THE DEVELOPMENT OF INTELLIGENT TECHNOLOGIES AND THE CONTEMPORARY PROBLEM OF ETHICAL RESPONSIBILITY

Cristinel UNGUREANU[1]

1. Lecturer, PhD, Dept. of Communication, Public Relations and Journalism, "Apollonia" University of Iaşi, Romania
Corresponding author: c_ungureanu2003@yahoo.com

Abstract

In order to find the way in the contemporary debate of robotic ethics, it is necessary to clarify what kind of ethics we are talking about. Is it human ethics applied to the use of technologies? Is it an ethics specific only to robots? Or is it a universal code common to all forms of advanced intelligent creatures? In this paper we try to argue that the decision for one ethics or another depends first of all on how we conceive the artificial mind in relation to the human mind. Consequently, our task is to analyze the two paradigms of Artificial Intelligence (AI), the Classic and the New AI, and see what kind of ethics is in each case at stake. In our discussion we encounter the following dilemma: if we build more intelligent robots, we can exert less control on them. The solution seems to be that the robot must posses some fundamental ethical principles on the basis of which it could develop particular ethical behaviors.

Keywords: *Classic AI, New AI, the Frame Problem, Heideggerian Robot, Robotic Ethics*

The most important challenge for the science of artificial intelligence nowadays is to create robots able to take over more of human tasks, not just mechanical, but also complex activities which presuppose an increased autonomy. This may be a source of fear for those who believe that the more sophisticated the robots become, the less we can control them. Thus, it seems that we do not want autonomous intelligent robots without ethical rules guiding their behavior. A lot of questions arise and must be taken seriously by scientists and philosophers. For example, should a search engine automatically collect data considered to be private? Should a robot hurry up to help a stranger even if this means a delay of its proper tasks? Should this be established by the owner of the robot? What should a driverless bus do in case of an ethical dilemma? Should it avoid killing a child by "pushing another spectator off a bridge onto the tracks?"[1] This robotophobia may be justified "against the prospects of building machines without conscience more powerful than we."[2]

In order to find the way in the contemporary debate of intelligent technologies ethics, it is necessary to clarify what kind of ethics we are talking here about. Is it human ethics applied to the use of technologies? Is it ethics specific only to robots? Or is it a universal code common to all forms of advanced intelligent creatures? There are three kinds of answers to this problem:

1. Robotic morality is illusory because only humans are to be blamed or to be praised. Robots have no autonomy.
2. The robotic ethical dilemmas highlight a pseudo-moral problem – the robots lack something that could make them fully moral agents.
3. The robotic ethics is a real problem and must be taken seriously.[3]

We accept the third variant and we try in this paper to argue that the decision for one robotic ethics or another depends first of all on how we conceive the artificial mind in relation to human mind. Consequently, our task is to analyze the two paradigms of Artificial Intelligence (AI), the Classic and the New AI, and see what kind of ethics is in each case at stake.

I.

In the Classic Cognitive Science, the mind, either human or artificial, is conceived as symbolic processing. According to Simon and Newell, the main promoters of this view, „A physical-symbol system has the necessary and sufficient means for general intelligent action. (…) any system that exhibits general intelligence will prove upon analysis to be a physical-symbol system." (1976, p. 111). We will not dwell here on the details of this paradigm; we will merely emphasize those aspects relevant for robotic

ethics. According to Fodor (1987), one of the main arguments in favor of the symbolic paradigm is that it vindicates the folk psychology – that practice of ascribing thoughts (intentions, desires, beliefs etc.) to other people. The basic assumptions of folk psychology are that thoughts have an intentional content (they are about something) and they are causally decisive determinants of our behavior. If thoughts are intentional entities, and intentionality is a relational property, how can thoughts be causally potent? Causality is a real process in the world, implying intrinsic properties of objects. It seems that Fodor accepts this restriction of causation when he speaks of "a metaphysical prejudice" (Fodor 1987, p. 139), which consists in the fact that the content *per se* cannot have causal power.[4] The vindication of folk psychology imposes that, beside the intentional content, thoughts have syntactic properties encoded in their physical form. Given this dual structure, it follows that thoughts are symbols. Hence, the mind incorporates symbols instantiated as such in the brain. The semantic properties map systematically onto the syntactic ones. That is why Fodor affirms that the symbolic paradigm is a nice possibility to act on the content of the symbols *via* their syntax (1987, p. 17). The symbolic-processing paradigm proves to be successful especially when it comes to high-level properties of cognition, such as systematicity, compositionality, planning, search of a good move in a game etc. (cf. Fodor & Pylyshyn 1988).

