Proceedings of the International Symposium on AI-Inspired Biology, Jackie Chappell, Susannah Thorpe, Nick Hawes and Aaron Sloman (Eds.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Constructing Emotions

Epistemological groundings and applications in robotics for a synthetic approach to emotions

Luisa Damiano♦ and Lola Cañamero♦

Abstract. Can the sciences of the artificial positively contribute to the scientific exploration of life and cognition? Can they actually improve the scientific knowledge of natural living and cognitive processes, from biological metabolism to reproduction, from conceptual mapping of the environment to logic reasoning, language, or even emotional expression? To these kinds of questions our article aims to answer in the affirmative. Its main object is the scientific emergent methodology often called the "synthetic approach", which promotes the programmatic production of embodied and situated models of living and cognitive systems in order to explore aspects of life and cognition not accessible in natural systems and scenarios. The first part of this article presents and discusses the synthetic approach, and proposes an epistemological framework which promises to warrant genuine transmission of knowledge from the sciences of the artificial to the sciences of the natural. The second part of this article looks at the research applying the synthetic approach to the psychological study of emotional development. It shows how robotics, through the synthetic methodology, can develop a particular perspective on emotions, coherent with current psychological theories of emotional development and fitting well with the recent "cognitive extension" approach proposed by cognitive sciences and philosophy of mind.

INTRODUCTION

"Understanding by building" [41] is the promise of the "synthetic approach", currently presented as the methodology by which the sciences of the artificial can contribute to the scientific research on life and on cognition. Considerable work has been done by specialists of artificial sciences to define ways of "synthetically" modelling living and cognitive phenomena [7, 9, 16, 17, 33, 41, 42]. This renders even more important the question of how pertinent this modelling is for the scientific knowledge of life and cognition. What are the status and the value of the insights coming from the exploration of synthetic models of life and cognition? Is there a criterion to warrant a positive transmission of knowledge from the sciences of the artificial to the sciences of the natural? Can systems built with different materials, and endowed with different embodiments, be considered effective models of natural living and cognitive systems? Questions of this kind, if left unanswered, threaten the legitimacy of the synthetic approach, as well as its effective integration among the explorative practices accepted by the scientific community as a source of valuable insights.

The main aim of this article is to generate answers for the aforementioned questions, by offering an epistemological solution to the issue of the legitimization of the synthetic approach. We will try to present this solution not only through speculative dissertation, but also through a concrete application taken from current scientific practice. In Section 1 we will introduce and describe an epistemological framework able to provide the synthetic approach with a general epistemological legitimacy, as well as with useful criteria to evaluate the pertinence of its applications to the study of living and cognitive phenomena. In Section 2 we will present one concrete implementation of this approach, which, through the support of the epistemological framework previously proposed, explores a class of phenomena among the most problematic for science – emotions. The intent of this presentation is to show that the synthetic approach, when rigorously applied to emotional phenomena, can generate significant theoretical and methodological results. Indeed, it produces a rather innovative view of emotions, which heralds a generative two-way transmission of knowledge between epigenetic robotics and the psychology of emotional development.

1 AN EPISTEMOLOGICAL FRAMEWORK FOR AIIB AND AIICS[1]

Grounding the synthetic approach in constructivism

a) The synthetic approach: the methodology of a creative science

Increasingly, the sciences of the artificial claim they are able to go beyond a merely engineering approach. They manifest the ambition to structure a properly scientific approach, and to deal with the crucial questions at the basis of "sciences of the natural" such as biology, psychology and cognitive sciences. The intent is to offer to these sciences a privileged methodology to get to the hidden mechanisms of their objects: "understanding by building" [41]. In the current literature this emergent methodological principle promotes the programmatic production of embodied and situated models of living and cognitive systems in order to explore aspects of life and cognition usually not accessible in natural systems and scenarios. The procedural novelty of this approach, often called "constructive" or "synthetic" [7, 9, 16, 17, 33, 41, 42], is that it inverts the usual order between analysis of behaviour and construction of models. It requires the researcher firstly to "embed" the basic hypothesis on life and cognition in working artificial systems and then to examine the behaviours they produce [16]. It proposes a heuristics which opens to scientific investigation not only the question "how does it work?" but also "why this and not that?" [33] Besides this, it provides to science a new access to the simplicity of nature – better: a new concept of the simplicity of scientific explanation.

The current versions of the synthetic approach share a common thesis on the genesis of complex behaviours, paradigmatically exposed in Herbert Simon's "story of the ant" [46, 16, 40, 4]. This thesis, according to its most detailed elaborations, can be decomposed as follows:

♦ STRI and School of Computer Science, University of Hertfordshire, College Lane, Hatfield, Herts AL10 9, U.K.

[1] We use AIIB as an abbreviation of Artificial Intelligence Inspired Biology, and AIICS as an abbreviation of Artificial Intelligence Inspired Cognitive Science.

(1) Complex behaviours rest on the interaction of simple elements, but are not found in these elements taken separately.
(2) Complex behaviours arise from the interplay between the interacting elements, the global system constituted by them and the environment with which the system interacts.
(3) Complex behaviours tend to exceed the observer's possibilities of calculation and prevision, even when he built the system which manifests them.

This is clearly not a mechanistic perspective. The proponents of the synthetic approach express its "post-mechanistic" character through the notion of "emergence" – a sort of emblem of the arising "post-classical" science [14].[2] This notion not only disobeys the old mechanistic postulate "the whole is the sum of its parts". It also refuses the mechanistic heuristic pretension to a science able to know past, present and future behaviours of the objects it studies – Laplace's demon "omniscient science" [3, 43, 47]. And it is through the dismissal of these two traditional schemes of scientific rationality that the concept of emergence promises to the synthetic approach effective and simple models of life and cognition. Hypothesis (1) gives to science the possibility to plan the construction of artificial models of living and cognitive systems as the organisation of elemental components in integrated dynamical units – that is, as the manipulation of relations between elemental components.[3] Hypothesis (2) allows researchers to ideate and design for these systems quite simple internal mechanisms, as it grounds the complexity of living and cognitive behaviours not in the systems, but in the interplay between them and the environment. Hypothesis (3) authorizes scientists to consider these artificial models of living and cognitive systems as actual generators of knowledge, since, according to this hypothesis, once these models are situated in an environment, they can express behaviours unexpected by their constructor, and give him new insights and feedbacks about the thesis on life and cognition that they embed.

