RECEIVED: January 5, 2007, REVISED: April 12, 2007, ACCEPTED: May 12, 2007

Ascriptional and ‘genuine’ autonomy

Marieke Rohde[1,2] and John Stewart[1]

[1]COSTECH Université de Technologie de Compiègne, France

[2]Centre for Computational Neuroscience and Robotics (CCNR),
Department of Informatics, University of Sussex
Brighton, BN1 9QH, UK

Email:
m.rohde@sussex.ac.uk, john.stewart@utc.fr

Marieke Rohde is the corresponding author; please use the UK (CCNR) address for written correspondence. Telephone: 0044 1273 87-2948, Fax: 0044 1273 87-7873

Abstract: Knowing that human judgment can be fallible, we propose to distinguish the subjective *ascription* of a property, such as autonomy, from the *genuine fact* that an entity is characterised by a certain property, i.e., it *is* autonomous. In this paper, we take a closer look at this distinction and what it is grounded on, taking a constructivist stance that sees the scientist as an observing subject. We arrive at a notion of fortified ascription, in which knowledge and scientific study of generative mechanisms play an important role, and look at some models of autonomy in the light of this distinction.

Keywords: Autonomy, Observer, Ascription, Artificial Life, Modelling

1. Introduction

Maturana points out that “everything said is said by an observer” [20]. The well-known example of a skilful submarine pilot [22], who manoeuvres a submarine relying just on the readings of different kinds of meters, on the basis of which he decides which levers to pull, will help us illustrate this phrase. Maturana and Varela argue that this submarine pilot, who supposedly never in his life has left the submarine, can skilfully master the passage through a reef full of obstacles. However, he would not know what an obstacle is, or a reef, or even a submarine. These concepts can be used to describe and perceive the situation from the outside, for example by an observer standing on the seashore. The concepts the pilot himself will use to describe and perceive the situation will be different, and will probably rely on meter readings and levers, not on reefs and submarines. Maturana and Varela invoke this metaphor to illustrate how, in the scientific study of life, the biologist’s point of view differs from the organism’s point of view. It expresses a deep constructivist belief, which rejects the idea of an objective world out there, with a pre-given ontology of events, objects and facts that the organism aspires to represent. It is the organism that creates its objects and their meaning, in accordance with its needs, desires and the history of its sensorimotor engagement with the world.

Science, as an activity exercised by human organisms, is therefore not about real objects that exist in an observer independent reality either. Maturana and Varela are *not actually themselves adopting the organism’s point of view*, it is their very point that the view from within another organism is not attainable for an observer. A scientist’s experiential world is a product of his own conceptual space and the distinctions he decides to undertake, and they will necessarily impact on the results from and interpretation of scientific activity. This is immediately and obviously true for the operational distinction between the ascription of autonomy and the genuine reality of a system’s autonomy in the scientific study of autonomy. In this light, it seems justified to raise the question of the nature of this distinction. The recognition of our status as observers transforms our conceptual world in a way that blurs the boundaries of what we normally consider a belief and what we consider a fact. How to taxonomise judgments into mere ascriptions and recognitions of genuine truths seems problematic, or at least unclear, if the idea of the observer is taken seriously.

Basically, in this paper we argue that this distinction can indeed not be maintained in its strict sense. However, recognising that matters are not quite what they seem to be does not automatically imply that the distinction under investigation is not a useful distinction to be made. Acknowledging that, in many cases, it has served well to clarify matters, even just as a first approximation, we investigate what is at its core, and carefully try to set it onto new feet, in agreement with the idea of an observer science. In doing so, we will consider empirical evidence from experiments in minimal perceptual crossing and different approaches to explaining life, to closely examine the role that underlying mechanisms play for this distinction. Our analysis will be discussed as regards implications for the study of autonomy through artificial means and an observer science in general.

2. Ascriptional autonomy

In 1950, in his now classic paper "Computing machinery and intelligence" [29], Alan Turing proposed a scenario that he called the 'imitation game', but which is now more commonly been known as the 'Turing test': Will a computer, via a language interface, be able to trick a human being into thinking that it was indeed another person? It may be arguable whether Turing's original rather gentle formulation of the test has been met, i.e., that towards the end of the 20th century "an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning [a computer]". The 'real' goal, i.e., computers that can reliably talk with humans and just like humans has clearly not been achieved. However, the proposal to tell whether an artefact is intelligent by relying on a statistical measure of human judgment, even though it seems so obviously limited, still holds the status of a benchmark test; Turing's paper is a must-read for every first year student in cognitive science or artificial intelligence.

This success is due, in part at least, to a shortage of convincing alternative proposals to scientifically identify intelligence, a concept that we understand intuitively, but that seems so hard, if not impossible, to pin down in terms of necessary and sufficient conditions, and that seems beyond direct scientific measurement (there is no intelligence-o-meter). The same reasoning applies to other properties of cognitive/living systems, such as intentionality and autonomy, which, therefore, are in the same way subject to the critical methodological analysis performed in this paper[1]. The question of whether there can be ever more than merely an ascriptional judgment about the reality of autonomy is not just some kind of post-modern fantasy; it has, ever since the birth of AI, penetrated and influenced the practice and theory of mainstream cognitive science.