The syntax of mental states is a higher level physical property. However, it is not reducible to physicalist explanations. Fodor's idea is grounded on the functionalist thesis that claims the existence of a level of causality which, although physically realized, it cannot be explained by physical laws. For example, currency exchange cannot be explained on the basis of the physical properties of the banknotes, but taking into account specific concepts and laws (price, demand, central bank etc.). (Fodor 1974, 55-56; Rudder-Baker 1995, 133).

There are two features of the Classic Cognition very important for our discussion. As symbolic processing, the mind, either human or artificial, is a kind of software which works in the same way, whether it is instantiated in a biological brain, or it is implemented in silicon hardware. The embodied aspects, such as sensory-motor coupling, feedback loops, temporal rates etc. are irrelevant to the core operations of intelligence. Another feature of symbolic paradigm is that thinking does not operate directly in and upon the real things in the world, but through their representations. Of course, action takes place in the world, but its contact with objects takes place first at the moment of perception and second when the behavioral output is produced.[5] For this reason, the main task of the Classic AI researchers was to build up robots with sufficient symbolic knowledge to cope with various situations in the world. They believe that the difference between human and artificial mind lies in the amount of information it possess (for example, CYC, the robot built in the U.S. between 1984 and 1994, was endowed with a giant encyclopedia of knowledge about everyday situations, such as: the capital of Missouri is Jefferson City, in UK the cars run on the left side of the road and so on).[6]

II.

There are numerous critiques against this theory, because of its heteronomic consequences. John Searle, in his famous critique against Strong Artificial Intelligence, argues that symbols cannot account for intrinsic intentionality. The Chinese Room Argument (CRA) (Searle 1980) shows, first, that the syntactic operations are blind to semantics; the semantics is assumed in the programming by its designer or user. Hence, the second very important conclusion of CRA, developed by Searle in detail in his later work, is the distinction between intrinsic and derived properties.[7] The intentionality of a map is derived as it depends on the interpretation of an observer. The property of being snow on Himalaya is intrinsic because even if all interpreters would die, there still will be snow on Himalaya. Resorting to this distinction, Searle criticizes the claim that the brain is a digital

computer.[8] According to symbolic paradigm, the brain does operate with symbols in accordance with syntactic rules. But notions like symbol, syntax, program, bits etc. point always to interpretation. If it were true that, for example, syntax is intrinsic to the physical world, then, everything would instantiate a syntax, even the wall behind us would instantiate the program Word Star. (Searle 1992, pp. 208-209). Hence, two important critiques are addressed to the symbolic paradigm: homunculus fallacy and causal impotence of mental states. We will not insist here on the huge intrinsic/derivative intentionality debate. What interests us is the idea that symbolic processing is heteronomic (homuncular). In essence, Searle's thesis maintains that only human mind is characterized by intrinsic intentionality. Although a human mental state depends on the existence of a subject, it is not a derivative property, because "I have it, regardless of what anyone thinks about it." (Searle 1999, p. 93). Just as digestion is an intrinsic property of the digestive system, so the mind is an intrinsic property of the brain and should be studied as such (1992, p. 227-8).

Some authors point out that it is a little bit mysterious that only the biological brains emanate intentionality, but the silicon based devices do not. However, Searle accepts that a device as complex as the human brain can produce intrinsic intentionality (*ibid*., p. 92). Thus, it seems that the difference lies in the degree of complexity. Unfortunately, Searle does not clarify what that means. A symbolic device could be very complex, and still lacks the access by itself to intrinsic intentionality.