But this theoretical structure, although offering to the synthetic approach the advantage of a simple explanation, does not spare it the disadvantages of a controversial explanation of living and cognitive processes. This emergentist framework, relying on one of the most problematic and debated notions of contemporary science, is not a good candidate for solving doubts about the contribution of the science of the artificial to the scientific understanding of natural phenomena.

Even if we could obtain full agreement on the idea that the creativity of nature can be conceived in emergentist terms, many unanswered critical questions would remain. Can the artificial construction of natural phenomena actually reproduce the creative action of nature? Can it actually accede to the natural mechanisms creating life and cognition? Can it genuinely penetrate and reconstruct – and not roughly imitate – complex processes such as biological metabolism and reproduction, or conceptual mapping of the environment, linguistic expression, emotion? In short: in what sense is a significant transmission of knowledge from the sciences of the artificial to the sciences of the natural possible?

History of science shows that this kind of transmissions has taken place since the birth of the sciences of the artificial, and with remarkable results. Two of the most significant examples are the application in biology of the cybernetic concept of feedback, fundamental for the scientific modelling of cellular metabolism and biological stability [14], and the transfer of the scheme of the digital computer from the engineered production of artefacts to the scientific description of natural cognition [4]. This transfer of knowledge has the merit to have given rise to the classic or computationalist form of cognitive sciences, and to have created the paradigm which oriented their development until recent times – actually it still does, albeit partially.[4] Besides, philosophy of science has lately tended to acknowledge and enhance the positive contribution of the sciences of the artificial to the scientific understanding of nature [1, 44]. For example, some of the latter's new trends, such as philosophy of scientific instruments and philosophy of technology, are developing the thesis that the stable production of phenomena through artefacts could allow one to overcome the problem of induction [20]. Moreover, current scientific literature is increasingly engaged in elaborating procedures directed to implement and make rigorous the application of the sciences of the artificial to research on natural phenomena. But history, philosophy and methodology leave open the issue of the epistemological legitimacy of this emerging way of doing science.

The synthetic approach does not simply subordinate analysis – the classically privileged direction of scientific investigation – to synthesis – indeed, a synthesis which intends to diverge from the mechanistic "sum of parts". This approach challenges the "representationalist" epistemology typical of modern science.[5] It

[2] The divergence between emergentism and mechanism we invoke in the text refers to mechanism as the philosophical (ontological and epistemological) view at the basis of "modern" or "Newtonian" scientific tradition. To simplify, this view can be related to two main elements: (a) the idea of the natural universe as a deterministic succession of qualitatively homogeneous states (a "clockwork universe" wherein everything can be deduced from a certain starting point), and (b) the idea of the observer as a cognitive agent endowed with an external and neutral point of view (an "absolute spectator of nature"). These elements (which, for the sake of brevity, are over-schematized here) ground the scientific heuristics which defines what some epistemological literature, as well as this article, calls "classical science" [3, 43, 47]. This heuristics corresponds to the ideal of elaborating a description of nature endowed with five primary characteristics: determinism, objectivity, simplicity, homogeneity, completeness, well-described in [43]. With the adjective "post-classical" or "non-classical" we refer to every form of science which departs from this heuristics as well as from the view of nature and of science on which it rests. In this article emergentism is defined post-classical in this sense, and not in the sense that it implies the refusal of "mechanistic explanations", i.e. explanations which rely on the identification of mechanisms and avoid the introduction of meta-physical forces (Sect. 2.c).

[3] According to the emergentist point of view, organisational interrelations between elemental components can inhibit some of their individual behaviours and generate collective behaviours which are not expressed in the elements taken separately [14].

[4] Since the 1980s, the computationalist paradigm is strongly criticised, mainly for its poor biological plausibility [38]. In the 1990s, from the criticisms to computationalism arose a new candidate to the guide of cognitive sciences, namely, the paradigm of "embodiment" [12]. This is supported by trends of research which intend to overcome the Cartesian dichotomy inherited by computationalism – that is, they intend to "re-embody the mind" [50]. For the most part, current research uses descriptive solutions which hybrid the two paradigms [12].

[5] Representationalsim is the way of conceiving knowledge typical of classical science as defined before (Footnote 2). It grounds the ideal of

promotes a way of describing nature which diverges from the old-fashioned ideal of an exhaustive reflection of the natural world free from subjective alterations. "Understanding by building" proposes to practice science as a deliberate act of construction, exercised on systems which would not exist without this act. It demands that researchers think and implement science as a form of knowledge which actively creates, and does not passively reflect, the phenomena explored. It stimulates the elaboration of new epistemological groundings for the scientific enterprise – epistemological principles able to generate answers for the critical questions which threaten the legitimacy of the new creative approach to the modelling of nature.[6]

b) Scientific constructivism: an epistemological framework for the synthetic approach

The hypothesis we propose in this article is that some epistemological groundings appropriate for the new creative science already exist. They have been elaborated by a heterodox branch of twentieth-century science, made up of the first groups of research to introduce and develop the notions of autonomy and self-organisation in biology and cognitive sciences. They are a few mutually independent groups, among which some were dedicated to naturalistic inquiries (i.e. organicistic embryology groups, such as, the Cambridge group and the Brussels group, and the thermodynamics of dissipative structures group, that is, the School of Brussels founded by Ilya Prigogine), and some to cybernetic research (e.g. the Biological Computer Laboratory group founded by Heinz von Foerster[7] and the French neoconnectionist group of Henri Atlan). Usually the main contribution attributed them is the new vision of nature generated by their exploration of the "endogenously controlled organised systems" – i.e. the systems that they labelled as "autonomous" or "self-organising". But these first explorers offered to contemporary science another relevant contribution, often neglected. It consists in a new scientific epistemology, aimed at underpinning a tradition of research which conceives and practices scientific knowledge not as the "representation of nature", but as its "construction" (i.e. a "co-construction" due to the interaction of the observer with the reality he explores).[8]