One of the most prominent and passionate proponents of a merely ascriptional approach (in the case of intentionality) is Daniel Dennett (e.g., [7]), who has introduced the notion of the *intentional stance*. To adopt the intentional stance is to successfully explain an entity's behaviour as rational action towards a goal. Machines, animals and humans differ with respect to the scope of applicability of the intentional stance to understand their behaviour. To the present day, living organisms, and, in particular, human beings are the only entities whose behaviour can be best and most generally be understood through the adoption of the intentional stance. According to Dennett, for robots, computers and other artefacts to become truly intentional, it is necessary to overcome this threshold of behaviour qualifying for the successful application of the intentional stance. It seems clear that in order to establish the nature of intentionality as it is laid down in Dennett's theory, a kind of Turing test would be necessary, to measure to what extend we can successfully ascribe intentionality to a system.

These views are to be taken very seriously. They are radical and consequent pursuits of the idea that there can never be more to intentionality (autonomy, intelligence) than what the human eye sees in a system's behaviour; and they have strongly impacted the agenda of artificial agent research, a scientific community that seems most in need of a formal criterion to identify such properties. However, they are debateable. Di Paolo [11], for instance, believes that large parts of the robotics research on emotion (e.g., MIT's 'Kismet' [5]) are misled in following these kinds of proposals. Instead of investigating the nature and origin of emotions, this sort of research concentrates on imitation or simulation. Even if an 'emotional robot' can be convincing, its emotional

[1] We do not wish to imply that all these concepts (i.e., intentionality, autonomy, intelligence, emotions…) are the same, or could be used interchangeably across contexts. Though there clearly are relations and interdependencies between these properties, their exact nature is to date not at all resolved or obvious. However, in the context of this paper we will treat them as interchangeable because our criticism of simulation/imitation based approaches, on the one hand, and (implicitly) realist approaches, on the other hand, applies to them in the same way.

space is externally defined. In these models, the meaning of implemented emotions like ‘joy’, ‘fear’ or ‘anger’ are completely arbitrary in their relation to the robot architecture (i.e. inverting/changing meanings of implemented emotions is of no consequence for the underlying mechanism and its functionality). By contrast, the emotional space of a living creature relates directly, in a meaningful way, to its internal states, its organisation and dynamics, and the consequences of an emotion-evoking event for the material processes that bring about an emotion. “A real animal […] can be trained to do lots of things, but never to treat a punishment as a reward.” [11]. Imitation based approaches to values and intentionality (Rohde & Di Paolo [25]) and autonomy (Di Paolo, this volume [9]) have been criticised analoguously.

Figure 1 here

Intuitively, the idea that there is more to autonomy than just successful imitation seems appealing. But this demand for ‘something more’ relies on the possible distinction between the ascription and the reality of autonomy. Such criticisms seem to rely on the idea of an observer independent and objective existence of an entity’s autonomy that can be distinguished from mistaken human ascription. The questions to be asked are twofold. There is a conceptual question: can these demands for more than just successful imitation be held up if one accepts that everything said is said by an observer? And the complementary instrumental question: How can this ‘more’ be fathomed out, if not with some kind of variant of a Turing test?

3. Beyond ascription at first sight: generative mechanisms

The aim of this section is to present a central thesis of this paper: contrary to the spirit of the ‘Turing-test’ and Dennett’s ‘intentional stance’, we consider that it is indeed possible to go beyond mere ascription. Our proposal, in general terms, is this: what can be done, is to elaborate a serious hypothesis concerning the *generative mechanism underlying the phenomenal appearances,* on which the first-blush subjective ascription (of ‘intelligence’, of ‘intentionality’, of ‘autonomy’ or whatever) is made (or not, as the case may be). We wish to emphasize that proposing a mechanism which could generate the phenomenon in question involves a substantial amount of scientific work. First, it must be shown that the hypothetical mechanism, if it works in the way proposed, would indeed generate the phenomenon. In addition, independent evidence should be provided to show that the various components of the mechanism do in fact exist. Ideally, the components should be measured quantitatively, and it should be shown that the phenomenon does emerge from the system as a whole given the measured numerical values of the parameters. Of course, even in the ideal case, this would not positively prove that the phenomenon really is generated by the mechanism in question; but in the Popperian spirit, if the above-mentioned conditions are to be met the hypothesis of the mechanism is eminently refutable, and if it turns out not be refuted it is worth taking seriously.

Figure 2 here

What can this do for us? Well, we consider that knowing the mechanism (or at least having a serious hypothesis) can alter our judgement as to whether the phenomenon ‘really is’ a case of autonomy (or whatever). It is clear that in many cases, knowledge of the mechanism will serve to disqualify the initial impression. Consider the example of a conjuring trick: when ‘we know how it is done’, in most cases the magical impression is dispelled and the phenomenon appears as a mere ‘illusion’ (this is, of course, why conjurors - amateur and professional alike - are generally reluctant to reveal the ‘trick’). Searle’s well known ‘Chinese room’ argument [27] can indeed be seen as an application of this disenchanting effect that knowledge of the mechanism can have in the context of the Turing test scenario.

