCarthy, P.J. Hayes, Some philosophical problems from the standpoint of artificial intelligence, in: B. Meltzer, D. Michie (Eds.), Machine Intelligence 4, Edinburgh University Press, Edinburgh, UK, 1969, pp. 463–502.
[85] M. McGann, Enactive theorists do it on purpose: Toward an enactive account of goals and goal-directedness, Phenomenology and the Cognitive Sciences 6 (4) (2007) 463–483.
[86] B. McMullin, Thirty years of computational autopoiesis: A review, Artificial Life 10 (3) (2004) 277–295.
[87] C. Melhuish, I. Ieropoulos, J. Greenman, I. Horsfield, Energetically autonomous robots: Food for thought, Autonomous Robots 21 (3) (2006) 187–198.
[88] M. Merleau-Ponty, Phénomènologie de la perception (Phenomenology of Perception), Routledge & Kegan Paul, New York, NY, 1962, trans. by: C. Smith. First edition 1945.
[89] R.G. Millikan, Biosemantics, The Journal of Philosophy 86 (6) (1989) 281–297.
[90] A. Moreno, A. Etxeberria, Agency in natural and artificial systems, Artificial Life 11 (1–2) (2005) 161–175.
[91] A.F. Morse, T. Ziemke, On the role(s) of modeling in cognitive science, Pragmatics & Cognition 16 (1) (2008) 37–56.
[92] A.F. Morse, R. Lowe, T. Ziemke, Towards an enactive cognitive architecture, in: Proc. of the Int. Conf. on Cognitive Systems (CogSys 2008), Karlsruhe, Germany, 2008.
[93] T. Nagel, What is it like to be a bat? Philosophical Review 83 (4) (1974) 435–450.
[94] E. Nagel, Teleology revisited: Goal-directed processes in biology, The Journal of Philosophy 74 (5) (1977) 261–279.
[95] A. Noë, Action in Perception, The MIT Press, Cambridge, MA, 2004.
[96] S. Nolfi, D. Floreano, Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines, The MIT Press, Cambridge, MA, 2000.
[97] J.K. O'Regan, A. Noë, A sensorimotor account of vision and visual consciousness, Behavioral and Brain Sciences 24 (5) (2001) 939–1031.
[98] D. Parisi, Internal robotics, Connection Science 16 (4) (2004) 325–338.
[99] R. Pfeifer, Building 'fungus Eaters': Design principles of autonomous agents, in: P. Maes, et al. (Eds.), From Animals to Animats 4: Proc. of the 4th Int. Conf. on Simulation of Adaptive Behavior, The MIT Press, Cambridge, MA, 1996, pp. 3–12.
[100] R. Pfeifer, J. Bongard, How the Body Shapes the Way We Think: A New View of Intelligence, The MIT Press, Cambridge, MA, 2007.
[101] R. Pfeifer, F. Iida, J. Bongard, New robotics: Design principles for intelligent systems, Artificial Life 11 (1–2) (2005) 99–120.
[102] R. Pfeifer, G. Gómez, Interacting with the real world: Design principles for intelligent systems, Artificial Life and Robotics 9 (1) (2005) 1–6.
[103] R. Pfeifer, C. Scheier, Understanding Intelligence, The MIT Press, Cambridge, MA, 1999.
[104] J. Prinz, Putting the brakes on enactive perception, Psyche 12 (1) (2006) 1–19.
[105] A.N. Rosenblueth, N. Wiener, J. Bigelow, Behavior, purpose and teleology, Philosophy of Science 10 (1943) 18–24.
[106] K. Ruiz-Mirazo, A. Moreno, Basic autonomy as a fundamental step in the synthesis of life, Artificial Life 10 (3) (2004) 235–259.
[107] G. Sandini, G. Metta, D. Vernon, The iCub cognitive humanoid robot: An open-system research platform for enactive cognition, in: M. Lungarella, et al. (Eds.), 50 Years of AI, Springer-Verlag, Berlin, Germany, 2007, pp. 358–369.