Indeed, the notion of self-organisation – a notion which is closely linked to that of emergence – is strongly heterodox. It is not merely the core concept of a post-mechanistic scenario in which natural evolution manifests a property denied by the Newtonian tradition: the creativity required to overcome the physico-chemical level of reality and give rise to "qualitatively different" or "emergent" levels – life, cognition, man, human scientific and technological creations. The notion of self-organisation, by supporting this view of natural evolution, is also at the heart of the pioneers' transition to a new scientific epistemology. As they adopted this emergentist theoretical background, the early researchers on self-organisation rejected the classical image of the observer as an "absolute spectator" of nature, without localization and perspective. They described the scientific observer as belonging to the natural world and approaching nature from within – from a "limited and situated point of view". According to their evolutionary conception, the observer is an embodied material system belonging to the class of the natural self-organising systems. With every other system of this kind, he can experience external events insofar as they destabilize its internal dynamics and to the extent he can attribute them operational meanings through self-regulation. His knowledge is made of interpretations determined not only by its physico-biological organisation, but also by the specific characteristics of the socio-cultural systems to which he belongs. He cannot develop categories able to generate neutral representations of reality. His theoretical categories play a selective and constructive role, and limit his descriptive domain to those aspects of nature they can define – "build" – as determined objects of research. These categories offer to the scientist not an objective knowledge in the classical sense, but a "pertinent" knowledge, that is, knowledge able to warrant him effective operationality in the domain to which they can be applied.[9]

This transition from representationalism to constructivism was not only due to speculation. For the pioneers it was primarily due to the need of providing an effective description of the systems they studied.

Actually, the "internal determination" typical of autonomous systems – the fact that, by self-regulation, they can break the causality of external variations and determine their internal variations – corresponds for science to a limit of intelligibility. This kind of "endogenous control" implies that autonomous systems can resist not only environmental pressures, but also scientific knowing actions. The observation and the manipulation of this class of systems go with an uncertainty which deprive of any plausibility the classical scientific heuristics. With respect to autonomous systems science cannot satisfy the request of knowing past, present and future behaviours of the objects it explores – "capturing their dynamical law". In other words: with regard to these systems science cannot produce a standard classic description [3, 43].[10] Their unpredictable transformations oblige the observer to permanently (re)negotiate with its objects the

---

the objectivity of scientific knowledge as independence from its subject or subjective aspects [3, 43, 45].

[6] The underlying hypothesis is that objectivist representationalism (i.e., to put it roughly, the doctrine according to which the knowing subject, via his sensory apparatus, can dispose of internal representations of external and pre-determined objects) is not the best gnoseological option for the synthetic approach. This epistemological doctrine tends to describe human knowledge as the passive internalisation of the external world and, as a consequence, tends to see any active role of determination exercised by the subject on his objects of knowledge as a source of alteration. This leads us to propose for the grounding of the synthetic approach a constructivist epistemology, as it sees the active role of the subject in the process of cognition not as a source of subjective contamination of his knowledge of reality, but as a necessary ingredient of it, which, in certain conditions, can produce pertinent knowledge. Our option, which orients the development of this article, does not intend to imply that all the proponents of the synthetic approach are or must be constructivist, nor it deny that often they adhere to a represetationalist epistemology.

[7] As well-known, this laboratory hosted researchers such as Gordon Pask, Georg Zopf and Ross Ashby.

[8] The contents of the present sub-section of the article are dealt with in detail in [14].

[9] The primary references are [3, 21, 25, 39, 43, 47].

[10] As pointed out by some of the pioneers [3, 21, 39, 43], to describe autonomous systems requires to depart from the classical ideal of the scientific "representation" of nature (Footnote 2) – i.e. to produce a description which is not deterministic, not simple (no fundamental levels of observation result available; to accede to all the observable content of the system different levels of observation are needed), not homogeneous (more than one system of description is needed), not complete (depending on the descriptive system used, different observables are available), not objective in the classical sense (dependant on the subjective levels of observation and systems of description).

categories and the pertinent observables of their scientific characterisation. When they are applied to autonomous systems, scientific categories can only produce partial and revocable descriptive solutions. They express a kind of functioning which forces the observer to definitively decline the traditional idea of panoramic inspections of nature, and practice scientific research as the construction of plausible referents for entities which in themselves are not accessible. Concretely, this practice corresponds to a post-classical procedure of characterisation: to structure and coordinate an increasing multiplicity of theoretical levels of description, each able to define as a definite object of research a different aspect of the system explored. In short: to find and to study carefully an autonomous system's resistances to the application of a model; to exploit this study for the development of new theoretical points of view on the system; to move from one to another of these point of views, tracking the system's manifestations – its transformations. The pioneers thought this heuristics as a style of knowledge which requires the researcher to learn from nature how to build it as a set of definite objects of research.[11] And they conceived the artificial construction as a useful complement of this theoretical construction of nature.[12]

The epistemological thesis at the basis of this new heuristics radicalises the outcomes of the "crisis" which in the last century weakened the foundations of classical science. According to the pioneers' view, the action of determination exercised by the observer on the objects he explores is not the contingent alteration of a reality which is neutrally accessible. Indeed, it is a positive and constitutive ingredient of the scientific description of nature: an essential creative activity, without which reality cannot manifest the form of defined objects of research.

On this thesis the first explorers of self-organisation built a new and complex epistemological framework for science, which deeply transforms the structure of the classical one. At a conceptual level, it integrates the traditional epistemological dichotomies, but modifies their internal configuration [3, 43, 47]. It transforms their opposite terms in complementary terms, and allows science to adopt a kind of intelligibility in which *discovery* can be *invention*, scientific *facts* can be *artefacts*, *objective evidence* can converge with *subjective construction*, and the old-fashioned "spontaneous *manifestations* of nature" can be found in its theoretical and artificial scientific *reifications*.

