Purposiveness and causal explanation.

Carlos Herrera Pérez
School of computing and mathematical sciences
Glasgow Caledonian University
Glasgow, Scotland
c.herrera@gcal.ac.uk

Major themes: Emergence, Emotions, Robotic and computational models of interaction and cognition

Abstract

This paper has three distinct parts. First we will explore some issues concerning the nature of behaviour and its explanation in terms of causes. We will argue in favour of recognising the multiplicity of causal explanation, in particular teleological explanation, rather than classifying all non-standard causes as emergent properties with no causal power. In the second part will summarise a theoretical account of the emotions, in which we try to recognise the different dimensions of emotions and in what sense they are causes of a holistic order. In the third part we will develop a methodology to implement robots inspired by the insights gained. We will illustrate this methodology or architecture with an experiment on the evolution of predator/prey Khepera robots.

Introduction: the explanation of behaviour.

One of the main questions to be resolved in Cognitive Science is in what form and at what level of abstraction should the explanation of behaviour in man and animals develop. Many scientists and philosophers have argued "that human behaviour, and for that matter the behaviour of animals and even living organisms in general [and artificial agents, we will argue], is in some way fundamentally different from the processes in nature which are studied by the natural sciences" [Taylor 1964]. Behaviour seems to have a quality of purposiveness or intrinsic meaning that the blind accidents of nature have not. Sometimes we qualify these behaviours as having a goal, as being directed towards a future state of affairs - in a sense we claim that the goal is the cause of the behaviour.

Naturalism rejects regarding the goal as a cause because the explanation of reality follows the cannon traditional since the Newtonian revolution in physics, especially through Hume's reflection on the nature of cause and our capacity to acknowledge it. In only a few worlds, whereas Aristotelian science finds different classes of necessary causes in nature (efficient, formal, material or final), naturalism and empiricism cannot accept such causes because all we perceive is the succession of two events A and B, and we have no way of perceiving “the necessary connection”. Therefore logical and necessary causes must be left aside in scientific explanation, in such a way that we are left with what can be called Humean cause: A causes B if we are used to see B happening every-time A happens. The fact that Newtonian science does not follow this model (we talk about gravity as a cause although it is not a precedent event or entity, and we have principles as inertia), and that recent developments in Physics (theory of relativity, dynamical systems, quantum theory) give a different picture of "reality", is not an obstacle to a commitment to Humean cause in the sciences of behaviour. The underlying metaphysical commitment is that it conceives reality as processes (whether physical or mental) that are constituted on a linear time, i.e., something has a causal power on the world if it *happening* has effects on other entities *happening after*[1]. Naturalism claims therefore that the scientific explanation of an event can only be sustained by causes that happened prior in time and/or structure. Something of this sort is asserted when behaviourism talks of drives and forces whereas cognitivism talks about mental representations - both accounts have to pose non-empirical entities to support this type of explanation.

The notion of teleological cause is still considered to introduce some sort of unreality in the picture, a magical or metaphysical entity that obscures human reason. These convictions are more often than not product of the tradition and folk understanding of science and metaphysics that a matter of empirical research. In our approach we are not committed to an explanation only by antecedent causes - the reason for this is that behaviour is empirically directed, and as such requires a teleological explanation. There is nothing unreal about goals, they are not some “isolatable and independent grounds of reality” - we cannot locate them in the physical space because they are neither events to happen nor entities to perceive, and nevertheless an explanation by goals is still an empirical claim. When we say someone as doing x, (for example fixing a bike) we are not only describing the situation but also giving an account of the cause of the behaviour in terms of its goal (to fix the bike), which it is not an antecedent entity because in the logic of time the bike can only be fixed in the future. As the description in terms of goals is also a

[1] Being = presence = happening

causal explanation (a teleological one), we are not talking of mere epiphenomena or global effects, but of the causes of the behaviour, and therefore, reality in a natural sense. The goal, despite not being antecedent in time, has a causal role to play in the behaviour, and therefore when we claim that a behaviour has this or that goal, it is not enough that this is an intelligible description. If I say that P is going to Manchester from Glasgow, I am not only describing the dynamics of the behaviour - it is different claim to say that the person is going to London, even though the dynamics between Glasgow and Manchester are the same. When we refer to the goal, the end point of the journey, as the cause of the behaviour, we are making a claim than can be empirically proven right or false. There must always some kind of interpretation, because the goal is not an entity we can perceive, but it would be wrong to claim that the goal is just in the eye of the observer.

