Chapter 7.4
Moral Emotions for Autonomous Agents

Antoni Gomila
University Illes Balears, Spain

Alberto Amengual
International Computer Science Institute, USA

ABSTRACT

In this chapter we raise some of the moral issues involved in the current development of robotic autonomous agents. Starting from the connection between autonomy and responsibility, we distinguish two sorts of problems: those having to do with guaranteeing that the behavior of the artificial cognitive system is going to fall within the area of the permissible, and those having to do with endowing such systems with whatever abilities are required for engaging in moral interaction. Only in the second case can we speak of full blown autonomy, or moral autonomy. We illustrate the first type of case with Arkin's proposal of a hybrid architecture for control of military robots. As for the second kind of case, that of full-blown autonomy, we argue that a motivational component is needed, to ground the self-orientation and the pattern of appraisal required, and outline how such motivational component might give rise to interaction in terms of moral emotions. We end suggesting limits to a straightforward analogy between natural and artificial cognitive systems from this standpoint.

1. INTRODUCTION

The increasing success of Robotics in building autonomous agents, with rising levels of intelligence and sophistication, has taken away the nightmare of "the devil robot" from the hands of science fiction writers, and turned it into a real pressure for roboticists to design control systems able to guarantee that the behavior of such robots comply with minimal ethical requirements. Autonomy goes with responsibility, in a nutshell.

DOI: 10.4018/978-1-60960-818-7.ch7.4

Otherwise the designers risk having to be held themselves responsible for any wrong deeds of the autonomous systems. In a way, hence, predictability and reliability of artificial systems pull against its autonomy (flexibility, novelty in novel circumstances). The increase in autonomy rises high the issue of responsibility and, hence, the question of right and wrong, of moral reliability.

Which these minimal ethical requirements are may vary according to the kind of purpose these autonomous systems are build for. In the forthcoming years it is foreseable an increase in "service" robots: machines specially designed to deal with particularly risky or difficult tasks, in a flexible way. Thus, for instance, one of the leading areas of roboethical research concerns autonomous systems for military purposes; for such new systems, non-human supervision of use of lethal weapons may be a goal of the design, so that a guarantee must be clearly established that such robots will not kill innocent people, start firing combatants in surrender or attack fellow troops, before they are allowed to be turned on. In this area, the prescribed minimal requirements are those of the Laws of War made explicit in the Geneva Convention and the Rules of Engagement each army may establish for their troops. Other robots (for rescue, for fire intervention, for domestic tasks, for sexual intercourse) may also need to count on "moral" norms to constrain what to do in particular circumstances ("is it ok to let one person starve to feed other two?"). Much more so when we think of a middle range future and speculate about the possibility of really autonomous systems, or systems that "evolve" in the direction of higher autonomy: we really should start thinking about how to assure that such systems are going to respect our basic norms of humanity and social life, if they are to be autonomous in the fullest sense. So the question we want to focus on in this paper is: how should we deal with this particular challenge?

The usual way to deal with this challenge is a variation/extension of the existing deliberative/reactive autonomous robotic architectures, with the goal of providing the system with some kind of higher level control system, a reasoning moral system, based on moral principles and rules and some sort of inferential mechanism, to assess and judge the different situations in which the robot may enter, and act accordingly. The inspiration here is chess design: what's required is a way to anticipate the consequences of one's possible actions and of weighting those alternatives according to some sort of valuation algorithm, that excludes some of those possibilities from consideration altogether. Quite appart from the enormous difficulty of finding out which principles and rules can capture our "moral sense" in an explicit form, this project also faces the paradoxes and antinomies that lurk into any formal axiomatic system, well-known from the old days of Asimov's laws. So to speak, this approach inherits the same sort of difficulties known as the "symbol grounding" and "frame" problems in Cognitive Science.

However, it might turn out that there is a better way to face the challenge: instead of conceiving of morality as a higher level of control based on a specific kind of reasoning, it could be conceived instead as an emotional level of control, along the current trend in the Social Neurosciences and Psychology which point in such direction (for an illustration, the special double issue in volume 7 of the journal Social Neuroscience). From this point of view, which in fact resumes the "moral sense" tradition in Ethics, moral judgement is not a business of reason and truth, but of emotion in the first place, not of analytical pondering of rights and wrongs, but of intuitive, fast, immediate affective valuation of a situation (which may be submitted to a more careful, detailed, reflexive, analysis later on), at least at the ground level. From this point of view, it might be a better option in order to build systems with some sort of "moral" understanding and compliance, to start building systems with a practical understanding of emotions and emotional interaction, in particular moral emotions. Rights and norms, so the story