This is the epistemological framework we propose as pertinent for the grounding of the current synthetic approach. In the remainder of this section we will focus on two of its basic epistemological principles, introduced to support a modelling devoted to "generate" the natural behaviours explored, and not to "formally represent their law". We will propose these principles not in their early versions, but in the most elaborated one. It was introduced by the School of Santiago founded by Humberto Maturana and Francisco Varela – i.e. the autopoietic biology group, which, in a sense, can be considered a descendant of the pioneer groups [33-34].[13] We believe that these two principles could offer to the synthetic approach not only a general epistemological legitimacy, but also some criteria useful to evaluate the pertinence of its applications to the study of life and cognition.

c) Two epistemological principles for AIIB and AIICS

I. The first principle proposes a constructivist definition of scientific explanation, paradigmatically expressed in many places of Maturana and Varela's literature.

"If you want to explain lightning, you must provide a mechanism that generates it." (H. Maturana)[14]

The epistemological content of this principle can be conceived as an *operational* concept of scientific explanation, according to which explaining a phenomenon amounts to proposing a mechanism able to produce it. The pioneers' constructivist style of knowledge finds here an evolved expression. This postulate does more than presenting the equation in which is grounded the first explorers' creative way of doing science – to know scientifically is to "build" ("construct", "invent", "fabricate"). It introduces a specific "post-classical" procedure of description, which integrates and enhances the heuristic attitude of the early research on autonomy. This descriptive solution can be seen as particularly appropriate to deal with autonomous systems, since it cannot be affected by their unpredictability. Requiring models able not to predict, but to generate natural processes, the principle locates the focus of the scientific research not anymore on actual, but on possible natural behaviours. It discards the classical demand of predicting and controlling nature, and proposes a constructive explanation which is destined to be revocable. In the School of Santiago's literature, it is associated to the heuristic imperative of evaluating the progressive character of the explanation, that is, establishing if the mechanism that the explanation proposes is able to produce other phenomena belonging to the same domain. If the explanation does not result progressive, it has to be substituted with a more generative one; otherwise, it has to be re-tested. This re-proposes through procedural terms the main epistemological imperative of the constructivist logic of description: do not impose to reality a structure which belongs to human creativity.

"If I'd like to provide a scientific explanation of cognition, I must provide a mechanism that generates an appropriate (animal or human) behaviour, and other behaviours which are susceptible to be observed in the same domain." (H. Maturana)[15]

This operational concept of scientific explanation (with its new emphasis on *possible* rather than *actual* behaviours; on *generation* rather than *prevision* of phenomena; on *construction* rather than *representation* of objects) produces the epistemological shift that we consider appropriate for the

[11] See the references given at Footnotes 9 and 10. There is not enough room here to describe in detail the constructivist conception of science. A good description, focused on the substitution of the classical idea of objectivity with that of pertinence or viability, can be found in [24, 45].

[12] This is particularly true for the cybernetic pioneer groups, which, in their scientific explorations, tried to systematically couple artificial and theoretical construction. In this sense, they can be considered as the founders and pioneer implementers of the current synthetic approach.

[13] The School of Santiago (in particular Maturana) denied or minimized the influence of the pioneer groups' productions on its theory of life, although this influence is difficult to underestimate at many levels. Autopoietic biology exhibits very strong convergences, both theoretical and heuristic, with the pioneers' production, and the contacts of its authors with the pioneers (e.g. Heinz von Foerster) and their literature are well-known, as well as acknowledged by Maturana and Varela.

[14] See [34], p. 80

[15] See [34], p. 81

legitimization of the synthetic approach.[16] Its pertinence for the epistemological grounding of this methodology becomes flagrant if we consider its autopoietic application to the description of life.

On the basis of this epistemological principle, autopoietic biology formulated a procedurally new definition of life, which, instead of listing the main features of living systems, provides the theoretical draw of a dynamical mechanism able to produce their phenomenology [35]. Maturana and Varela called their definition of life "synthetic", to distinguish it from the traditional "analytic" definitions presenting detailed lists of properties. And, as it is manifest in their work, they grounded this synthetic definition in an emergentist logic, made up of the same thesis nowadays adopted by the supporters of the synthetic approach: (1) although living phenomenology rests on physico-chemical components, it cannot be found in them; (2) it emerges from the interaction of these components, the unitary system they compose and its *ambience*; (3) resulting from this interplay, living phenomenology exceed science's power of calculation and prevision [14, 32].

Affinity with the current synthetic approach is so strong that it authorizes us to hypothesize a direct link: autopoietic biology could be a historical source of methodological inspiration for the current synthetic approach, although its proponents usually do not refer to it as such. One of the rare case in which autopoietic biology is quoted in their literature is provided by the works of a nascent trend of synthetic biology (SB) called "chemical autopoiesis" [32], in which autopoietic biology figures explicitly as a theoretical source, and implicitly as a methodological one. This emerging branch is strongly engaged in overcoming the dominant bioengineer approach, and tries to answer to crucial scientific questions on life and cognition through the chemical implementation of the autopoietic definition of the minimal living system [33].

For these kinds of attempts, as to every synthetic approach expressed in SB or AI, the autopoietic operational concept of the scientific explanation offers a general epistemological legitimization, but with an imperative clause. To produce genuine insights, the artificial models of living and cognitive systems must display the same organisation of natural living and cognitive systems (see below).

II. The second autopoietic principle we consider pertinent for the epistemological groundings of the synthetic approach is indeed a theoretical postulate; but, as we will propose, in the context of the sciences of the artificial it assumes a significant epistemological value.

The basic content of this theoretical principle is the distinction between two notions, i.e. *organization* and *structure*. Simplifying the original autopoietic formulation, we can put it as follows:

-the *organisation* of a living system is its relational frame, that is, the network of functional relations which define the system as a unity of components;

-the *structure* of a living system is its materialization, given by the actual components and their interconnections [35].