This previous argument goes against the classification of properties as emergent, when that makes it impossible to distinguish whether these properties are mere side effects or have a causal role to play in the explanation of behaviour. For example, imagine a robot that removes objects from a field and needs to recharge a few times a day. We could say that the robot can only remove x objects per day, and this is an emergent property of the robot/environment interaction. We also say that the robot has a goal, and if such goal is not an antecedent condition (a representation), we say that the goal is an emergent property of the robot/environment interaction. And we also say that the behaviour itself is emergent, in the sense that it is dynamical and cannot be reduced to blind efficient causes. But the capacity of the robot to remove objects, the teleological cause or goal of the behaviour and the order of the behaviour itself are three very different categories with a different causal power on the behaviour to be explained. Under the category of emergent they lose what makes them different, their causal power, so they are regarded as in-the-eye-of-the-observer. We would miss the point that the goal is in a teleological sense cause of the behaviour, that capacity is an attribute of the agent, and the behaviour constitutes the natural relationship between agent and environment.

The Emotions

An interactivist analysis of emotional phenomena should therefore analyse the elements involved in the emotions, not only whether as 'real' or 'emergent', but attending to the causal role that such elements play in the phenomena questioned. We will therefore first expose the conclusions from extensive analysis of the phenomena involved in the emotions, and the different relationships with the emotion as a holistic phenomenon (empirical relationships which not necessarily are causal connections in the Humean or temporal sense). The phenomena analysed follows Frijda's structure (Frijda 1986), analysis of experience, which means a certain way of experiencing and coming to know the world (different from reflective experience). Analysis of emotional behaviour, as phenomena of interaction. Analysis of physiology either as antecedent or constituent elements of the emotion.

The first hypothesis stemming from the analysis of emotion is that emotions are functional, in the sense that they regulate behaviour often in an effective way (for example Darwin's principles of expression were: associated serviceable habits; antithesis; direct action of the nervous system, Darwin 1872). This will lead to the hypothesis that an evolutionary process should be able to find solutions that resemble emotional behaviour in its functional role. Even though some emergent behaviours might be classified as emotional, we believe a suitable architecture can sometimes be provided. In the experimental part we will provide an example of such architecture.

From the cognitive perspective, we can learn that the experience of emotions involve a certain kind of evaluation, a "hot cognition" in the sense that it evaluates a situation or object as baring importance for the agent, and this constitutes some kind of acknowledgement [Lyons 1986]. From an interactivist perspective this shows that we can only ascribe emotions in situations in which the agent can be affected, affect not just like an accident that shapes behaviour, but as an event which defines what kind of agent this is. For that we need recognising some sort of agent's nature, or in objective terms we can say that there are certain defining or essential properties of the agent in its relationship with the environment, that are at question in certain situations. World and agent are not symmetrical entities in the explanation of emotion, but there are certain conditions that apply to the agent's constitution. [Frijda 1986] calls these essential properties *concerns*, being the primary concern life. He claims that they are antecedents of emotion, thus the question of whether a situation will bare importance to such agent or not depends on the antecedent conditions[2]. In our analysis, concerns are not necessarily "a disposition that the subject carried with him prior to that moment in time", and that encounters an event resulting in an emotion, because concerns relate to defining properties of the agent, not temporal ones (see discussion on the role of fitness function in the conclusion section). Nevertheless, in our synthesis we may want to design such an agent, for example with body variables like energy level that is both a defining property and a disposition objectively identifiable in the time line.

From the analysis of behaviour we can note that emotional behaviour is expressive in the sense that it is relational. Relational here does not mean interactive, which is a property of all behaviours generally. Rather in the sense that "is behaviour that establishes or enhances, weakens or breaks, some form of contact with some aspect of the environment or that aims at doing so or is accessory in doing so." Consummatory behaviour or instrumental behaviour, for example, are interactive but not relational in this later sense.The importance of emotional behaviour is in which sense it modifies the actual relationship between agent and the world, not necessarily whether it aims at attaining a goal or not.

Emotional behaviour can be further analysed in the quality of the agent/world relationship and its dynamics. Emotional behaviour consists in changes in action readiness. This relational dynamics often have an agent-centred direction, approach or avoidance, for example, and action readiness is then called action tendency. Action readiness can often be inferred from antecedent physiological states, but action tendencies as such have a direction that is relational and dynamic.