Another critique, the Frame Problem, is more effective in making evident the failure of the Classic AI. According to this view, the designer establishes what information the robot must receive. Robots are just input-output devices, designed to solve predefined problems, with a pregiven set of instructions, in a predetermined environment. The Frame Problem emphasizes not just that robots have insufficient knowledge for accomplishing their tasks, but also that they are unable to choose the relevant knowledge for that task (Dennett 1987, Dreyfus 2008). Robots act on the basis of a world model stored in their chips. But in a changing world, this model gets always out of date. According to Mackworth, the robot's world must be fully deterministic and observable; the robot must posses a perfect internal model of its infallible actions and of the deterministic world. The perception has the role of determining the initial world state. Knowing the world's laws of change, the robot builds up a plan in order to reach its goals (Mackworth 2011, p. 337).[9] For Mackworth, the robot of the Classic AI works with "closed eyes": " So, with its eyes closed, it can just do action A, then B, then C, then D, then E. If it happened to open its eyes again, it would realize «Oh, I did achieve my goal, great!» However, there is no need for it to open its eyes because it had a perfect internal model of these actions that have been performed, and they are deterministic and so the plan was guaranteed to succeed with no feedback from the world."

Researchers (Minsky, Schank and many others) realized that they could not implement a perfect model of the entire world, so they decided to implement frames of knowledge for each situation encountered by the robot (for example, Shank's "Restaurant Script").[10] Beside the fact that this is a very impoverished world (cf. Brooks 1991, pp. 398-399),[11] the robots are caught in a *regressus ad infinitum* fallacy: in order to choose the relevant frame of knowledge for a situation, they need another frame for establishing the relevance of that frame. (Dreyfus 2008, p. 333).

Given the Frame Problem, it is difficult to see how such robots could behave in an ethically acceptable manner. For example, we want our robot to follow the rule of keeping its promises. If, in a certain context, this would cause suffering to other persons, should the robot still obey the rule? How should we program the robot? Should we create a list with all the situations in which telling lies will cause suffering to the others? But this would be a list with an indefinite number of items.

Moreover, given the heteronomous critique, the ethics of these robots refers mainly to their programmers. Programmers are obliged to take

into account that these robots, in their activities, can encounter ethical dilemmas. Therefore, they must program robots so that they behave ethically correct. The discussion here will concern only what kind of ethics these programmers must adopt (consequentialist, normative etc. ethics) and how it could be implemented. Regarding the implementational aspects, they have to foresee every ethically problematic situation the robots may encounter, and design a detailed plan for each situation. Hence, the Frame Problem: the slightest environmental change leads to the collapse of their actions.

III.

The New AI, based on the theory of dynamical systems, starts from the principle of autonomy: intelligence means first and foremost the ability to cope with new, unexpected situations in a flexible manner, while maintaining the internal working at a survival threshold.[12] Anticipated by Dreyfus (1972, 1992), the change begins with Rodney Brooks' new bottom-up perspective on building robots. His intention is to bring intelligence back in the real world. In his view, a system is intelligent when it copes in real time with the environment without being influenced by the programmer. From a technical point of view, Brooks breaks down the tasks of the robot, called also Creatures, not horizontally by function (perception – reasoning – action) like the Classic Computation, but vertically by activity, such as moving, avoiding obstacles, identifying objects, collecting objects etc. The activities run in parallel and each individually connects sensing to action. The advantage of this approach, as Brooks says, consists in the fact "that it gives an incremental path from very simple systems to complex autonomous intelligent systems. At each step of the way, it is only necessary to build one small piece, and interface it to an existing, working, complete intelligence." (Brooks 1991, p. 403).

The current goals are not to mimic complex cognitive abilities, but to set the coordinates within which the robot develops its own actions, starting with the simplest ones. (*Ibid.*, p. 410). Perception is direct, not mediated by representations. Its result is not taken by another module in order to build up a detailed map of the environment (*Ibidem*, p. 404). Perception and action are simultaneous, they form a causal loop. The robot is connected to the world in a much simpler way, by an ongoing sensing of it. With this approach, the role of the environment has changed: from being the scene in which the intelligent act takes place, it becomes a decisive part of the act itself. The robot needs no internal world model. The world is its own model. (*Ibidem*, p. 406).[13] Hence, the Brooks' robots can handle the Frame Problem because there is no internal model which could get out of date given the continuous changing of the world (*Ibid.*, p. 417).

There is a lot of discussion whether Brooks solved the Frame Problem indeed. Dreyfus claims that Brooks' Creatures act in a fixed world and reacts to a small number of features that their receptors can pick up (Dreyfus 2008, p. 335). He acknowledges, however, that Brooks' theory makes a significant advance in avoiding the Frame Problem, but not in solving it (*ibid.*).

