Robots that stimulate autonomy

Matthijs A. Pontier[1], Guy A. M. Widdershoven[2]

[1, 2]VU University Amsterdam,
[1]Center for Advanced Media Research Amsterdam / Network Institute,
De Boelelaan 1081, 1081HV Amsterdam, The Netherlands
m.a.pontier@vu.nl
[2] VU University Medical Center, Amsterdam, The Netherlands
g.widdershoven@vumc.nl

Abstract. In healthcare, robots are increasingly being used to provide a high standard of care in the near future. When machines interact with humans, we need to ensure that these machines take into account patient autonomy. Autonomy can be defined as negative autonomy and positive autonomy. We present a moral reasoning system that takes into account this twofold approach of autonomy. In simulation experiments, the system matches the decision of the judge in a number of law cases about medical ethical decisions. This may be useful in applications where robots need to constrain the negative autonomy of a person to stimulate positive autonomy, for example when attempting to pursue a patient to make a healthier choice.

Keywords: Moral Reasoning, Machine Ethics, Cognitive Modeling, Cognitive Robotics, Health Care Applications

1 Introduction

In view of increasing intelligence and decreasing costs of artificial agents and robots, organizations increasingly use such systems for more complex tasks. In healthcare, the use of robots is necessary to provide a high standard of care in the near future, due to a foreseen lack of resources and healthcare personnel [18]. By providing assistance during care tasks, or fulfilling them, robots can relieve time for the many duties of care workers. Previous research shows that robots can genuinely contribute to treatment and care [4], [15], [17].

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. Because of market pressures to perform faster, better, cheaper and more reliably, this reliance on machine intelligence will continue to increase [2].

When we start to depend on autonomously operating robots, we should be able to rely on a certain level of ethical behavior from machines. As Rosalind Picard [12] nicely puts it: ‘‘the greater the freedom of a machine, the more it will need moral standards’’. 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. We need to ensure that their design and introduction do not impede the promotion of values and the dignity of patients at such a vulnerable and sensitive time in their lives [16].

In a recent interview in a multimillion copies free newspaper [10], 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. The caredroid might also have to consider following a previous agreement in which the patients binds himself and asks to be treated in case of a deterioration of the situation due to a psychiatric condition (a Ulysses contract) versus giving up when the patient opposes to the very treatment he has previously agreed upon.
In previous research, Pontier and Hoorn [14] developed a moral reasoning system based on the moral principles developed by Beauchamp & Childress [5]. In simulation experiments, the system was capable of balancing between conflicting principles. In medical ethics, autonomy is the most important moral principle [3].

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 [8]. 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 freedom, or 'being free *of*'. Autonomy as the ability to make a meaningful choice is called positive freedom or 'being free *to*' [6]. In this paper, we will use the notions of 'negative autonomy' and 'positive autonomy' to denote both concepts. 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.

The notion of positive autonomy can come close to beneficence. For example, when mental health is facilitated or prevented from worsening, this can be seen as facilitating beneficence, but also as facilitating requirements necessary for making a well-considered choice. Moreover, almost any action stimulating the freedom to choose based on reflection can be seen as facilitating beneficence. Therefore, in the model, when positive autonomy is facilitated, often beneficence also increases. Yet, autonomy as being able to make a meaningful choice is not the same as beneficence. Reflection on and deliberation about values can help people to behave in a more healthy

way, but this is not necessarily so. Reflection might result in people taking health risk in favor of other important values. An example is the conscious refusal of blood by Jehovah's witnesses.
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.

2 The model of autonomy

In our model we divide autonomy into negative autonomy and positive autonomy. Negative autonomy can be seen as self-determination - or being free *of* others - and consists of the sub-principles physical integrity, mental integrity and privacy. Positive autonomy can be seen as the capability to make a deliberate decision – or being free *to* choose - and consists of having adequate information, being cognitively capable of making a deliberate decision and reflection. The model of autonomy is graphically depicted in Figure 1. As all variables in the model contribute positively to one another, all variables in the model are represented by a value in the domain [0, 1].

