Robots that stimulate autonomy (extended abstract)\textsuperscript{1}

Matthijs A. Pontier\textsuperscript{a}, Guy A. M. Widdershoven\textsuperscript{b}

\textsuperscript{1, 2}VU University Amsterdam, 
\textsuperscript{a}CAMeRA / Network Institute, De Boelelaan 1081, 1081HV Amsterdam, The Netherlands, 
\textsuperscript{b}VU University Medical Center, Amsterdam, The Netherlands, 
matthijsp@icloud.com, g.widdershoven@vumc.nl

Robots are increasingly being used to provide a high standard of care in the near future, due to a foreseen lack of resources and healthcare personnel. By providing assistance during care tasks, or fulfilling them, robots can relieve time for the many duties of care workers.

As their intelligence increases, the amount of human supervision decreases and robots increasingly operate autonomously. With this development, we increasingly rely on the intelligence of these robots. When we start to depend on autonomously operating robots, we should be able to rely on a certain level of ethical behavior from machines. Particularly when machines interact with humans, which they increasingly do, we need to ensure that these machines do not harm us or threaten our autonomy. In complex and changing environments, externally defining ethical rules unambiguously becomes difficult. Therefore, autonomously operating care robots require moral reasoning.

In a recent interview in a multimillion copies free newspaper [5], we presented a humanoid robot for healthcare (a “Caredroid”) in which we will implement the moral reasoning system. The caredroid will assist people in finding suitable care, and assist them in making choices concerning healthcare.

Caredroids will encounter moral dilemmas. For example, when supporting a patient in making choices, the caredroid should balance between accepting unhealthy choices or trying to persuade the patient to reconsider them.

In previous research, Pontier and Hoorn [7] developed a moral reasoning system based on the moral principles developed by Beauchamp & Childress [1]. In simulation experiments, the system was capable of balancing between conflicting principles. In medical ethics, autonomy is the most important moral principle.

Often autonomy is equated with self-determination. In this view, people are autonomous when they are not influenced by others. However, autonomy is not just being free from external constraints. Autonomy can also be conceptualized as being able to make a meaningful choice, which fits in with one’s life-plan [3]. In this view, a person is autonomous when he acts in line with well-considered preferences. This implies that the patient is able to reflect on fundamental values in life. Core aspects of autonomy as self-determination are mental and physical integrity and privacy. Central in autonomy as ability to make a meaningful choice are having adequate information about the consequences of decision options, the cognitive capability to make deliberate decisions, and the ability to reflect on the values behind one’s choices. Autonomy as self-determination can be called negative autonomy, or ‘being free of’. Autonomy as the ability to make a meaningful choice is called positive autonomy or ‘being free to’ [2]. To be able to reflect this more complex view of autonomy in the moral reasoning system, we decided to expand the moral principle of autonomy.

In medical practice, a conflict between negative autonomy and positive autonomy can play a role. Sometimes, the self-determination of the patient needs to be constrained on the short-term to achieve positive autonomy on the longer term. For example, when a patient goes into rehab his freedom can be limited for a limited period of time to achieve better cognitive functioning and self-reflection in the future.

We present a moral reasoning system including a twofold approach of autonomy. The system extends a previous moral reasoning system [7] in which autonomy consisted of a single variable. The behavior of the current system matches the behavior of the previous system. Moreover, simulation of legal cases for courts in the Netherlands showed a congruency between the verdicts of the judges and the decisions of the presented moral reasoning system.

\textsuperscript{1} The full version of this paper appeared in: Artificial Intelligence Applications and Innovations 2013, IFIP Advances in Information and Communication Technology, AIAI’13
including the twofold model of autonomy. Finally, the experiments showed that in some cases long-term positive autonomy was seen as more important than negative autonomy on the short-term.

Case 1 showed that, both according to the judge and to the model, assertive outreach was a morally justifiable option to prevent judicial coercion. By assertive outreach, the mental integrity and privacy of the patient were constrained. However, this prevented worsening of the situation, which would have raised the need for judicial coercion, a measure that would constrain the privacy of the patient more heavily.

In case 2, the psychiatrist should have informed the ambulatory care team instead of the parents of the patient. Informing the ambulatory care team constrained the privacy of the patient less than informing the parents, and had more potential to prevent worsening of the situation and improve the cognitive functioning of the patient. Because of its advantages for positive autonomy, also when only taking into account the principle of autonomy, informing the ambulatory care team is in this situation a better option than doing nothing. Thus, in this case, constraining negative autonomy in benefit of positive autonomy improves the overall level of autonomy.

Finally, case 3 showed that negative autonomy can sometimes be constrained to stimulate positive autonomy. In this case, the patient had agreed to that under certain conditions. Because the conditions of a self-binding declaration were met, judicial coercion was justified. During detoxification, the cognitive function and reflection of the patient could be restored. Because of this stimulation of positive autonomy on the longer term, the constraints of negative autonomy on the short-term in the end positively influence the level of overall autonomy.

The moral reasoning system presented in this paper can be used by robots and software agents to prefer actions that prevent users from being harmed, improve the users’ well-being and stimulate the users’ autonomy. By adding the twofold approach of autonomy, the system can balance between positive autonomy and negative autonomy.

In future work, we intend to integrate the model of autonomy into Moral Coppélia [6], an integration of the previously developed moral reasoning system [7] and Silicon Coppélia - a computational model of emotional intelligence [4]. Adding the twofold model of autonomy to Moral Coppélia may be useful in many applications, especially where machines interact with humans in a medical context.

After doing so, the level of involvement and distance (cf. [4]) will influence the way the robot tries to improve the autonomy of a patient. A robot that is more involved with the patient will focus more on improving positive autonomy and especially on reflection.

Acknowledgements
This study is part of the SELEMCA project within CRISP (grant number: NWO 646.000.003).

References