Action readiness can be described as the underlying embodied and situated structure of a behaviour, which involves regulation, activation and often a direction. In fear, we experience a readiness to act in a certain way, the pulse increases, the adrenaline circulates making muscles ready to contract, our senses are alert regulating the input space. Our body shows as an antecedent condition the readiness for action, but it is not a sufficient description of emotion: it lacks orientation. Singer & Schachter's experiments show that the

[2] As (Taylor 1985) shows, humans are self-interpreting animals, our essence in constituted in the way we interpret what is worth for us, thus the essential properties of a human agent are not natural kinds, but narrative. Emotions like humiliation or shame do not spring from antecedent concerns and states of the world, they cannot be objectively matched but in the course of a narrative life, in which our own being is at question. We don't act bravely because that is our natural constitution, but because each action constitutes narratively who we are and we don't want

injection of chemical substances can change the action readiness towards new events in the subjects, but cannot define what kind of action tendency will emerge. Some patients just claim to have a strange feeling with no quality, while other subjects may react to a new event that can bear a remote importance to the subject in an emotional way.

Physiological states are states of action readiness and are symptoms of emotions. It would be rush to say that they are the efficient causes of the emotional reaction, because they alone cannot cause an emotion. We would like to provide a formal description that accounted for the causal role of physiological states of action readiness in the generation of behaviour. This description will be based on dynamical systems theory and the notion of global variable (Clark, 1997) - regarding action tendency as a dynamical process of the interaction between agent and environment. Global variables are antecedent conditions that somehow represent the overall dynamics. For example, we can say that the temperature of a liquid somehow summarises the whole dynamics of the liquid. If a global variable is controllable, that opens a space of solutions to the design of an agent that dynamically interacts with its environment. If the change in the direction of behaviour depends on some higher level properties of the dynamical interaction, and we may find a controllable global variable (or we can design it) for such dynamics, then to a certain extent the behaviour can be controlled, at least in its action readiness. Thus our attempt is to describe or design certain physiological states as global variables for the dynamical interaction in which the agent is involved. In our case the global variable will be an input node to the neural system that only depends on current perceptions - ideally these global variables should represent action readiness in a richer physiological way than electronic control.

We can summarise our theory of emotions in a few points:

1. Emotional situations are those that bare importance to the agents
2. These are relational situations, therefore dynamical, and some dynamical systems can be described and even controlled by global variables
3. Physiological states are global variables of emotional situations, and in that sense we can say they are "hot representations" of the situation and are precedent causes of a change in action readiness.

to be cowards.

4. The direction of the action, nevertheless, is an interactive phenomenon, and requires a teleological explanation such as being an action tendency.

Experiments

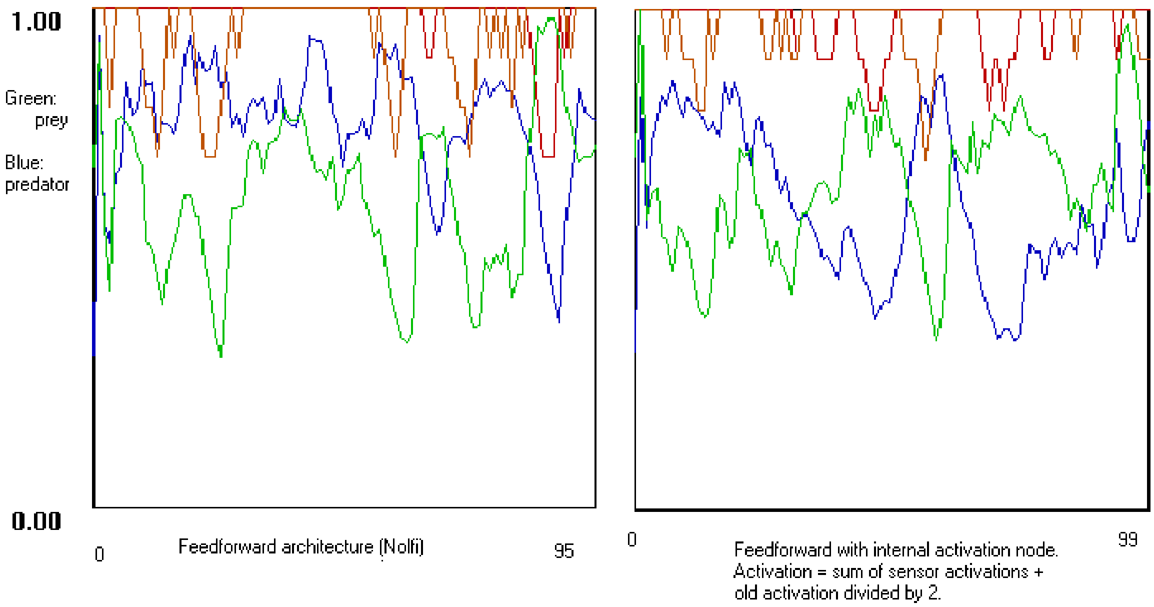
We will try to find an application for this model in the evolution of robots - but this does not preclude its application to, for example, learning in robots (although the relation between emotion and learning has not been in the scope of this paper). We have used the Evorobot toolkit and we have worked upon Floreano and Nolfi's experiments on the co-evolution of prey and predator Khepera robots controlled by feedforward neural networks.