Another and scientifically more significant example of the way in which knowledge of mechanisms can lead to a disqualification is provided by reductionist Molecular Biology with respect to the phenomenon of ‘Life’. François Jacob, a major molecular biologist remarkable for his perspicacity and historical awareness, explicitly acknowledges that: “Today, life is no longer an object of questioning in the laboratory” [17]. Henri Atlan, who has made major contributions to the theory of information in the context of complex systems, confirms the diagnosis. Quoting (approvingly) the Hungarian biochemist Szent-Györgi, he writes: ‘‘Life as such does not exist, no-one has ever seen it… The term ‘life’ does not mean anything, because no such thing exists”.[2] Atlan continues: “The object of biology is a physical and

[2] We will not comment on the dubious epistemology of this: no-one has ever seen a gravitational force, or an electron, but these are nevertheless perfectly ‘real’ scientific objects.

chemical object. From the moment when one starts doing biochemistry and biophysics and when one understands the physical and chemical mechanisms that account for the properties of living beings, life as such disappears! Today, a molecular biologist has no use for the word ‘life’ in his work. This means that biology studies an object, the object of its science, that is not life!” [1]. We thus have here a clear example of a case where the identification of a mechanism leads to the disqualification of the original phenomenon.

However, this brings us to a key point in our argument: we wish to emphasize that disillusionment is not *necessarily* the case; on the contrary, knowing the mechanism can sometimes actually serve to *fortify* an initial subjective impression. We can sometimes have the reaction: “Well, my initial impression was purely subjective, and I myself did not feel that it was very reliable; but if that is how it works, then the phenomenon may be ‘real’ after all”. A good example of this is the theory of autopoïesis: Maturana and Varela [21] explicitly put this theory forward as a *mechanism* which, if it functioned in the manner specified, would indeed generate the phenomenon of life. This can serve as a vigorous rejoinder to the impression resulting from reductionist Molecular Biology, according to which ‘life does not exist’[3].

4. Perceptual crossing and intentionality

To show how our proposal can work in practice, we shall now illustrate it by a recent experiment carried out by the ‘Perceptual Supplementation Group’ at the University of Compiègne. ‘Intentionality’ is an important aspect of autonomy; and the *recognition* of intentionality in another entity is a most interesting question. Auvray, Lenay and Stewart ([2], personal communication, 2006) have investigated the dynamics of human perceptual crossing in a minimal shared virtual environment, in a Turing-test-style situation. Each of two blindfolded human subjects could move a cursor left and right in a one-dimensional virtual space (the ends were joined so that the space was finite but had no boundaries). When the cursor encountered an object in

[3] In the Workshop at San Sebastian, a near-consensus emerged according to which autopoïesis is a necessary but not sufficient condition for “life”. “Constitutive autonomy” (autopoïesis) must be complemented by “interactive autonomy”: the system must use sensory feedback to guide its actions in such a way as to maintain the boundary conditions necessary for continued autopoïesis [4]. But this merely strengthens the general point: certain mechanisms can actually *establish* the reality of a phenomenon.

the virtual space, this caused a sensory return in the form of a tactile stimulation. For each subject, there were 3 objects (of equal size) in the virtual space (see figure 3):

1. The receptor field of the other subject, henceforth the ‘avatar’. Thus, when the two subjects’ avatars overlapped, both of them received an all-or-none tactile stimulation. This situation is termed ‘perceptual interaction’.
2. A fixed object, henceforth the ‘fixed lure’. The fixed object perceived by subject 1 was invisible for subject 2, and vice versa; the two fixed lures were in different positions.
3. A mobile object, henceforth the ‘mobile lure’. In order to ensure that this object and the avatar have similar objective trajectories of displacement, the mobile lure was attached by a virtual rigid link at a certain distance from the avatar.

The task for the subjects was to click on the mouse button when they considered that they were in the presence of the other human subject. It is, therefore, the button press that represents the ascription of autonomy and intentionality in this task. It should be noted that when subject 1 explored subject 2’s mobile lure, subject 2 did not receive any tactile feedback (and conversely). Moreover, the tactile stimulation caused by encounters with an object constituted the only sensory information received by the subjects; in particular, they had no information as to whether the other partner gave a click on the mouse-button.

Figure 3 here

The result of this experiment is that human subjects, without the need of training, were quite successful at solving the task: 70% of clicks were correct, i.e. followed a genuine perceptual interaction; 20% of clicks were due to encounters with the mobile lure, and only 10% to the fixed lure. This is rather remarkable, given the poor sensory information that this minimalist virtual environment provides, and despite the difficulty of distinguishing two identically moving entities, only one of which is the interacting[4].

[4] A further exploration into the dynamics and generative mechanisms of this experiment with simulation experiments in evolutionary robotics has been done by Di Paolo, Rohde and Iizuka [9].

How is this discrimination achieved? This is where we move on to a consideration of the underlying mechanisms. A closer look at the data reveals that there are two parts to the explanation. Firstly, subjects were able to distinguish between a fixed object and a mobile object. Stimulations due to encounters with the fixed lure were quite frequent, but only rarely induced a click. The mechanism for this was that subjects moved back and forth across all objects they encountered. If the pattern of stimulation was reversible - i.e. the stimulations always occurred at the same position - the object was perceived as ‘fixed’ and did not induce a click.

Secondly, subjects did not seem generally able to distinguish directly between the avatar and the mobile lure: the probabilities of a click following these two types of stimulation were virtually identical (of course this is an analysis made by an ‘external observer’ with full access to all the data: for the subjects themselves, the stimulations were identical whatever the type of object). The fact that 70% of the clicks were due to encounters with the avatar, and only 20% to encounters with the mobile lure was due to the fact that encounters with the avatar were much more frequent than encounters with the mobile lure. The mechanism for this lies in the dynamical properties of the interaction process as a whole. Simulations confirm that, if the strategy of subjects is to pass back and forth whenever an object is encountered, the basin of attraction around the situation of encounter is wider and deeper when the two avatars are in front of each other, than when one subject encounters the mobile lure of the other. Therefore, the solution to this second aspect of the task does not derive from individuals performing a deliberate perceptual discrimination between different patterns of sensory stimulation; rather, it emerges from the mutual perceptual activity of the subjects which is oriented towards each other.