Dreyfus considers that Artificial Intelligence can handle the Frame Problem only if it adopts the Heideggerian philosophical ideas as basis for programming. Using the Heideggerian distinction between "readiness-to-hand" of equipments (when we are using them) and "presence-at-hand" of objects (when we reflect upon them), he affirms that the basic interaction with things is not intellectualized, as a whole tradition stemming from Descartes and continuing up to the Classic AI believed; the things are not first experienced as meaningless and then the robot confers them meaning (Descartes), or function (Searle). For Brooks' robot, the first relation to the world is not theoretical (to know the world), but practical (to act in the world). In skilful coping, the distinction between subject and world gets blurred. Things are seen as "solicitations" for action (Dreyfus 2008, pp. 348-9); in James Gibson's terms, they are "affordances". This means that they are neither fully subjective, nor fully objective. As Francisco Varela says,

"The key point is that such systems do not operate by representation. Instead of representing an independent world, they *enact* a world as a domain of distinctions that is inseparable from the structure embodied by the cognitive system." (Varela et al. 1991, p. 140). For example, Varela built a cellular automaton, Bittorio, and put it in a chemical soup of zeros and ones bits. Bittorio acts on the rule of assimilating a certain sequence of bits (such as 10010000). The internal structure changes and rebuilds itself depending on the chemical environment, not randomly, but when it encounters that sequence. Thus, Bittorio "makes" already a distinction in that soup of zeros and ones. The soup is the background where the structural coupling between Bittorio and its world (the sequence of 10010000s) emerge. Even at this level, Bittorio already enacts a world of significance, that of all strings it can assimilate. This world determines its actions, but the world is enacted and modified by Bittorio's actions. (Varela et al.1991, pp. 155-156).

Brooks' robots represent a big step forward in solving the Frame Problem, as they act directly in and upon the world. But a Heideggerian robot acts in a more adaptive way. Its readiness-to-hand is not a final function involving a predefined response, but a "flexible response" in accordance with the changing world (Dreyfus 2008, p. 340). This flexibility is more visible in the fact that the significance and the relevance of the next situation is determined by the experience of the current situation through a highly sensitive feedback mechanism. Thus, for such a robot the relevance could not be established "beforehand". (*Ibid.*).

Dreyfus identifies in the work of the neurobiologist Walter Freeman the right tools to solve the Frame Problem. Resorting to the concepts of dynamical systems theory, Freeman has shown how the rabbit's brain works when it comes to significance and relevance. Analyzing the brain when the rabbit perceives significant stimuli, the researcher will observe that strong bursts of energy cross the nervous system. These states tend toward an energy minimum, which in the terms of the dynamic theory is called attractor (Freeman 2000, ch. 4). The entire activity of the system can be seen as a transition from one attractor to another. (Dupuy 2001, p. 104). The totality of the states tending toward the same attractor forms the attractor's basin. The brain forms a basin of attraction for each significant class of inputs. Other experiences tend to integrate these basins of attraction, forming an attractor landscape. This landscape governs the selection of the appropriate behavioral answer. For example, the attractor responsible for looking for food it is not grounded in a representation of a carrot, but it is the sum of all past experiences of acting with carrots. (Dreyfus 2008, p. 351).

The rabbit interacts with the world in a fully dynamic, non-representational way. Its neurons fire according to the current state of the organism. If the rabbit is hungry, the neurons responsible for food are "primed to respond" (*ibid.*, p. 350). If the rabbit has just eaten, it is ready to mate, so the neurons looking for females are active and those looking for food are switched off. There is an "optimal body-environment gestalt" (*ibid.*, p. 343) which guides the action in a non-representational way. When the agent deviates from this optimum, tension appears which forces the organism to lower it.[14]

Complex cases, such as imagining counterfactual situations, planning vacations, arranging objects by their value etc., which, given the absence of direct environmental stimuli, require complex mental representations and serial processing. (Clark and Toribio 1994, pp. 419-420). Accepting this argument, some adepts of the dynamic explanation argue that the mental representation is not a fundamental mental cognitive performance, but it emerges from a dynamical substrate. (van Gelder 1997, p. 448).[15]

The Heideggerian robot can act in a flexible way, not only because of its highly sensitive feedback mechanism, but also because of its special relation to the world. Its "familiarity" with the world does not arise from storing piece-by-piece information. The world is not a database of explicit knowledge. To be a true world, it must function as a background of opportunities for action, structured primarily by previous experiences of that agent. The intentional relation to a specific situation emerges from this

background. It is the basis for choosing the right action in the right context. "Action" means here not the selection and the application of knowledge to solve a predefined problem in a pregiven situation, but the response structured according to internal principles of self-organization, given the significant environmental stimuli. Thus, the problem is not to identify the relevant knowledge for coping in a specific situation, but to select the right response. A Heideggerian robot responds adequately to significant stimuli because it is able to structure the world according to its own needs and capacities.