> The predators have 8 infrared sensors, 5 simulated photoreceptors that can detect the black protuberance of the prey, and two motors that control the two wheels. This fitness function reward individuals with 1.0 point for each epoch in which predators are able to catch the prey by simply touching it. In additions predators can move with a maximum speed that is half of the normal speed. Preys have 8 infrared sensors and two motors that control the two wheels. This fitness function rewards individuals with 1.0 point for each epoch in which they are able to escape predators (i.e. reach the end of the epoch without being touched by predators). Both predators and prey are tested against 10 different individuals taken from the best competitors of the previous 10 generations. The environment consists of an arena of 47x47cms.(extracted from Nolfi 2001).

This experiment is analysed in depth in (Floreano and Nolfi 2000). Our goal is to enhance these robots with that their underlying architecture to support emotional behaviour in the sense described. The same methodology has been used for preys and predators, but here we will concentrate only in the prey case, as the results seem more interesting (although the improvement in functionality is similar). For this, we first need to predict at an abstract level a global variable the represents certain kinds of interaction between the robot and the world. Such prediction might be easy in some cases (for example, if we give the robot a rechargeable source of energy, the level of energy would seem to represent in the general case a global variable in the interaction between agent and world no matter what the interaction is). The observation of behaviour evolved from a previous architecture might enlighten what kind of interactions the agents will be involved in, in which situations a defining property of the agent is at question. For example, if we evolve a prey, we want to look for global variables that reflect the dynamics of dangerous interactions.

We have chosen an activation unit in the robot, some sort of energy level, that increases by adding up the activation of the IR sensors, and decreases to half each lifecycle. Giving the continuity of behaviour, the value of this variable should be high for the prey near the predator and the walls. If the prey does not follow walls but avoids them, then the activation unit will reflect mostly interaction with the predator. The value of the activation unit will increase to its highest rate when the prey is caught between one or two walls and the predator. So its seems that this variable should be a global variable of the dynamics of the interaction for the prey, even though we don't know yet what the interaction will be like. The interaction might be such that the prey's behaviour is such that it never gets close to walls. It is expected, nevertheless, that if evolution exploits this internal variable, it should engage in interactions for which this represents a global variable, having its arousal a causal effect on the action readiness of the agent.

First, if we compare the functionality of these robots against the previous architecture, we see an improvement for both preys and predators against previous architecture predators and preys (fig 1). The green line shows the performance of the prey, while the blue line the performance of the predator. In Nolfi's experiment, predators clearly outperform preys, while in the co-evolution with the "emotional" architecture, preys slightly outperform predators. This could be seen as a proof that our model adds functionality to the robot, at least in this case (the claim that this kind of enhancement would be functionally positive in any agent cannot be proven empirically, because it depends on the function of the robot and its natural form of interaction with its environment).

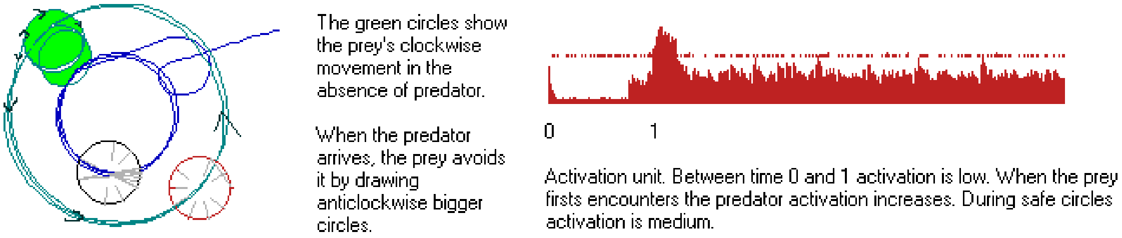


Feedforward architecture (Nolfi)

Feedforward with internal activation node. Activation = sum of sensor activations + old activation divided by 2.