This brings us to the crucial question: supposing that these are indeed the mechanisms that generate the phenomenon, ‘does it count’ as a ‘genuine’ recognition of intentionality? Interestingly, initial scientific opinion, at this stage of the debate, is divided. Some reactions are of the type: ‘Well, if that is all that is going on, it certainly does *not* correspond to what I mean by ‘intentionality’. These sceptics consider that a ‘mobile vs. fixed discrimination’, plus a non-deliberate feature of mutual dynamics, does not correspond to what they mean by the ‘perception of intentionality’. To spell this out, these sceptics address the question of intentionality in terms of a ‘Theory of Mind’, involving the attribution of mental states and propositional attitudes such as beliefs and desires. In this perspective, the question of the recognition of the other as an intentional subject [8] is considered as the result of a process of cognitive inference based on objectively determined behaviours [Premack 24, Gergely 13].

The rejoinder of Auvray et al. is framed in terms of dynamic systems theory of sensory-motor coupling; they suggest that all such processes involving the manipulation of symbolic representations may be simply superfluous, at least in the first instance. They argue that the recognition of mutual perception occurs ‘directly’, in the sense of Gibson [14]. They maintain that a recognition of mutual perception, in the concrete sense of *knowing-how* to find the other participant’s avatar, does occur in this experiment. Indeed, the difference between the mobile object and the other’s avatar is that the latter’s behaviour changes when crossing my avatar. Furthermore, this change in behaviour corresponds to a perceptual activity oriented towards me. In other words, the active perceptual activities attract each other; just as in everyday situations there is an attraction to the situation where two people catch each other’s eye. The fact that the other’s reaction to my presence corresponds to a perceptual activity oriented towards me can be considered as a recognition of perceptual interaction, in the sense of a practical and concrete recognition, it is neither a deductive reasoning, nor an analysis of the stimulation in itself.

The point we want to make here is not so much to argue in favour of one or other interpretation; rather, we wish to remark that a discussion based on generative mechanisms leads to a genuine clarification of what we mean by ‘intentionality’ (or, more generally, by ‘autonomy’). A discussion of the described findings merely on the basis of surface behaviour would probably not even have generated the controversy, or if it would, it would not have provided the vocabulary and concepts to distinguish the basis of divergent beliefs.

5. Implications for Artificial Autonomy

Computational models can play an important role in the study of generative mechanisms, and are therefore in principle very interesting as regards our proposal of informed mechanism based ascription. An impressive example is the computational model by Hinton and Nowlan [15], which succeeded in finally putting the ‘Baldwin effect’ in evolutionary theory beyond doubt, and above suspicion that it may be just some form of closet-Lamarckianism. A proposed mechanism that had not been perceived as convincing because it was counterintuitive and difficult to understand had been made credible with the help of a computational model. Even if we do not really understand a model at first, we know that the results of the simulation follow logically from our in-built premises, i.e., there is no magic involved[5]. As a consequence, the Baldwin effect has been integrated into the canon of evolutionary theory. Similarly, the computational model of autopoiesis by Varela, Maturana and Uribe [30] demonstrates how organizational unity can emerge from local distributed processes and a constantly changing material substrate. Computational models can generate proofs of concept against human prejudice and cognitive limitations, because they are guaranteed to be sound, and they can illustrate links between the behavioural and the mechanistic domain. This is what we see is their key advantage for the study of autonomy in order to arrive at 'scientifically informed ascriptions', as described in this paper.

At the same time, our argument also points towards possible pitfalls for the modeller. To model autonomy computationally or formally is to walk on a tightrope. We have to avoid a naïve objectivist reduction of autonomy to just a mechanism, or even to a computational model of a mechanism; and at the same time to remain wary of imitation based approaches that reduce autonomy to the behaviour of an artificial system and are focussed on evoking certain ascriptional responses in human observers. We will now examine some of the approaches taken by other contributors to this issue and analyse their models. To start with, we shall discuss the formalisations proposed by Bertschinger et al. [3] and Chemero and Turvey [6] with respect to our previous analysis.

Bertschinger et al. propose an information theoretic measure of the amount by which the behaviour of an organism is determined by inner processes rather than external influences as a measure of interactive autonomy. As regards our postulate about the importance of studying the generative mechanisms, the first thing to mark about Bertschinger et al.'s model is that it relies on an *a priori* distinction between organism and environment and focuses on the surface behaviour, a limitation that the authors themselves are aware of by pointing out that they do not account for constitutive autonomy. To this extent, their framework is independent from actual physical

[5] See Di Paolo, Noble & Bullock [12] for a more detailed discussion on the role of "simulation models as opaque thought experiments" in science.

realization, the embodiment of an entity. To take the authors' example formalization of a glider in the game of life, there are many conceivable systems that would fit the FSA description of its behaviour, that we would not for a moment consider calling autonomous (whether or not we want to do so with a glider), even if we could call their behaviour equally independent from the environment. What is interesting about a glider is how its form and motion emerge from local rules that specify how grid cells are being switched on or off, but do not directly specify the glider's behaviour, and this aspect seems to get lost in the formalization. Or, to put it differently, few people would be impressed seeing a glider glide around, collide and decompose in the game of life, unless they knew about the rules of the game of life. If we are right with our observations about the importance of the (dis)enchanting role that comprehension of the mechanism can play, such surface descriptions leave out a very important aspect of autonomy, because they are fully independent of internal mechanisms. This is, of course, not to say that this measure, which is intuitively very appealing, is not potentially very useful or important for the study of autonomy, as part of a larger explanatory framework that includes aspects of generative mechanisms.