[108] J. Searle, Minds, brains and programs, Behavioral and Brain Sciences 3 (3) (1980) 417–457.
[109] N.E. Sharkey, T. Ziemke, A consideration of the biological and psychological foundations of autonomous robotics, Connection Science 10 (3–4) (1998) 361–391.
[110] T. Smithers, Autonomy in robots and other agents, Brain and Cognition 34 (1) (1997) 88–106.
[111] J. Stewart, Life = Cognition: The epistemological and ontological significance of Artificial Life, in: F.J. Varela, P. Bourgine (Eds.), Toward a Practice of Autonomous Systems: Proc. of the 1st Euro. Conf. on Artificial Life, The MIT Press, Cambridge, MA, 1992, pp. 475–483.
[112] J. Stewart, Cognition = life: Implications for higher-level cognition, Behavioral Processes 35 (1996) 311–326.
[113] J. Stewart, E.A. Di Paolo, O. Gapenne, Introduction, in: J. Stewart, O. Gapenne, E.A. Di Paolo (Eds.), Enaction: Towards a New Paradigm for Cognitive Science, The MIT Press, Cambridge, MA, in press.
[114] M. Suzuki, D. Floreano, Enactive robot vision, Adaptive Behavior 16 (2/3) (2008) 122–128.
[115] J.G. Taylor, The Behavioral Basis of Perception, Yale University Press, New Haven, CT, 1962.
[116] E. Thompson, Life and mind: From autopoiesis to neurophenomenology. A tribute to Francisco Varela, Phenomenology and the Cognitive Sciences 3 (4) (2004) 381–398.
[117] E. Thompson, Sensorimotor subjectivity and the enactive approach to experience, Phenomenology and the Cognitive Sciences 4 (4) (2005) 407–427.
[118] E. Thompson, Mind in Life: Biology, Phenomenology, and the Sciences of Mind, The MIT Press, Cambridge, MA, 2007.
[119] E. Thompson, D. Zahavi, Philosophical issues: Phenomenology, in: P.D. Zelazo, M. Moscovitch, E. Thompson (Eds.), The Cambridge Handbook of Consciousness, Cambridge University Press, Cambridge, UK, 2007, pp. 67–87.
[120] S. Torrance, In search of the enactive: Introduction to special issue on enactive experience, Phenomenology and the Cognitive Sciences 4 (4) (2005) 357–368.
[121] S. Torrance, Introduction to the second special issue on enactive experience, Phenomenology and the Cognitive Sciences 6 (4) (2007) 425.
[122] F.J. Varela, Principles of Biological Autonomy, Elsevier North Holland, New York, NY, 1979.
[123] F.J. Varela, Organism: A meshwork of selfless selves, in: A.I. Tauber (Ed.), Organisms and the Origins of Self, Kluwer Academic Publishers, Dordrecht, Netherlands, 1991, pp. 79–107.
[124] F.J. Varela, Autopoiesis and a biology of intentionality, in: B. McMullin, N. Murphy (Eds.), Proc. of Autopoiesis and Perception: A Workshop with ESPRIT BRA 3352, Dublin City University, Dublin, Ireland, 1992, pp. 4–14.
[125] F.J. Varela, The early days of autopoiesis: Heinz and Chile, Systems Research 13 (3) (1996) 407–416.
[126] F.J. Varela, Patterns of life: Intertwining identity and cognition, Brain and Cognition 34 (1) (1997) 72–87.
[127] F.J. Varela, E. Thompson, E. Rosch, The Embodied Mind: Cognitive Science and Human Experience, The MIT Press, Cambridge, MA, 1991.
[128] F.J. Varela, H.R. Maturana, R. Uribe, Autopoiesis: The organization of living systems, its characterization and a model, BioSystems 5 (1974) 187–196.
[129] M. Velmans, Where experiences are: Dualist, physicalist, enactive and reflexive accounts of phenomenal consciousness, Phenomenology and the Cognitive Sciences 6 (4) (2007) 547–563.