IV.

The New AI shows that autonomy is the very condition for intelligence. In this case, if robots act under the autonomy principle, should their ethics obey the same principle? Given that it is possible for robots to be more intelligent than us, is that recommendable? Also, because of the Frame Problem, we cannot implement ethical algorithms for every encountered situation.[16] However, if we want more intelligent robots, we have to exert less control on them. That is why the discussion about intelligent systems ethics becomes now very important. But how should an autonomous intelligent robot behave in an ethical acceptable way? Some authors consider that for ethical behavior to occur there is no need for a moral agent (Deborah Jouhnson 2006), or a free will (Allen et al. 2006). Other authors consider, on the contrary, that ethical behavior presupposes the existence of conscience (Storrs Hall 2011) or meta-intentionality (Dennett 1998).

We should begin to study the robotic ethics not by asking whether the robots could be fully moral agents. In this respect, authors like Deborah Johnson and John Sullins are right. Even we, humans, are not fully autonomous agents, because we take over by socialization many of the ideas of others (Sullins 2011, p. 156). The interesting problem is, given the examples at the beginning of this paper, whether robots are able to solve ethical dilemmas. The Classic AI solves the problem by foreseeing every situation and building specific algorithms. The Frame Problem shows that this is impossible. The New AI has made significant progress in building up autonomous robots and some authors argue that this theory already has the prerequisites to solve the frame Problem. So, the question is whether and how the New AI deals with ethical dilemmas.

Taking autonomy seriously and supposing that it is possible to build Heideggerian robots that solve the Frame Problem,[17] it seems that the designer does not have to implement ethical algorithms for every encountered ethical dilemma. In the New AI, the appropriate ethics for robots is grounded in their autonomy. Therefore, robots must be able to behave ethically in a flexible way, that is, they must be able to develop their own ethical behavior.

The solution seems to be that the robot possesses some fundamental principles on the basis of which it adopts further particular ethical behavior. The corresponding ethics could be the Kantian deontological ethics, because the categorical imperative could be that basis for particular ethical actions. This is easier said than done.[18] A robot may act under the economical rule "maximize your profit" and at the same time wish for this rule to become a universal law. But we, humans, probably won't accept this rule, because, in comparison to us, the robot could acquire a better ability to foresee the consequences of its actions. The rule "maximize your profit" would perhaps lead to our definitive impoverishment and slavery. Therefore, we have to choose whether we build not so smart robots, which we are able to control, or smart robots that do not harm us. Supposing that we choose the latter option, it is important to emphasize that the condition for robotic general intelligence, namely the ability to solve the Frame Problem, applies also to ethical thinking. The categorical imperative functions only when the robot is able to wish something for itself, and only afterwards something for the others. The lesson of the Frame Problem says that the robot is able to wish something for itself in a flexible way (that is, not in a predetermined block-world style) when its world functions as a background organized according to its own capacities and

past experiences with things. The Heideggerian robot learns from its own experiences and it builds its own perspective on the world.

In this case, the categorical imperative could be implemented as that internal condition, as the control parameter encoded by an attractor which, though it offers no detailed algorithm for each ethical action, modulates the interaction with the world; the categorical imperative modulates the formation of the robot's own ethical behavior. In other words, the designer will implement some ethical principles, such as the observance of the rights of any intelligent being, and then establishes these principles as control conditions for each action, allowing the robot to apply them in different situations. In the terms of the dynamic theory, these attractor-principles will have the largest basins of attraction, that is, all the states of the system will naturally tend to reach those attractors (equilibrium energy states). The robot will learn by experience that there are permitted and forbidden behaviors; it will learn to obey the laws of the community in which it lives; it will learn the respect for the human beings, and so on. Don't we do the same thing when we educate our children? Don't we educate them in the spirit of some values and then let them act depending on different contexts?