Chemero and Turvey introduce a graph theoretic (GT) formalisation of Rosen's idea of living organisms being closed under efficient causation, which formalises aspects of internal mechanistic organisation of the system and its coupling to the environment. Using the concept of hypersets and recursive graph structures, it points out similarities and differences between this proposed basis of life and autonomy, and others, such as autopoiesis and Kaufmann's autocatalytic cycles. It serves the authors as a tool to clarify conceptual properties and similarities of different mechanistic approaches to life and autonomy that are not apparent at first sight. To this extent, it is in accordance with our postulate about formal models as tools for better understanding generative mechanisms, because the analogies observed after formalisation are not at all obvious in the stand-alone description of these theories.

Without making any judgments about the value of each of these proposals, an important methodological difference between them is that Chemero and Turvey resist the temptation to reduce autonomy to either a mechanistic or a behavioural description. Their GT model is used to describe a *mechanistic* concept, i.e., closure under efficient causation, which is then *put into relation to autonomy*. Bertschinger et al., by contrast, propose their measure as a direct indicator of interactive autonomy, which is a reduction (the behaviourist kind), and therefore makes it to some extent vulnerable to our criticism of leaving aside issues of generative mechanisms. If, however, they had used their measure as an indicator of behavioural independence to then assess how it relates to autonomy, the fact that it leaves out issues of generative mechanisms would not have posed a problem.

This is not to question that simplifying reductionist definitions of autonomy can be useful in many situations. For a robot engineer, “autonomy” may be a useful term to refer, for instance, to the capacity of a robot to move without being remote controlled or with its own power supply (as it is necessary for, e.g., a mars rover) and there is nothing to object about such pragmatic definitions in many research contexts. The problems appear if autonomy itself is the subject of one’s studies, because one will lose the richness of the strong and widely applicable concept of autonomy when starting with a reductionist definition. By contrast, taking a similarly pragmatic approach to other concepts, mechanistic concepts, such as ‘behavioural independence’, to then investigate their relation to the phenomenality of autonomy and its generative mechanisms, one gets the best of both worlds.

Another contribution to this issue will help us to further illustrate our point: Di Paolo and Iizuka [9] present two simulated robotic models that are similar in their surface behaviour, but different in internal organisation. Both agents can be described as changing between two different modes of behaviour, and the times at which these changes occur depend on both internal and external factors. In one of these agents, the internal factor is realised by a built-in oscillating sub-module, in the other agent, the two modes of behaviour are associated with two different homeostatic boxes (or hyperboxes) in the state space of the neural controller. Therefore, in the first agent, change of ‘behavioural preference’ and motor behaviour are generated by functionally and structurally separate modules; in the second agent, however, these two behavioural aspects are profoundly intertwined, and it is argued that this second model is a better model of autonomy, because “autonomy is not something that a system does, it is a property of how the system is organized and re-organizes itself so as to channel its functionalities towards newly generated intentions” and can therefore not derive from a homuncular and fully functionally detached ‘choice module’. Importantly, even if the authors talk about autonomy, the two synthetic models are not presented as artificially autonomous agents, but as models of one aspect of autonomous behaviour, i.e., goal generating activity. They are presented to appeal to the reader’s common sense about the kind of generative mechanisms that convincingly realise this kind of behaviour as opposed to others (those that have a goal changing module) that do not. The point is that, even in this simplified scenario, the solution that has a ‘random goal generator’ feels like ‘cheating’, and that other possible mechanisms exist to generate the same behaviour, but which we are intuitively more comfortable with. This is a direct example of how knowledge about generative mechanisms can influence and inform our ascriptional judgments, and how simulation experiments can be used as tools to make matters of generative mechanism clear.

As the state space is continuous and high dimensional in both cases, it is not obvious how these systems would be formalised in Bertschinger et al.’s framework, but it seems clear that it would assign them equal or similar levels of autonomy, because both of them have external and internal factors determining the switching between behavioural modes, and are therefore partially independent from the environment in their behaviour, whilst depending partially on their inner state. This illustrates our argument that issues about generative mechanisms, as those at the heart of the authors’ concern, are left outside in a merely superficial behavioural description. Note that, even though this is a point about internal organisation, it does not refer to issues of constitutive autonomy, which Bertschinger et al. have deliberately left out, because Di Paolo and Iizuka’s model is not self-constitutive either, they do not address issues of self-constitution.

Similarly, Ikegami and Suzuki’s contribution [16] on homeodynamics that investigates the nature of the link between homeostasis and self-movement is crucially concerned with matters of generative mechanisms. If autonomy was just about surface appearance, what would be the point of investigating this link, rather than simply building it in? Again, it is the profound intertwinedness of functions, the fact that, in both their models, one and the same mechanism gives rise to homeostatic self-preservation and goal-oriented motion that enchants about their model, not the simple chemo/thermotactic behaviour the model agents exhibit.