But we should also be able to describe the phenomena observed in the same terms we have account for the emotions generally. That is, we should be able to describe the agent's behaviour as relational, with changes in action tendency or action readiness in situations that could be evaluated as relevant or important to the defining properties of the agent. The designed activation unit should be a global variable of the interaction in concern-relevant terms, and it should also let itself be described as representing a change in action readiness in the prey.

There are two forms of relational behaviour of the prey. In the first behaviour, in the absence of a predator (safe), the prey moves forward in clockwise circles (Fig 2). These are small circles, and therefore a predator attacks more likely from the front because the prey is open to stimulus by its movement to the front, sensors 4 and 5 are the most used during the whole interaction. Once the predator is nearby, the prey moves in the totally opposite direction, anticlockwise wider circles.



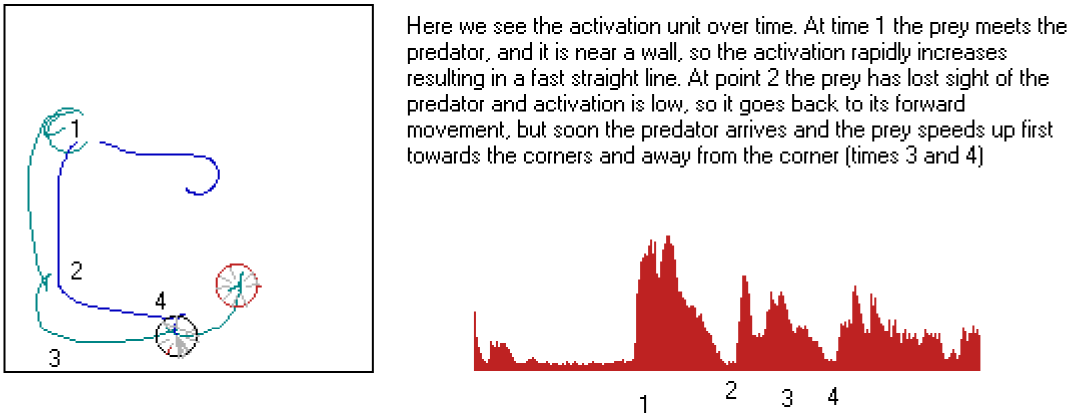
The green circles show the prey's clockwise movement in the absence of predator.

When the predator arrives, the prey avoids it by drawing anticlockwise bigger circles.

Activation unit. Between time 0 and 1 activation is low. When the prey firsts encounters the predator activation increases. During safe circles activation is medium.

Fig 2

At certain times, the prey finds itself between predator and the wall (fig 3). Nevertheless, the speed highly increases, and the prey "squeezes" between the predator and the wall, in a straight fast movement. The prey keeps running away in a straight line for some distance, far beyond the predators IR sensors range. During this change of behaviour from circular movement to straight, the activation unit has a very highly value until the prey is at a safe distance. The activation decreases to a low level and the prey turns back and approaches the predator in an instance of the safe interaction. When the predator enters again the sensor range, the prey engages in the same escape behaviour, as in figure 2 unless a new wall is found (which will normally be the case). The activation does not vanish when the situation is safe, but the agent is activated for a while in which the tendency is of escape. The activation decreases by half, so since its value was very high in danger, it takes a little while for the robot to a new change of action tendency, one that we can know evaluate as normal, not of special danger. This situation can be better described if we say that the activation unit increases when the prey is in danger, this activation represents a readiness to escape - in the interaction this behaviour shows itself as "squeezing" between wall and predator.


Here we see the activation unit over time. At time 1 the prey meets the predator, and it is near a wall, so the activation rapidly increases resulting in a fast straight line. At point 2 the prey has lost sight of the predator and activation is low, so it goes back to its forward movement, but soon the predator arrives and the prey speeds up first towards the corners and away from the corner (times 3 and 4)

Conclusion

Among other antecedents of this behaviour we have two designer's commitments, the underlying architecture and the fitness function. The underlying architecture is a cause of the behaviour, but it would be difficult to classify this cause. In an Aristotelian sense we could say that the underlying architecture that the designer provides the evolutionary process is the material cause of the agent behaviour. For living organisms, we could say that the material cause is the body. But here embodiment means more that the bundle of organs, skeleton and flesh, a certain embodiment is a perspective on the world. Only insofar living organisms inherit a body of a certain sort from its ancestors we can say that embodiment is an antecedent of an agent behaviour. But whilst in our experiment the architecture does not evolve, in real life there is no fixed embodiment but a creature develops in interaction with the environment. As Ziemke claims "today's situated robots still are radically different from living organisms... Mostly, this is due to the fact that artificial organisms are composed of mechanical parts (hardware) and control programs 'software'. The autonomy and subjectivity of living systems, on the other hand, emerges from the interaction of their components, i.e. autonomous cellular units." In our jargon, we could simply say that in living organisms embodiment is not necessarily an antecedent cause of behaviour, but the material cause of behaviour, therefore not necessarily an antecedent cause.