References

1. Allen, Colin, Wallach, Wendell, Smit, Iva (2006) "Why Machine Ethics?" in Anderson, Michael, Susan Leigh Anderson (eds.), 2011.
2. Anderson, Michael, Anderson, Susan, Leigh (eds.) (2011) *Machine Ethics*, Cambridge: Cambridge University Press.
3. Brooks, Rodney (1991) „Intelligence without Representation" in Haugeland (ed.), 1997.
4. Chemero, Anthony (2009) *Radical Embodied Cognitive Science*, Cambridge Mass.: MIT Press.'
5. Clark, Andy (1997) *Being There. Putting Brain, Body and World Together Again*, Cambridge, Mass.: MIT Press.
6. Clark, Andy, Toribio, Josefa (1994) „Doing Without Representing?", *Synthese* 101(3), pp. 401-431.
7. Dennett, Daniel (1987) "Cognitive Wheels: The Frame Problem in Artificial Intelligence", in Zenon Pylyshyn (ed.), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Norwood, NJ: Ablex.
8. Dennett, Daniel (1998) "When HAL Kills, Who's to Blame? Computer Ethics." in David Stork. *HAL's Legacy: 2001's Computer as Dream and Reality*, Cambridge Mass.: MIT Press.
9. Dreyfus, Hubert (1972) *What Computers Can't Do*, New York: Harper and Row.
10. Dreyfus, Hubert (1992) *What Computers Still Can't Do*, Cambridge Mass.: MIT Press.
11. Dreyfus, Hubert (2008) "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian" in Philip Husbands, Owen Holland, and Michael Wheeler (eds.), *The Mechanical Mind in History*. Cambridge, Mass.: MIT Press.
12. Dupuy, Jean-Pierre (2001) *The Mechanization of the Mind*, transl. by M. B. DeBevoise, Princeton: Princeton University Press.
13. Fodor, Jerry (1974) „Special Sciences" in Paul K. Moser, J. D. Trout (eds.), *Contemporary Materialism. A Reader*, London: Routledge, 1995.
14. Fodor, Jerry (1983) *The Modularity of Min,*. Cambridge, Mass.: MIT Press.
15. Fodor, Jerry (1987) *Psychosemantics. The Problem of Meaning in the Philosophy of Mind*, Cambridge, Mass.: MIT Press.
16. Freeman, Walter (2000) *How Brains make Up Their Minds*, New York: Columbia University Press.
17. Gibbs Jr., Raymond, (2007) *Embodiment and Cognitive Science*, Cambridge: Cambridge University Press.
18. Haugeland, John (1997) *Mind Design II. Philosophy, Psychology, Artificial Intelligence. Revised and enlarged edition*, Cambridge, Mass.: MIT Press.
19. Johnson, Deborah (2006) "Computer Systems: Moral Entities but Not Moral Agents" in Anderson, Michael, Susan Leigh Anderson (eds.). 2011.
20. Keijzer, Fred (2001) *Representation and Behavior*, Cambridge, Mass.: MIT Press.
21. Kelso, Scot (1995) *Dynamic Patterns*: *The Self-organization of the Development of Brain and Behavior*, Cambridge, Mass.: MIT Press.
22. Lakoff, George, Mark Johnson (1999) *Philosophy in the Flesh. The Embodied Mind and its Challenge to Western Thought*, New York: Basic Books, A Member of the Perseus Books Group.
23. Mackworth, Alan (2011) "Architectures and Ethics for Robots: Constraint Satisfaction as a Unitary Design Framework" in Anderson, Michael, Susan Leigh Anderson (eds.).
24. Marr, David (1982) *Vision. A Computational investigation into the human representation and processing of visual information*, Cambridge, Mass.: MIT Press, 2010.
25. Minsky, Marvin (1975) "A Framework for Representing Knowledge" in Patrick Henry Winston (ed.). *The Psychology of Computer Vision*, New York: McGraw-Hill, pp. 211-277.