6. Conclusion

To many researchers, a scientific approach to autonomy that is purely based on ascription, in a Turing test-like situation, seems insufficient. However, taking seriously Maturana’s insight that everything said is said by an observer, the question we pose in this paper is: Can there ever be more than a merely ascriptional judgment about whether something is autonomous? And what exactly does this more consist in, if not an objectivist and observer independent truth? Discussing examples from experiments in perceptual crossing to the theory of autopoïesis, we have argued that knowing the mechanism behind an apparently autonomous behaviour can alter our spontaneous ascriptional judgment. In some cases, such knowledge can sometimes clearly weaken our ascriptional judgment; but this is not necessarily the case, and sometimes knowledge of the mechanism can actually strengthen a positive ascriptional judgement. We propose this as a new basis for the traditional distinction between the 'ascriptional' and the 'genuine': We put forward the claim that an *ascription based on acquaintance with the underlying mechanism generating the behaviour* is a stronger form of ascriptional judgment than *naïve ascription based on observation of the behaviour alone*.

We also acknowledge that the effect of knowing the mechanism can differ from subject to subject as regarding the direction of the effect on an ascriptional judgement. However, this is mitigated by the fact that science is a social activity. If the disagreements remains within the scope of a single paradigm, the normal process of Popperian refutation (or not) will lead to progress. If the disagreement occurs between incommensurable Kuhnian paradigms, then an element of subjective choice may remain; we will come back to this important point at the end. But in either case, the collective dimension means that choices are not merely idiosyncratic individual whims.

Figure 4 here

This discussion focussing on the scientist as a subject is not just an argument about the importance of generative mechanisms. It is also not just a point about avoiding reductionism. Indeed, it is about both these issues together, about their synthesis: The idea of more robust 'informed ascription' on the basis of knowledge about generative mechanisms is something that we ourselves were unclear about before working on the ideas presented in this paper, even though each of the issues individually have been of concern for a long time. Reductionism seems to be in no way a marginalised activity (see, e.g., Di Paolo, Rohde, De Jaegher, forthcoming [10]), and reflex-like swaying between behaviourist and objectivist accounts seems to contribute to this trend. If one registers that behaviourism or imitation-based approaches do not capture the essence of what autonomy really is, an understandable counter-reaction is to stress the importance of the generative mechanism, and the temptation is strong to then go on and proclaim a mechanism of autonomy as ‘the real thing’, which, in the end, is just a different type of reduction. On the other hand, if one finds a mechanistic model of autonomy unconvincing because it seems to presuppose an objectivist worldview and seems to appeal to an observer-independent reality of autonomy, the temptation is strong to discard it in favour of a pragmatic human judgment-based model, which leads to a Turing-test-style behaviourist reduction, because this seems to be the best that can be achieved. Only by taking a step back from exercising science and recognising the subjective scientific study of the mechanism as an extension of naïve behavioural analysis, which is not ontologically superior, but has, pragmatically, turned out to generate more robust knowledge or belief, one can assess computational models of autonomy more modestly, but more confidently at the same time: Simulation experiments and formal models are tools that helps one to understand, express and discuss aspects of autonomy and its generative mechanisms that are otherwise difficult to grasp.

An objectivist premise is necessary to keep up a strict divide between explanans and explanandum. As Kurthen points out, a “hermeneutic cognitive science” can, and in our opinion should, be both a “science of hermeneutic cognition” and a “hermeneutic science of cognition” [18]. Issues like the one raised in this paper will arise again and again, in particular wherever supposedly *a priori* postulates lead to seeming dilemmas or antinomies. We therefore encourage the twodirectional flow of information between epistemological and scientific debate. We registered that the desire to go beyond ‘naïve’ ascription holds the danger of falling back into an objectivist worldview. We could have just done away with it as a matter of talking, rather than a metaphysical commitment. But by analysing what is really at the core of this distinction, which is indeed a useful one to make, we learned about the power of knowledge of the mechanism with respect to our ascriptions. We can now apply this knowledge to the study of autonomy, explicitly asking the question of how and why a mechanism convinces.

The ‘bad’ news resulting from our analysis is that, by replacing the idea of ‘genuine’ autonomy with one of informed ascription, which can be established in a kind of studying-behaviour-and-mechanism-and-scientific-debate Turing test, we lose the hope for an absolute criterion, providing timeless necessary and sufficient conditions for autonomy. But we can legitimately take comfort from the fact that this was only ever a fool’s paradise; it is not a part of the human condition to be able to have absolute certainty. Those who believe otherwise are living in an illusion[6].

The good news is that, with the ‘fortified’ concept of ‘informed ascription’, the generative mechanism is shifted more into the centre of scientific debate. It is not just some kind of necessary, but contingent and negligible detail, as it is the case in simulation/imitation based approaches that delimit the relevant aspects of ascription to the behavioural surface. Synthesising or modelling a mechanism can play a very important role in understanding it, and in making it understood. And finally, taking the ‘hard’ case, when judgement depends on a paradigm choice, as we illustrated it through the examples of intentionality and life, there is after all a sort of intrinsic, poetic justice. When a scientist makes a paradigm choice, (s)he has to live in the world that (s)he has participated in bringing forth. Biologists who choose to live in a world brought about by the paradigm of genetic determinism condemn themselves to living in a disenchanted, lifeless world. By contributing to bringing about a world in which life, and possible other forms of mechanism-generated autonomy, exist, the scientists live in a world where autonomy does exist. As you make your bed, so you shall lie in it.