This distinction is not a theoretical novelty introduced by Maturana and Varela. It was progressively elaborated by the early researchers on living and cognitive autonomy. A first complete formulation can be attributed to Jean Piaget [40], who realized an integrative elaboration of the pioneers' studies which, as showed elsewhere [14], seems to have strongly influenced Maturana and Varela's production. Piaget proposed this conceptual distinction as the theoretical key to the comprehension of biological systems as dynamical, since it corresponds to the distinction between the invariant and the variant aspects of their dynamics. As he remarked, living systems can be considered dynamical systems endowed with a peculiarity: all their elementary components permanently change, while systems, as relational unities of components, remain. This, as Piaget pointed out, can be affirmed at both the ontogenetic and the phylogenetic levels. The relational unity is what remains unchanged not only in the permanent flux of physico-chemical component typical of biological organisms, but also during the ontogenetic transformations which can make a living system unrecognisable from one observation to the next. Moreover, this relational unity is transmitted through reproduction and remains unchanged generation after generation. Indeed, this relational unity is the invariant of the biological dynamics and therefore the lowest common denominator of living systems. To distinguish this invariant relational frame from the changeable materializations of living systems, and to determine its configuration, amounts to isolating an element which defines the class of dynamical systems belonging to the biological domain.

Here lies the relevance of the distinction between organization and structure, which is at least double. Firstly, this distinction opens the possibility of giving an operational explication of life, that is, the chance, exploited by autopoietic biology, to define a mechanism able to generate the living dynamics.[17] Secondly, this distinction generates significant implications both for theoretical biology and the epistemology of the sciences of the artificial. Synthetically: (a) In principle the materialization of living systems can be manifold;[18] (b) An artificial model of living systems, which is built with different materials and therefore endowed with a different embodiment, can be considered belonging to the class of living systems if it shares their organization; (c) To obtain pertinent insights for the study of life, the sciences of the artificial must produce models of living systems which display the same organisation as natural living systems.[19] Given the autopoietic equation between life and

[16] As pointed out by some of the pioneers and of their descendants, this notion of explanation is not an absolute novelty, but has some philosophical precursors, such as Thomas Hobbes and Giambattista Vico [23].

[17] As Maturana and Varela pointed out [35], this must be a mechanism able to produce the living organizational invariance through permanent structural variation.

[18] The only clause is that the materialization of a system must allow its organizational invariance [2].

[19] This is the pertinence of insights which come from the exploration of systems belonging to the same class of living systems. The issue of establishing if the sciences of the artificial could produce systems endowed with the organisation of living and cognitive systems deserves a specific treatment. Here we would like to merely suggest that this can be done with different degrees of approximation, and according to different theoretical choices, in order to obtain insights whose pertinence is relative to these choices and has many possible degrees. If research wants to adopt autopoietic biology not only as a methodological, but also as a theoretical framework, then it has to take into account that autopoiesis defines in detail only the organisation of the minimal living cell. Indeed, Maturana and Varela's theory describes very vaguely the organisation of superior organisms. For organisms endowed with the nervous system, autopoiesis provides only the constructivist draw of a circular organisation which connects "sensorium" and "motorium" [35].

cognition, (a), (b) and (c) has to be considered valid for cognitive systems too.

III. The two autopoietic principles presented above produce two criteria which can be useful to evaluate the applications of the synthetic approach to the study of life and cognition. To sum up: in order to produce pertinent insights, the artificial models of living and cognitive systems must

1. have the same organisation of living and cognitive systems, although they can have a different structure;
2. embed a mechanism able to produce the living and cognitive phenomena they intend to explain.

These criteria, with the epistemological framework proposed, guide our group's application of the synthetic approach to the study of emotions – phenomena lately reintegrated within the proper phenomenology of cognition.[20] In the following section we will try to point out the theoretical and methodological contributions of this approach to the understanding of emotions.

2 MODELLING *EMERGENT MINDS*' EMOTIONS

The generative loop between robotics and psychology

It is worth noticing that the application of the synthetic approach to emotional phenomena is not theoretically neutral. It requires researchers to develop and implement a very specific conception of emotions – indeed, a promising one. Besides being quite original in comparison to the classic philosophical and scientific views of emotions, this theoretical perspective fits well with some of the most recent developments in cognitive science and neurophysiology. Moreover, it shows a strong affinity with some new theoretical trends of developmental psychology, to the extent that a science of the artificial like robotics seems able to provide to psychological theories on emotional development a pertinent and fruitful experimental test-bed.

In robotics the rigorous adoption of the synthetic methodology amounts to departing from those AI approaches that exploit symbolic architectures (i.e. architectures grounded in rule-based symbol systems) in order to generate emotions by specific modules devoted to their computation, and to make emotions perceived by introspection [7]. In this case the result is a computationalist view of emotions which re-propose a conceptual element typical of classical western philosophical view. According to this approach, emotions are conceived and built as private events, which arise within the intra-individual space as internal states resulting from internal evaluations. They are basically an individual creation, accessible primarily to the subject of emotional experience and then to other subjects, through the agent expression.

This aspect of the classical theory of emotions is rejected by the synthetic methodology, whose emergentist framework requires developing different approaches to the artificial generation of emotions. More specifically, it requires researchers to construct emotions not anymore as integral part of the agent architecture, but as arising from the interactions of elements which in themselves do not constitute a internal emotional machinery. Indeed, according to the emergentist background seen above, emotional behaviours must be modelled as emerging from the interactions of elements which do not belong only to the agent, but also the environment. This approach, developed by both symbolic and embodied tradition of AI [7, 42], finds a radical application in Valentino Braitenberg's fundamental text *Vehicles* [6]. In Braitenberg's drawings agents display architectures which contain only sensorimotor correlations and produce emotional behaviours through the interaction with some perceivable aspects of their environment. As pointed out elsewhere [7], this design of emotions has the merit of expressing very clearly the main characteristics of the synthetic approach, but has also the default of bringing them to their very limit, and risks resulting unproductive. Braitenberg draws architectures deprived of any elements to which the production of emotions, as known in natural systems, can be related. He models emotions as pure epiphenomena "in the eye of the observer", and therefore misses the possibility of studying the role played by emotional processes in natural and artificial agents' interactions with the environment. More productive applications of the synthetic approach tend instead to explore how emotional behaviours arise from the interaction of "underlying mechanisms" inspired by the scientific knowledge of natural emotional processes. This procedure does not implement in robotic architectures the notions used to describe emotions, but those used to describe lower level processes, and it tests these implementations in agents interacting with the environment.[21]

In both these versions, the synthetic approach supports a view of emotions which discards the elements belonging to their classical philosophical conception. It declines the idea of emotions as private events arising within an individual and separable organisation. It locates emotions' generation not in the agents in themselves, but in their relation of coupling with the environment. It finds the pertinent unit for the study and the modelling of emotional processes in the whole system-environment unit – not an individual, but a *relational* unit, and primarily an *inter-individual* unit.

