Chapter 20
The Functional Morality of Robots

Linda Johansson
Royal Institute of Technology, Sweden

ABSTRACT

It is often argued that a robot cannot be held morally responsible for its actions. The author suggests that one should use the same criteria for robots as for humans, regarding the ascription of moral responsibility. When deciding whether humans are moral agents one should look at their behaviour and listen to the reasons they give for their judgments in order to determine that they understood the situation properly. The author suggests that this should be done for robots as well. In this regard, if a robot passes a moral version of the Turing Test—a Moral Turing Test (MTT) we should hold the robot morally responsible for its actions. This is supported by the impossibility of deciding who actually has (semantic or only syntactic) understanding of a moral situation, and by two examples: the transferring of a human mind into a computer, and aliens who actually are robots.

INTRODUCTION

Technoethics focuses on the ethical aspects of technology in society, and attempts to devise principles to guide technological development in particular in relation to emerging new technologies that give rise to new ethical issues. Increasingly autonomous and intelligent robots represent one of these new technologies. Autonomy in robots raises questions about robot morality, and about how we can make sure that the autonomous robots behave ethically, in some sense (Bekey, 2005; Allen, Smit, & Wallach, 2006; Andersson, 2008; Andersson, Anderson, & Armen, 2004; Allen, Varner, & Zinser, 2000).

This is relevant for ongoing technological development. Today there are robotic research programs where internal "ethical governors" and

DOI: 10.4018/978-1-4666-1773-5.ch020

"guilt systems" are being developed, and there are discussions on potential "pull the trigger-autonomy" of unmanned aerial vehicles in war (Arkin, 2009).

It has sometimes been argued that a robot can never be held responsible for its actions. No matter how advanced it is it can never be autonomous; since it is always programmed by a human there is no question of a robot having alternative possibilities. Robots also seem to lack mental states, which are considered necessary in order to be an agent.

This paper argues that we do not need to know what goes on in a robot, in terms of being programmed or possessing mental states. If a robot can pass a so called Moral Turing Test, than we can hold it morally responsible for its actions.

An objection to the whole idea of trying to find criteria for whether a robot might be morally responsible, is the "what would be the point"-objection, that is, that it would be pointless to hold a machine responsible. We cannot punish a robot; it would be useless to send it to prison, for instance. If it misbehaves, we would simply turn it off or destroy it. But in a longer perspective, with robots becoming more advanced, this issue cannot be ignored. The potential responsibility of robots might have an impact on liability when something goes wrong. If the robot is considered responsible, we may not be able to punish it, but its responsibility may have implications for the responsibility of others, such as programmers. The ascription of responsibility to robots may also have influence on decisions whether and how to use such robots. There might also be implications for how robots should be programmed "ethically".

The idea of morality as a human *construction* in a moral community is a useful assumption when investigating matters regarding robot morality. Moral responsibility can be expected to be a central notion in the moral community since the whole point of morality is to promote right actions and prevent wrong actions. Members of the moral community might be moral agents or moral receivers, i.e., agents whose well-being is morally relevant but who cannot be held morally responsible.

The paper is outlined as follows. First, the moral community is described in terms of its members and how members decide whether other members are agents (which are a part of the other minds problem). The solution to the other minds problem seems—in this community—to be functionalistic; that is, we look at other humans' behaviour and assume that their mental states are what caused their outward behaviour. Then the moral community is discussed focusing on the case for robots. I argue that the nature of morality—the way humans actually behave in moral matters—supports the idea that *the passing of a moral Turing Test* (MTT) is a necessary and sufficient criterion for being held morally responsible. Support for this idea also comes from the so called "*deceiving robot-aliens"-example*. In summary I will conclude that the "functional morality" of robots allows us to hold them responsible. We have no reason to be biased towards nonorganic potential agents.

THE MORAL COMMUNITY

In order to decide whether we can hold *robots* morally responsible we should begin by considering when we hold *humans* morally responsible. On what criteria do we—or do we not— hold humans morally responsible?

Whom do we consider morally responsible? Not children, at least not very young children, and not animals, for instance. The same goes for the severely mentally ill, which is why we often test the mental health of defendants who are tried in court. Consider, for instance, a child hitting another child, not realizing that the other child feels pain—or someone who is mentally ill, and believed that he was fighting trolls and not innocent humans. The reason for not holding such people morally responsible is that we doubt their ability to properly understand what they do and what the consequences might be, or their ability