[6] Merlau-Ponty, perceptively, points out that although objectivism is an illusion, in some circumstances it is a ‘‘well-grounded illusion” [23]. In the case of science, objectivism (believing that your theory is a true reflection of an independent, pre-existing reality) is a collective illusion that comes about when a scientific theory has been consensually stabilized for a sufficient period. It is thus constructed; but for a constructivist, saying this does not discredit it. Aeroplanes are constructed; but this does not mean that they can be constructed just anyhow (if you believe differently, I would not get into an aeroplane that you had constructed).

References

[1] Atlan H. & Bousquet C. (1994). *Questions de vie*. Le Seuil, Paris.

[2] Auvray, M., Lenay, C., & Stewart, J. (2006). *The attribution of intentionality in a simulated environment: the case of minimalist devices.* In Tenth Meeting of the Association for the Scientific Study of Consciousnes, Oxford, UK, 23-26 June, 2006.

[3] Bertschinger, N., E. Olbrich, A. Nihat & J. Jost (forthcoming): *Autonomy: an information theoretic perspective.* This volume.

[4] Bourgine P. & Stewart J. (2004). *Autopoiesis and Cognition.* Artificial Life 10, 327-345.

[5] Breazel, C. (2001). *Affective interaction between humans and robots.* in: Kelemen, J., & Sosík, P. (eds.), *Advances in Artificial Life: Proceedings of the Sixth European Conference on Artificial Life* , Springer Verlag. pp. 582-591.

[6] Chemero, A. & M. Turvey (forthcoming): *Autonomy and Hypersets.* This volume.

[7] Dennett, D. (1989). *The intentional stance.* MIT Press, Cambridge MA.

[8] Di Paolo, E. A., M. Rohde and H. Iizuka (forthcoming). *Sensitivity to social contingency or stability of interaction? Modelling the dynamics of perceptual crossing*. submitted to *New Ideas in Psychology*. for a special issue on *Dynamics and Psychology*.

[9] Di Paolo, E. & H. Iizuka (forthcoming): *How (not) to model autonomous behaviour*. This volume.

[10] Di Paolo, E., M. Rohde & H. De Jaegher (forthcoming): *Horizons for the Enactive Mind: Values, Social Interaction, and Play.* To appear in Stewart, J., Gapenne, O. & Di Paolo, E. (eds). *Enaction: Towards a New. Paradigm for Cognitive Science*. Cambridge MA: MIT Press.

[11] Di Paolo, E. A., (2003). *Organismically-inspired robotics: Homeostatic adaptation and natural teleology beyond the closed sensorimotor loop.* , in: K. Murase & T. Asakura (Eds) *Dynamical Systems Approach to Embodiment and Sociality,* Advanced Knowledge International, Adelaide, Australia, pp 19 - 42.

[12] Di Paolo, E. A., Noble, J. & Bullock, S. (2000). *Simulation models as opaque thought experiments.* Artificial Life VII: The Seventh International Conference on the Simulation and Synthesis of Living Systems, Reed College, Portland, Oregon, USA, 1-6 August, 2000.

[13] Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences, 7*, 287-292.

[14] Gibson J.J. (1979). *The Ecological Approach to Visual Perception.* Houghton Mifflin Press, Boston.

[15] Hinton, G. & Nowlan, S. (1987): *How learning can guide evolution*. Complex Systems, 1. 495-502.

[16] Ikegami, T. and Suzuki, K (forthcoming). *From homeostatic to homeodynamic self*. This issue.

[17] Jacob F. (1987). *La statue intérieure*. Odile Jacob, Paris.

[18] Kurthen, M. (1994). *Hermeneutische Kognitionswissenschaft. Die Krise der Orthodoxie*. DJRE Verlag, Bonn

[19] Latour B. & Woolgar S. (1979). *Laboratory life: the social construction of scientific facts.* Sage, Beverly Hills.

[20] Maturana, H. (1978). *Kognition.* in: Hejl, P., W. Köck & G. Roth (Eds.): *Wahrnehmung und Kommunikation*. Frankfurt, Peter Lang. pp. 29-49.

[21] Maturana, H. & Varela, F. (1980). *Autopoiesis and cognition: the realization of the living.* Reidel, Boston.

[22] Maturana H., & Varela, F. (1987). *The tree of knowledge: The biological roots of human understanding*. Boston, Shambhala.

[23] Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Gallimard, Paris.

[24] Premack, D., & Premack, A. J. (1997). Infants attribute value ± to the goal-directed actions of self-propelled objects. *Journal of Cognitive Neuroscience*, *9*, 848-856.

[25] Rohde, M. & E. Di Paolo (2006): *'Value Signals' and Adaptation: An Exploration in Evolutionary Robotics*. Cognitive Science Research Paper 584, Department of Informatics, University of Sussex.

[26] Schrödinger E. (1944). *What is life?* Cambridge University Press.

[27] Searle, J. (1980)*Minds, brains, and programs.* Behavioral and Brain Sciences 3, 417-424.

[28] Stewart J. (2004). *La vie existe-t-elle?* Vuibert, Paris.

[29] Turing, A.M. (1950). *Computing machinery and intelligence.* Mind, 59, 433-460.

[30] Varela, F., Maturana, H & Uribe, R. (1974). Autopoiesis: the organization of living systems, its characterization and a model. Biosystems **5,** 187–196.