This idea, which in general is not explicitly formulated, connects the synthetic approach to emotional phenomena with a recent trend in the theory of emotions and, again, to the early research on self-organisation. Indeed, this rising trend is developed by researchers in different fields of natural and human sciences who share a strong interest for the theory of self-organisation [19, 31]. They inherited from the early sciences of autonomy the thesis that self-organising systems, such as living and cognitive systems, can experience dynamics of correlation which transform them in sub-units of higher level adaptive units. This idea grounded the pioneers' view of inter-subjectivity as the correlation of individuals in inter-individual units. To sum up: the mutual interaction between cognitive autonomous systems, for example human individuals, produces the co-dependence of their somatic and neural networks' dynamics, and couples their self-determined cognitive and affective behaviours at many levels.

Concepts such as Piaget's "inter-individual regulation" [40], autopoietic "behavioral coupling" [35], first Paskian and then Varelian "conversational unit" [39, 49] were introduced to express this idea. They can be seen as conceptual ancestors of the arising self-organisational theories of emotions and emotional development. The hypothesis that these theories develop is indeed one of the main implication of the self-organisational conception of inter-subjectivity: "emotions are social and [...] the body proper to emotions is the social body" [19]. In this picture emotions are not individual, but common –

[20] This re-integration is due to the recent attempts to "re-introduce the body in cognitive science" [50, 38].

[21] Typically this kind of design uses homeostatic control and neural networks modelled at different levels of abstraction [7].

"common works" [19]. They are aspects of an inter-individual process of coordination of intentions and actions in which social individual beings are involved from their birth and which, by affecting the most basic levels of their bodily organisation, deeply shapes their interaction with the environment. This thesis contrasts sharply with the computationalist view of emotions. For example, it excludes the possibility that the basic operation by which we accede to others' emotions can be found in a rational evaluation relying on the analysis of their expressions. It affirms that, before any rational calculation, we experience others' emotional processes, since we are essentially involved in these dynamics at the very basic levels of our biological and cognitive organisation [15]. There is no room here for the detailed explanation this view deserves. But it is worth noting that this kind of theory of emotions, besides converging with that implicit in the synthetic approach, is increasingly accepted and developed. It seems to express a basic element of inter-subjectivity which has lately been detected and explored by different branches of cognitive sciences.

Since the late Nineties, neuroscience has been studying the neural mechanisms underlying inter-subjective emotional processes – Vittorio Gallese, one of the discoverers of "mirror neurons", calls them "mirror mechanisms" [22].[22] The evidence pointed out by these studies is that the observation of one agent's emotional expression activates in the observer the neural pattern underlying this expression. This implies that in recurrent inter-subjective interactions there is such a co-dependence of the "respective" emotional dynamics that it requires us not to think of them as individual, but as intrinsically inter-individual processes.

This idea of an "inter-individual emotionality" converges with some of the last and most interesting theoretical advancements in cognitive sciences and philosophy of mind. They consist in new designs of the "embodied mind", i.e. the cognitive architecture by which researchers try to overcome the Cartesian dichotomy inherited by computational cognitive science. The current proponents of the *embodied mind* are engaged in a theoretical operation that philosophy of mind calls "cognitive extension" [51]. It is directed to redefine the boundaries of mind in line with this insight: if mind is the area where cognitive processes take place, then it cannot be relegated within the intra-individual space. According to the most widespread version of this thesis – a version labelled with the concept of "the extended mind" – the region of the mind overcomes the limits of "skull and skin" to include pertinent elements of the environment, i.e. elements which are external to individual body and brain, and which are necessary to the accomplishment of the agent's cognitive processes [13]. But there is also a more interesting version of the same thesis, which transforms the idea of the "cognitive extension". Its proponents give to the problem of "the embodiment of mind" not an anatomic, but a dynamic solution. They ground the mind not in the nervous system's anatomy, but in the regulative dynamics by which it links the body to the environment. Their "extension" of mind does not overcome spatially "skull and skin" [13]; instead, it re-defines the architecture of the cognitive mind. It describes the mind as the structure of coupling which inter-connects the dynamics of nervous system, body, environment and other organisms – other selves. In other words, it designs the mind not as a spatial – "extended" [13]– object, but as the dynamical co-determination which couples self and others' somatic and neural networks between them and to their environment. In this picture mind is essentially this: the emerging unit of self, others and environment's co-determination, which produces the cognition of the others and of the world. This view has been developed rigorously by a current trend of embodied cognitive sciences called "radical embodiment" [12, 14], which explicitly includes in its genealogy the early research on autonomy and autopoietic biology. Its concept of mind, often called "radically embodied mind", expresses with the tools of dynamical systems theory the theoretical content that the notion of "extended mind" tries to express too: cognitive processes – emotional processes included – do not belong to an individual, but to a *relational* unit – for social beings primarily the "self-other" *inter-individual* unit.

This emergentist and inter-individual concept of mind offers a theoretical expression not only to the synthetic approach to the study of emotions, but also to a new trend of current psychology of emotional development [36]. More specifically, this notion offers to them a conceptual link, which nowadays seems able to connect robotics and psychology in a circuit of exchanges susceptible to improve the scientific understanding of emotional development.

