Artificial Ethics

Laura L. Pană
Polytechnic University of Bucharest, Romania

INTRODUCTION

The possibility to conceive and to effectively apply a new ethics, deeply explicit and valid in theory, strongly relevant in practice, and suitable for both artificial and human intelligent agents is argued in this chapter.

The whole intellectual history shows that long ago humans were aware of their specific gift to add artificial objects of social (material or ideal) kind to those natural, i.e. to create culture, as an extension of nature. The oldest cyberneticist, Plato, in his *Republic*, offered the first description of society as a human design and an artificial product. Aristotle, in his turn, wrote in *Politics* (II. 5) about the automatic tools and installations of Daedal and even about the mentally controlled tripods created by Hephaestus, which served the "band" of gods.

For the specific field of moral culture, P. Danielson outlined the artificiality of morals and analyzed the possible degrees of creativity by which moral values are invented and moral culture is renewed (Danielson, 1998). He explicitly writes: "important parts of morality are artificial cognitive and social devices" (p.292). Even before, J. Bentham, who explored both the human mind and the *Table of Springs of Action* (1812), understood the artificial character of morality and described "the whole fabric of morals". He elaborated a both comprehensive and operational moral theory, and coined a specific calculus, based on the central moral value of his vision.

This study identifies theoretical possibilities to model moral conduct, and aims to find and to develop a set of moral prerequisites (mental aptitudes and practical skills), available and suitable for any kind of moral agents. In this way, the axiological foundation of the new ethics may be continued by an attempt to identify an appropriate and operable value-set, selected from a complete ethical system.

The main practical contributions covered by chapter: indicates a well-founded way to effectively connect moral theory and moral practice; proposes a complex but feasible strategy of designing artificial moral agents, detailed in a few operational phases; makes a concrete proposal of modeling and implementing moral action in the behavior of artificial agents; shows the new possibilities offered by the present changes in value systems for effective moral agents designing and training.

BACKGROUND: SCIENTIFIC, TECHNICAL, AND PHILOSOPHICAL PREMISES OF ARTIFICIAL ETHICS

Some theoretical deficiencies of great ethical systems and some practical difficulties of applying abstract moral values in concrete conditions by individual agents have been frequently discussed by ethicists, in their common effort to establish a new foundational theory of moral choice, moral freedom and then of a deep moral conduct.

The nature of artificial agents able to behave ethically has been extensively studied. This study leaded to different conclusions, which can however be considered convergent. Among other features of moral agents McDermott emphasizes the capacity to take specific, moral decisions, just like (Moor, 2005 and 2011), and analyses ethical reasoning, seen as the main decisional sub-structure (McDermott, 2011). The role of free will is also

DOI: 10.4018/978-1-5225-2255-3.ch008

detailed by him, while the will itself is refined up to temptation. J. Gips also emphasizes the role of free will, as of conscious choice (Gips, 2011), and shows the necessity to develop perceptual and non-symbolic aspects of morality, and to favor training, not teaching of abstract theories. (p. 250). Wallach and Allen correlated autonomy and sensitivity (to moral considerations) as defining dimensions for artificial moral agents (Wallach & Allen, 2009). For J. P. Sullins, the relevant aspects of moral agency are autonomy, intentionality and responsibility (Sullins, 2006). L. Floridi studies the distribution of the necessary factors for an ethical behavior - interactivity, autonomy and adaptability -, between a large category of agents, such as natural objects, ecosystems, technical systems, organizations and humans. Intentionality is also associated, as the most important, responsibility, which is here distinguished from accountability (Floridi, 2011, p. 205). The list of the required capacities is also extended in (Anderson, 2011) by sentience, self-consciousness, reasoning, and emotionality. Moral agency is also considered here, although some of the above indicated elements are integrated into the very internal structure of action, as aspects or parts of interests, motivations, decisions and goals. Other authors dedicated to define agency include, among other features, normativity and asymmetry (Barandiaran, Paolo, & Rohde, 2009). Spatio-temporality, also assigned to agents by them, actually is a universal property. Moral agent's study, as part of the artificial intelligent agent's theory, may be illustrated including by (Pană, 2005c), a study on artificial cognitive agents, followed by (Pană, 2008b), on cognitive and moral agents, viewed in their evolution (Pană, 2006b). This chapter continues and deepens the growing list of the possible features previously assigned to artificial moral agents by (Pană, 2006a; Pană, 2012), especially by attributes related to consciousness and value systems.

Many scholars believe consciousness is a condition of moral judgment and moral conduct. Here the approaches range goes from those philosophical, neurobiological and psycho-sociological up to computational ones, such as those which mark the progress in the field from diverse perspectives (Gamez, 2008), which ambition to build a conscious machine (Angel, 1989) or which explore large areas of the field but also thoroughly study a series of essential topics (Holland, 2003). A study on axioms and tests for the presence of minimal consciousness in agents establishes as eligibility criteria perception, imagination, attention, planning capacity and emotion (Alexander & Dunmall, 2003), in conformity with their set of axioms (p. 9-10), which is adequate for the consciousness understood, *apud* Dennett, as "a virtual machine running on a parallel neural computer" (p.8). The contribution of the present chapter in this matter is based on studies concerning an integrative model of brain, mind, cognition and consciousness (Pană, 2008a), as well as on a few forms of social consciousness (Pană, 2000), and consists of an analysis of the internal structure of moral conscience, in order to find which of its components and levels are suitable to be modeled and implemented in artificial moral agent programs.

The project of building an artificial ethics, suitable both for human and artificial moral agents, have to include, from the start, a value-centered approach and vision. In the today dedicated literature, the values under discussion are those human, and mainly those already recognized in the moral field, such as good and evil, right and wrong, but also a few others, implied both in current human activities, as in the human-computer interaction, e.g. egoism and altruism, treated in (Floridi, 2011), together with values occurring in organization environments, such as equal opportunity, financial stability, good working and holiday conditions for employees, good service and value to their customers or shareholders, and honesty, integrity and reliability to other companies (p. 205). Inspired by competing and cooperation needing contexts, other authors study the relationship between altruism and reciprocity, as in (Danielson, 2002), where he develops evolutionary agent-based models in order to test their conjecture in various cases. Responsibility, which remains a core value