The fitness function, on the other hand, has a logical relationship with the behaviour of the system. When we evaluate a robot's behaviour high if it touches another robot, this later robot's behaviour as high if it is not touched, we are representing the logic of predator/prey behaviour. Following this metaphor, we could say that the logic of behaviour is in the evolutionary pressure over the species. In our artificial cases, we provide a fitness function that relates to one or various behaviours. In life, nevertheless, there is no such fitness function. Some would argue that the logic of natural behaviour is survival, but if we follow that line we would rather say that the logic of survival is life. So, in this sense, we could argue that evolutionary robotics depends always upon an antecedent logic, which cannot be the logic of life, because life "creates itself in living". Should we conclude that this imposes a limit on evolutionary robots? All will depend of the causal relationship that fitness functions will have with evolved behaviours, and the role of fitness function is an open question in current research (See Nolfi 2000).

.

1. Arnold M. B (1960) Emotion and personality(Vols I & II) New York, Columbia University Press.
2. Charland, Louis C. Reconciling Cognitive and Perceptual Theories of Emotion: A representation Proposal. *Philosophy of Science*, 64 pp 555-579. (1997) Damasio, A., *Descartes' error: Emotion, Reason and the Human Brain.* Picador, Cambridge, MA, USA(1990)
3. Clark, A. (1997) *Being There - Putting Brain, Body and World Together Again.* Cambridge, MA: MIT Press.
4. Darwin, Ch. 1872. The expression of emotions in man and animals. London: John Murray.
5. Frijda, N.H. (1986). *The emotions*. Cambridge University Press.
6. Frijda, N.H. & Swagerman J. (1987). Can Ccomputers Feel? Theory and Design of an Emotional System. *Cognition and Emotion 1987, 3.*
7. Frijda, N.H. & Moffat, D. (1993). A model of emotions and emotion communication. In *Proceedings of RO-MAN'93: 2nd IEEE international workshop on robot and human communication.* 29-34.
8. Frijda (1993). The place of aprraisal in emotion. In *Appraisal and Beyond*, Frijda. N. (Ed.) Lawrence Erlbaum Associates Ltd.
9. Lazarus, R.S. (1991). *Emotion and adaptation.* Oxford University Press.
10. Lyons, W. *Emotion*. Cambridge University Press. Cambridge, UK. (1980)
11. Nolfi S., Floreano D., Miglino O. & Mondada F. (1994) How to evolve autonomous robots: different approaches in evolutionary robotics. In R.A. Brooks & P. Maes (Eds.), *Proceedings of the Fourth International Conference on Artificial Life.* Cambridge, MA: MIT Press.
12. Nolfi S. and Floreano D. (2000). *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machine*s. Cambridge, MA: MIT Press/Bradford Books
13. Nolfi (2001), *Evorobot 1.1 User Manual* http://gral.ip.rm.cnr.it/nolfi
14. Penrose, R. *The emperor's new mind.* Oxford University Press (1989).

15.Pfeifer, R. (1988). Artificial intelligence models of emotion. In V. Hamilton, G.E. Bower, and N. Frijda (eds.). *Cognitive perspectives on emotion and motivation* (Proceedings of the NATO Advanced Research Workshop). Amsterdam: Kluwer, 287-320.

16.Picard, R. *Affective Computing*. MIT press (1997).

17.Schachter, S & Singer, J. (1962) Cognitive, social and physiological determinants of emotional state. Psychological Review, 65, 379-399.

18.Taylor, C. (1964) The explanation of behaviour. Routledge & Kegan Paul Ltd. London.

19.Taylor, C. (1985) *Human Agency and Language.* Self-interpreting animals, Pages 47-55. Cambridge University Press.

20.Zajonc, R. Feeling and thinking: preferences need no inferences. *American Psychologist*, 39, pp. 151-175. (1980).

.