The specificity of this trend of psychology of emotional development is that it elaborates the idea of the inter-individual character of these developmental processes. A remarkable example is offered by the theory of attachment [11], in which the child and the caretaker(s) are not presented as independently defined individuals who undertake recurrent interactions. Instead, they are described as the polarities of a "dyad": an inter-individual system whose components are involved in a dynamics of co-determination which shapes the child's way of cognitively and affectively interact with his (social) environment – and re-shapes the caregiver's way. This inter-individual view of emotional development is typical not only of the original version of the theory of attachment [5], but also of its critical and revised versions [30]. Some of these latter are formulated in self-organisational terms, [48] and offer to synthetic approach and developmental psychology, besides strong theoretical affinities, a partially shared scientific genealogy.

This convergence between the synthetic approach to the study of emotions and psychology of emotional development is the premise of the collaboration of our epigenetic robotics' group with psychologists. The main goal is structuring an interdisciplinary exploration of emotional development and, in particular, of the mechanisms underlying the creation and the evolution of attachment bonds. This co-exploration rests on the joint, interdisciplinary design of robotic experimental scenarios, which are inspired by psychological insights and aim to produce feedback useful for psychology of emotional development and attachment.

Concretely, this collaboration is producing a series of studies focused on attachment bonds linking a "baby" robot and one (or more) human caregiver(s). Several crucial aspects of attachment are under inquiry: the development of different attachment profiles, the influence of these different profiles on exploratory behaviours, the role of attachment bonds in the development of sensorimotor associations, the development of attachment bonds in presence of multiple caregivers... [10, 26-29] A good premise for the pertinence of these robotics experiments for psychology is in our attempt to respect the two criteria presented above: (1)

[22] The notion of "mirror mechanisms" is introduced by Gallese to avoid the vague use of the notion of "mirror neurons" criticised in the current debate [18].

to give to the robots, as much as possible, a cognitive organisation which reproduces that of the natural cognitive systems they intend to model; (2) to make the robots generate, in interaction with humans, the phenomenology described as typical of attachment in specialists' literature. We are trying to satisfy criterion (1) through the implementation of artificial nervous systems which are not bio-mimetic, but are based on a perception-action organisation.[23] In these architectures, nothing plays the function of a internal emotional machinery – there are only mechanisms which allow the robots to modify their behaviours according to the interaction with the human user. These mechanisms, described in detail elsewhere, can satisfy criterion (2), since they are able to produce the fundamental aspects of attachment, such as the active research of the presence and the attention of the caregiver, distress in situation of separation, changes in attachment profile and exploratory behaviours in function of the behaviour of the caregiver etc... What this kind of architectures tries to implement is the thesis of the inter-individual character of emotional and attachment processes, which are not built as integral part of the robots, but as the product of their interactions with the users within a dynamical environment.

We believe the results we are obtaining are useful for both robotics and psychology, and moreover for the interdisciplinary unit they constitute. For robotics they consist primarily in significant steps towards the design of robots capable of an emotional development which make them adaptable to their specific user(s) and therefore more apt to interactive roles, such as the role of companions. With respect to psychology, these studies seem very useful for psychologists to evaluate and improve its theories of emotional development by testing what usually cannot be tested because of ethical constraints (e.g. the production of negative attachment bonds) or limits of intelligibility (e.g. the mechanisms underlying different attachment bonds). Finally, this collaboration heralds the development of a two-way interdisciplinary enterprise able to improve the scientific understanding of emotional development and, more generally, human cognition and mind. These studies allow us not only to test and possibly improve (or reformulate) the idea of inter-individual nature of emotional and cognitive processes, but also to offer our implementation and testing of this notion to new fields of research, such as the psychology of human-machine (human-robot) interaction.

We believe that this kind of cooperation – even if currently it is only at a starting point – can give a picture of what a science integrative of the synthetic approach could be: an interplay between sciences of artificial and sciences of the natural which makes productive the constructivist equation between *scientific facts* and *artefacts*.

CONCLUSION

In this article we have dealt with the issue of warranting a positive transmission of knowledge from the science of the artificial to the natural sciences of life and cognition. More specifically, we have tried to show how it is possible to warrant the pertinence of the "synthetic methodology" – "understanding by building" – for the scientific research on living and cognitive phenomena. Our contribution has not been only speculative, but has integrated an epistemological proposal for the legitimization of the synthetic approach with a concrete example of its application in the scientific practice. In Section 1 we proposed for the epistemological groundings of the synthetic approach the constructivist epistemological framework defined by the early research on self-organisation. In particular, we have presented two principles able to offer an epistemological basis to this approach. They are principles of intelligibility extracted from autopoietic biology, which respectively define: (a) an operational concept of scientific explanation and (b) the conceptual and methodological distinction between organisation and structure. From these two principles we have derived two criteria useful to evaluate the applications of the synthetic approach. In Section 2 we presented a concrete application of this approach, which, through the support of the proposed epistemological framework, explores a class of phenomena among the most problematic for science – emotions. We have shown that the synthetic approach, when applied rigorously, produces an inter-individual view of emotional processes which heralds a generative two-way transmission of knowledge between epigenetic robotics and the psychology of emotional development.

[23] See Footnote 15. See [10, 26-29]

ACKNOWLEDGMENTS

This research is supported by the European Commission as part of the FEELIX GROWING project (http://www.feelix-growing.org) under contract FP6 IST-045169. The views expressed in this paper are those of the authors, and not necessarily those of the consortium.