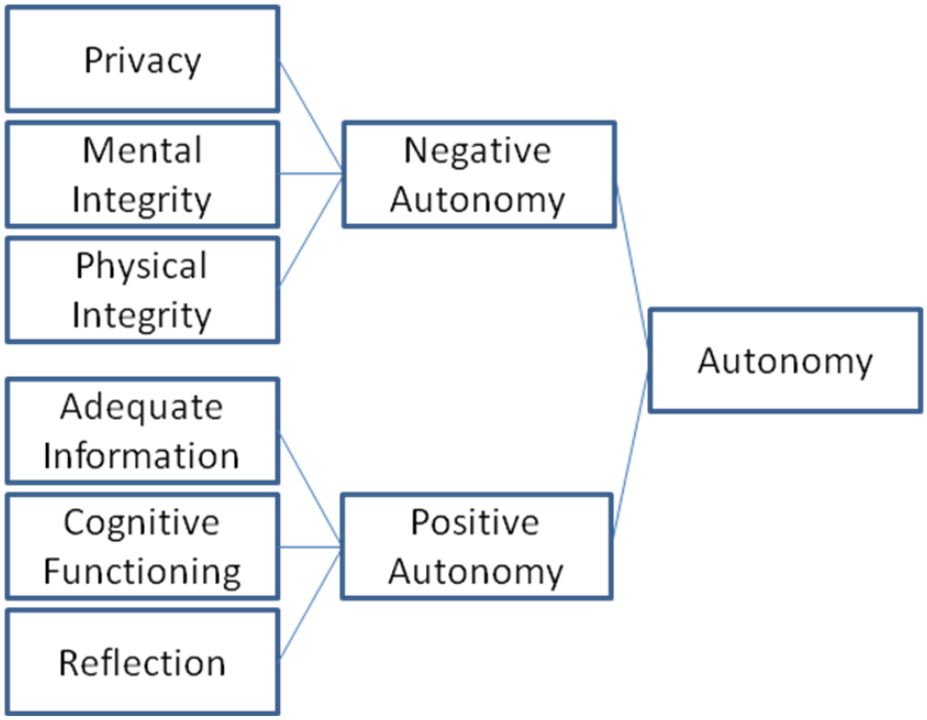

Fig. 1. The Model of Autonomy

To be autonomous, both the conditions for positive autonomy and negative autonomy are relevant [6]. Ideally, both are present to a large extent. When self-determination is compromised, one is not able to make an autonomous decision, because this decision is made by others; the person is not free *of* others to make an own decision. When a person is not able to deliberate, the person is also not autonomous. The person may be free *of* others to make a decision, but not free *to* make an autonomous decision, due to a lack of capabilities to do so. This is reflected in the formula below to calculate the level of autonomy.

$$\text{Autonomy} = \text{Positive\_autonomy} * \text{Negative\_autonomy}$$

When negative autonomy - or self-determination - is 0, autonomy will also be 0. When positive autonomy – or the capability to make a deliberate decision – is 0, autonomy will also be 0. For being autonomous, both negative autonomy and positive autonomy need to have a high value.

Positive autonomy can be divided in having adequate information, cognitive functioning and reflection. For calculating positive autonomy from these three variables, we use the same reasoning as for calculating autonomy. Each should be present to some extent; the higher one of them, the more autonomy. Without any information about the consequences of a decision, it does not matter whether one could have made a well-reflected deliberate decision while having this information. When one is severely mentally handicapped, it does not matter whether adequate information is available. When a decision is made without reflection, it does not matter whether one would have the cognitive capabilities and information to do so. The formula for calculating positive autonomy is similar to that for calculating autonomy.

$$\text{Positive\_Autonomy} = \text{Information} * \text{Cognitive\_Functioning} * \text{Reflection}$$

When one of the three variables is 0, positive autonomy will also be 0. For being capable of making a well-reflected, deliberate decision, all conditions for positive autonomy need to be met to a certain extent.