List of Figures

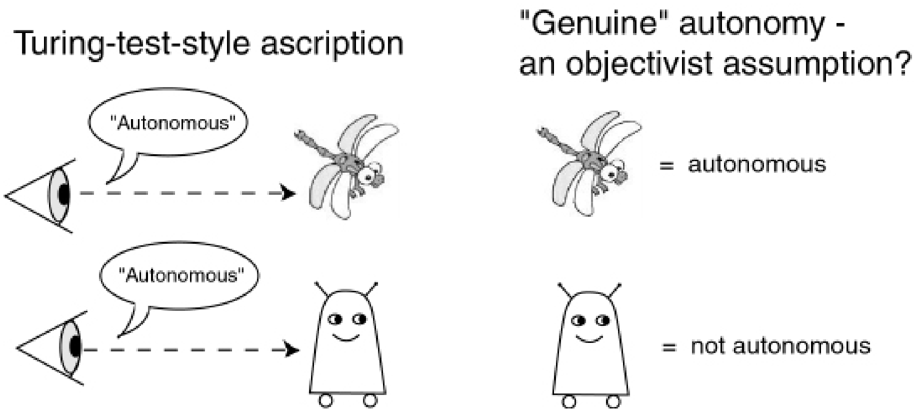

Fig. 1: Intuitively, imitation/simulation based approaches seem limited. But does asking for more than just Turing-test-style ascription presuppose an objectivist

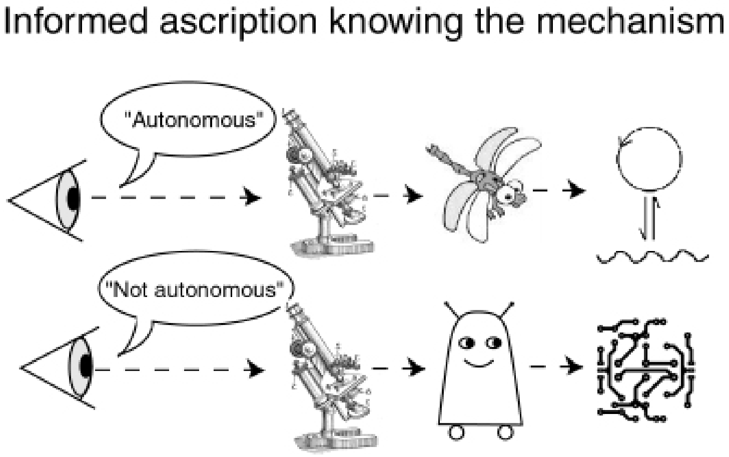

Fig. 2.: Sometimes, our ascriptional judgments are altered if we observe not just the macro behaviour of a system, but also scientifically investigate the mechanism generating this behaviour.

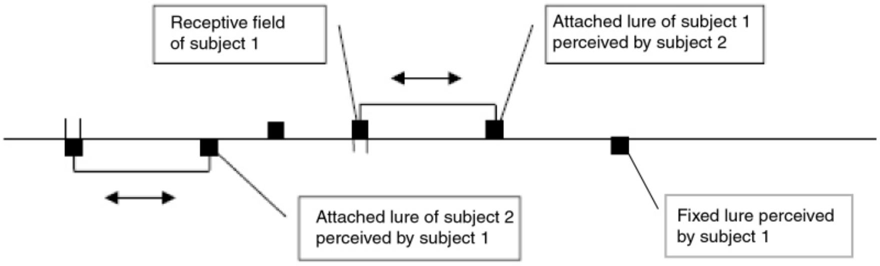
Receptive field of subject 1
Attached lure of subject 1 perceived by subject 2
Attached lure of subject 2 perceived by subject 1
Fixed lure perceived by subject 1

Fig. 3: A schematic diagram of the experimental minimal virtual environment.

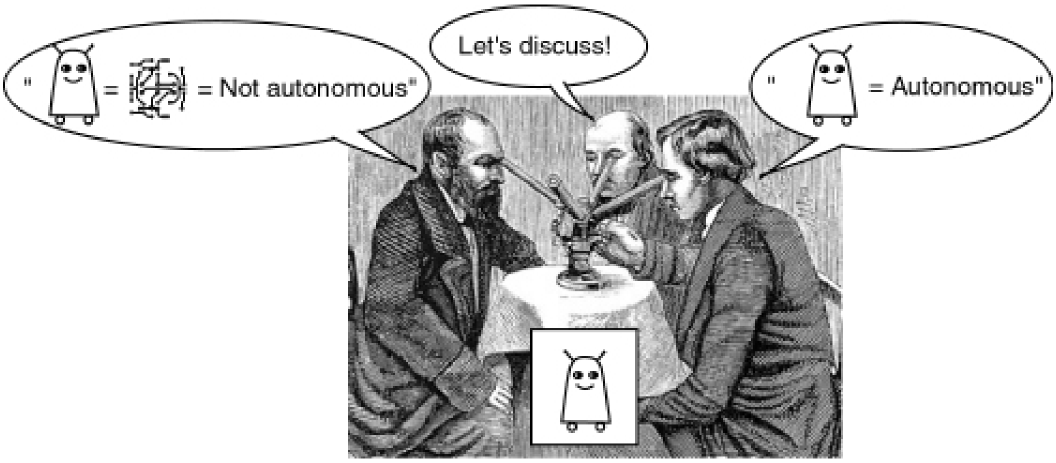

Fig. 4: In an observer science, disagreements are treated applying science-theoretically established standards, like, e.g., Popperian refutation.