REFERENCES

[1] D. Baird. *Thing knowledge*. University of California Press (2004).
[2] L. Bich and L. Damiano. Theoretical and artificial construction of the living. *Origins of Life and Evolution of Biospheres*, 34:459-464 (2007).
[3] G. Bocchi and M. Ceruti (Eds.). *La sfida della complessità.* Mondadori (2007).
[4] M. A. Boden. *Mind as machine.* Oxford University Press (2006).
[5] J. Bowlby. *A secure base.* Basic Books (1988).
[6] V. Braintenberg. *Vehicles.* MIT (1984)
[7] L. Cañamero. Emotion understanding from the perspective of autonomous robots research. *Neural networks*, 18:445-455 (2005).
[8] L. Cañamero. Playing the emotion game with Feelix. What can a Lego robot tell us about emotion? In: K. Dautenhahn, A.H. Bond, L. Cañamero, B. Edmonds. *Socially Intelligent Agents*. Kluwer (2002)
[9] L. Cañamero and R. Aylett (Eds.). *Expressive characters for social interaction*. John Benjamins (2008).
[10] L. Cañamero, A. Blanchard and J. Nadel. Attachment bonds for human-like robots. *International Journal of Humanoid Robotics*, 3:1-20 (2003).
[11] J. Cassidy and P. R. Shaver. (Eds.). *Handbook of attachment*. The Guildford Press (1999).
[12] A. Clark. An embodied cognitive science? *Trends in Cognitive Science*, 3, 9:345-351 (1999)
[13] A. Clark and D. J. Chalmers. The Extended Mind. *Analysis*, 58:10-23 (1988).
[14] L. Damiano, *Unità in dialogo*. Mondadori (2009).
[15] L. Damiano and P. Dumouchel, Epigenetic embodiment. In L. Cañamero, P. Y. Oudeyer and C. Balkenius, *Epigenetic Robotics*, Lund University Cognitive Studies, 146: 41-48 (2009).
[16] M. R. W. Dawson. From embodied cognitive science to synthetic psychology. In: *Proceedings of the First IEEE International*. (2002).
[17] M. R. W. Dawson. *Minds and machines.* Blackwell (2004).
[18] J. Dokic and J. Proust (Eds.). *Simulation and knowledge of action*. John Benjamins (2002).
[19] P. Dumouchel. *Émotions.* Institut Synthélabo (1995).
[20] P. Dumouchel. Sul modo d'esistenza degli strumenti scientifici: la conoscenza oggetto. *Il Protagora*, V, 12:327-332 (2008).
[21] H. Foerster, von, and G. W. Zopf. (Eds.). *Principles of Self-Organisation.* Pergamom (1962).
[22] V. Gallese. Intentional attunement. *Interdisciplines*. http://www.interdisciplines.org/mirror/papers/1

[23] E. Glasersfeld, von. *Radical constructivism*, The Falmer Press, London (1995).
[24] E. Glasersfeld, von. The radical constructivist view of science. In A Riegler (Ed.). The impact of radical constructivism on science. Special issue of Foundations of Science 6(1–3). (2001).
[25] E. Jantcsh. *The Self-organizing Universe*. Pergamom (1980).
[26] A. Hiolle, K. Bard and L. Cañamero. Assessing human responses to different robot attachment profiles. *Proceedings of the 18th International Symposium RHIC*, pp. 251-257 (2006).
[27] A. Hiolle and L. Cañamero. Why should you care? An arousal-based model of exploratory behavior for autonomous robots. In *Proceedings of the 11th International Conference on Artificial Life*, 242-248, MIT (2008).
[28] A. Hiolle and L. & Cañamero. Developing sensorimotor associations through attachment bonds. *In Proceedings of the 7th International Workshop on Epigenetic Robotics,* 45–52 (2007).
[29] A. Hiolle, L. Cañamero and A. Blanchard. Learning to interact with the caretaker: a developemental approach, *In Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interactions*, 425–-436 (2007).
[30] H. Keller. Attachment-Past and present. But what about the future? *Integr Psych Beav* 42: 406-415 (2008).
[31] M. D. Lewis and I. Granic (Eds.). *Emotion, development and self-organization*. Cambridge University Press (2002).
[32] P.L. Luisi. *The emergence of life*. Cambridge University Press (2006).
[33]P. L. Luisi. The synthetic approach in biology. Manuscript. (2010).
[34] H. Maturana. *Tutto ciò che è detto è detto da un osservatore*. In Thompson W. I. (Ed.). *Ecologia e autonomia*, 79-93, Feltrinelli (1988).
[35] H. Maturana and F. Varela. *The Three of Knowledge*. Shimbhala (1987).
[36] J. Nadel and D. Muir (Eds). *Emotional development*. Oxford University Press (2005).
[37] C. L. Neahaniv and K. Dautenhahn (Eds.). *Imitation and social learning in robots, humans and animals*. Cambridge University Press (2007).
[38] R. Nunez R. and W. J. Freeman. (Eds.). *Reclaming cognition*. Imprint (1999)
[39] G. Pask. The Natural History of Networks. In M. C. Yovits and S. Cameron (Eds.), *Self-organizing systems*. 232-263, Pergamom (1960).
[40] J. Piaget, *Biologie et connaissance*. Gallimard (1967).
[41] R. Pfeifer, M. Lungarella and O. Sporns. The synthetic approach to embodied cognition. In P. Calvo and T. Gomila (Eds.). *Handbook of cognitive science. An embodied approach.* Elsevier (2008).
[42] R. Pfeifer and C. Scheier. *Understanding intelligence*. MIT (2000).
[43] I. Prigogine and I. Stengers. *La nouvelle alliance*. Gallimard (1986)
[44] H. Radder (Ed.). *The philosophy of scientific experimentation*. University of Pittsburgh Press (2003).
[45] A. Riegler. (Ed.). The impact of radical constructivism on science. Special issue of Foundations of Science 6(1–3). (2001).
[46] H. A. Simon. *The sciences of the artificial*. MIT (1982).
[47] I. Stengers. *Cosmopolitiques*, vol. 6, La Découverte (2003).
[48] E. Tronick. *The neurobehavioural and socio-emotional development of infants and children*. Northon & Company (2007)
[49] F. Varela. *Principles of Biological Autonomy*.North-Holland (1979).
[50] F. Varela, E. Thompson and E. Rosch. *The embodied mind*. MIT (1991).
[51] R. A. Wilson and A. Clark. How to situate cognition. In P. Robbins and M, Aydede (Eds.). *The Cambridge handbook of situated cognition*. Cambridge University Press (2009)