26. Newell, Allen, Herbert Simon (1976) "Computer Science as Empirical Enquiry. Symbols and Search" in Margaret Boden (ed.), *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press, 1990.
27. Powers, Thomas (2006) "Prospects for a Kantian Machine" in Anderson, Michael, Susan Leigh Anderson (eds.).
28. Pylyshyn, Zenon (2003) *Seeing and Visualizing. It's Not What You Think*, Cambridge, Mass.: MIT Press.
29. Rudder-Baker, Lynne (1995) *Explaining Attitudes. A Practical Approach to the Mind*, Cambridge, Mass.: MIT Press.
30. Schank, Roger, Robert Abelson (1977) *Scripts, Plans, Goals and Understanding*, Hilsdale, NJ: Erlbaum.
31. Searle, John (1980) "Minds, Brains, and Programs" in *The Behavioural and Brain Sciences* 3: 417-424.
32. Searle, John (1990) "Is the Brain's Mind a Computer Program?" in *Scientific American*, pp. 26-37.
33. Searle, John (1992) *The Rediscovery of the Mind*, Cambridge, Mass.: MIT Press.
34. Searle, John (1999) *Mind, Language and Society. Doing Philosophy in the Real World*, London: Weidenfeld & Nicholson.
35. Storrs-Hall, John (2011) "Ethics for Machines" in Anderson, Michael, Susan Leigh Anderson (eds.).
36. Sullins, John (2011) "When Is a Robot a Moral Agent?" in Anderson, Michael, Susan Leigh Anderson (eds.).
37. Thompson, Evan (2007) *Mind in Life*, Cambridge, Mass.: Harvard University Press.
38. van Gelder, Thimothy (1997) „Dynamics and Cognition" in John Haugeland (ed.).
39. Wheeler, Michael (2005) *Reconstructing the Cognitive World. The Next Step*, Cambridge, Mass.: MIT Press.

Endnotes

1 Allen et all. 2006, p. 53.
2 These dilemmas are from Storrs-Hall 2011, p. 42.
3 See in Sullins 2011, p. 152.
4 Beside the metaphysical restriction, Fodor argues that the causal role of thoughts is thinner than the intentional content. For example, P and ÎÎP have the same content, but they may have different causal roles. Fodor 1987, p. 139.
5 During perception, after the physical input stimulates the sensorial interface of the cognitive system, the visual cortex computes this input in order to produce a three-dimensional representation from the two-dimensional projection of things on the retina; these representations are taken over then by other modules of the cognitive system – searching for solutions and building action plans. Cf., for example, Marr 1982, p. 23; Fodor 1983, pp. 102-103; Pylyshyn 2003, ch. 2.
6 Clark 1997, pp. 2-3.
7 See Searle 1992, pp. 78-80; Searle 1995, pp. 9-13; Searle 1999, p. 93-97.
8 See Searle 1990; Searle 1992, pp. 209-219.
9 Cf. also Wheeler 2005. For Classic AI, the environment is: "(i) a furnisher of problems for the agent to solve, (ii) a source of informational inputs to the mind (via sensing), and, most distinctively, (iii) a kind of stage on which sequences of preplanned actions (outputs of the faculty of reason) are simply executed)." (p. 45).
10 Cf. Minsky 1975, Schank and Abelson 1977.
11 This world is called "block-world" – the world created in the laboratory in order to test the abilities of the robot by using simple blocks as obstacles.
12 Brooks 1991; Varela et al. 1991; Kelso 1995, Clark 1997; Keijzer 2001; Wheeler 2005; Thompson 2007; Chemero 2009.
13 Viewed from the outside, the behavior of Brooks' Creatures seems so complex, that we assign them representations and high-level cognitive processes. The complexity of their behavior is, however, explained not by reference to complex programming, but to the world itself. Brooks 1991, p. 406.
14 "One does not need to know what the optimum is in order to move toward it. One's body is simply drawn to lower the tension." Dreyfus 2008, p. 343.
15 Lakoff and Johnson (1999, p. 34) show that even the concepts that are most distant from the structural coupling originate themselves in bodily states. For example, happiness and sadness are usually understood as proprioceptive states of feeling high or down. Cf. also a similar argument in Gibbs 2007, pp. 2 ff.
16 Some recent works (for example, Allen et al. 2006) argue that for the robotic ethical behavior to occur, a free will is not necessary; it is enough to implement the ethics in the basic programming.
17 Dreyfus has a dual attitude: on the one hand, he affirms that the discreteness of transitions from one attractor to another makes possible to model the human mind on a computer (2008, p. 357); on the other hand, he acknowledges that Heidegger's and Freeman's accounts are about us, our embodiment, our cultural interpretation and that a Heideggerian robot "can't get off the ground", if it is sensitive to significance in the way the human beings are. (pp. 361-362).
18 Powers argues that the Kantian ethics is more suitable for robots because it is based on rules and rules could easily run on computers. Cf. Powers 2006, p. 465.