[130] D. Vernon, D. Furlong, Philosophical foundations of AI, in: M. Lungarella, et al. (Eds.), 50 Years of AI, Springer-Verlag, Berlin, Germany, 2007, pp. 53–62.
[131] D. Vernon, G. Metta, G. Sandini, A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents, IEEE Trans. on Evolutionary Computation 11 (2) (2007) 151–180.
[132] J. von Uexküll, Theoretische Biologie, Springer-Verlag, Berlin, Germany, 1928.
[133] J. von Uexküll, Streifzüge durch die Umwelten von Tieren und Menschen: Ein Bilderbuch unsichtbarer Welten (A stroll through the worlds of animals and men: a picture book of invisible worlds), in: C.H. Schiller (Ed.), Instinctive Behavior: The Development of a Modern Concept, International Universities Press, New York, NY, 1957, pp. 5–80, trans. by: C.H. Schiller. First edition 1934. Also appeared in Semiotica 89 (4) (1992) 319–391.
[134] J. von Uexküll, The theory of meaning, Semiotica 42 (1) (1982) 25–82.
[135] W.G. Walter, The Living Brain, W.W. Norton & Company, New York, NY, 1953.
[136] A. Weber, F.J. Varela, Life after Kant: Natural purposes and the autopoietic foundations of biological individuality, Phenomenology and the Cognitive Sciences 1 (2002) 97–125.
[137] M. Wheeler, Cognition's coming home: The reunion of life and mind, in: P. Husbands, I. Harvey (Eds.), Proc. of the 4th Euro. Conf. on Artificial Life, The MIT Press, Cambridge, MA, 1997, pp. 10–19.

[138] M. Wheeler, Reconstructing the Cognitive World: The Next Step, The MIT Press, Cambridge, MA, 2005.
[139] M. Wheeler, Cognition in context: Phenomenology, situated robotics and the frame problem, International Journal of Philosophical Studies 16 (3) (2008) 323–349.
[140] R. Wood, E.A. Di Paolo, New models for old questions: Evolutionary robotics and the 'A not B' error, in: F. Almeida e Costa, et al. (Eds.), Advances in Artificial Life: Proc of the 9th Euro. Conf. on Artificial Life, Springer-Verlag, Berlin, Germany, 2007, pp. 1141–1150.
[141] T. Ziemke, Rethinking grounding, in: A. Riegler, A. von Stein, M. Peschl (Eds.), Understanding Representation in the Cognitive Sciences, Plenum Press, New York, NY, 1999, pp. 177–190.
[142] T. Ziemke, The construction of 'reality' in the robot: Constructivist perspectives on situated artificial intelligence and robotics, Foundations of Science 6 (1) (2001) 163–233.
[143] T. Ziemke, Are robots embodied?, in: C. Balkenius, et al. (Eds.), Proc. of the 1st Int. Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Lund, Sweden, Lund University Cognitive Studies, vol. 85, 2001, pp. 75–93.
[144] T. Ziemke, Embodied AI as science: Models of embodied cognition, embodied models of cognition, or both?, in: F. Iida, et al. (Eds.), Embodied Artificial Intelligence, Springer-Verlag, Heidelberg, Germany, 2004, pp. 27–36.
[145] T. Ziemke, Cybernetics and embodied cognition: On the construction of realities in organisms and robots, Kybernetes 34 (1/2) (2005) 118–128.
[146] T. Ziemke, What's life got to do with it?, in: A. Chella, R. Manzotti (Eds.), Artificial Consciousness, Imprint Academic, Exeter, UK, 2007, pp. 48–66.
[147] T. Ziemke, On the role of emotion in biological and robotic autonomy, BioSystems 91 (2) (2008) 401–408.
[148] T. Ziemke, N.E. Sharkey, A stroll through the worlds of robots and animals: Applying Jakob von Uexküll's theory of meaning to adaptive robots and artificial life, Semiotica 134 (1–4) (2001) 701–746.
[149] T. Ziemke, M. Thieme, Neuromodulation of reactive sensorimotor mappings as a memory mechanism in delayed response tasks, Adaptive Behavior 10 (3/4) (2002) 185–199.