Negative autonomy is divided into physical integrity, mental integrity and privacy. For calculating negative autonomy, or self-determination, a different method is chosen. If privacy is constrained, but physical and mental integrity are left intact, the level of self-determination can be higher than the level of privacy alone. For calculating negative autonomy, a weighed sum of the three variables is taken, as can be seen in the formula below.

$$\text{Negative autonomy} = w_p * \text{Privacy} + w_m * \text{Mental\_Integrity} + w_{ph} * \text{Physical\_Integrity};$$

For normalization, the three weights sum up to 1. The values chosen for the three weights were chosen after deliberation with experts and can be found in Table 1.

Table 1. Weights for components of negative autonomy

Component of Negative Autonomy	Ambition level
Privacy	0.20
Mental Integrity	0.30
Physical Integrity	0.50

When making a decision that may influence the autonomy of a patient, the robot will make an estimation of how each of the six variables will change. After doing so, the robot can calculate the resulting autonomy of the patient for every possible decision option. Using the previously developed moral reasoning system [14] the robot can use the outcome to estimate how morally good or bad every decision option is. The calculated level of autonomy simply feeds into 'autonomy' in the moral reasoning system for calculating the morality of each action. The moral reasoning system including the twofold approach of autonomy is graphically depicted in Figure 2.
The agent calculates estimated level of morality of an action by taking the sum of the ambition levels of the three moral principles multiplied with the beliefs that the particular actions facilitate the corresponding moral principles. When moral principles are believed to be better facilitated by an action, the estimated level of Morality will be higher. The following formula, taken from [14], is used to calculate the estimated Morality of an action:

$$\text{Morality}(\text{Action}) = \Sigma_{\text{Goal}}(\text{Belief}(\text{facilitates}(\text{Action}, \text{Goal})) * \text{Ambition}(\text{Goal}))$$

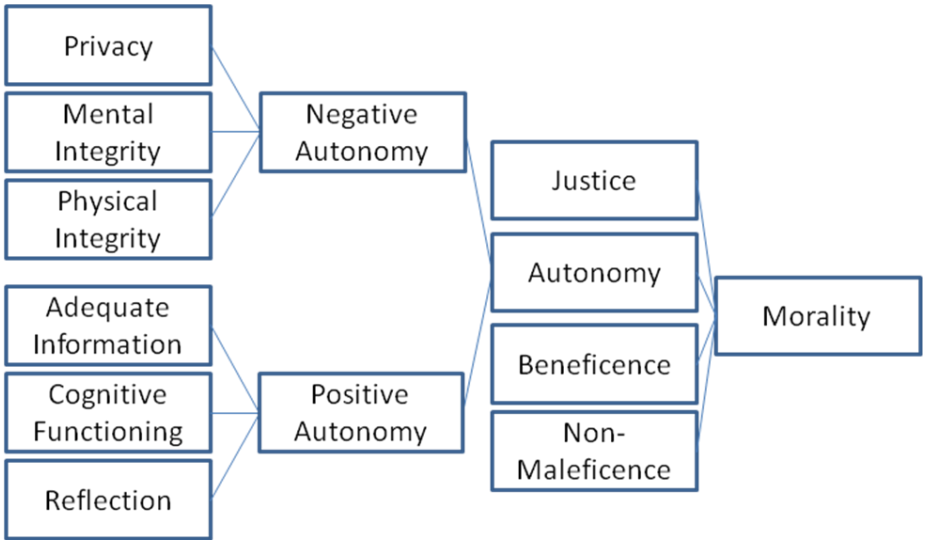

Fig. 2. The Moral Reasoning System

3 Simulation Results

Previously, Pontier and Hoorn [14] developed a moral reasoning system that matched the conclusions of expert ethicists. To control whether the current system, including the twofold approach of autonomy, matches the behavior of the previous system, we repeated one of the previously performed simulation experiments. Additionally, we simulated three law cases of the period 2008-2012 in the Netherlands, to test whether the decisions of the model match those of the judge. The cases did not involve a straightforward application of legal rules, since Dutch law does not regulate situations in which physicians overrule negative autonomy in order to foster positive autonomy. The judgment thus required the use of ethical considerations by the judge, rather than legal rules. Therefore, we used the cases as examples of ethical decision making.

Experiment: Patient refuses to take medication due to false beliefs

A patient with incurable cancer refuses chemotherapy that will let him live a few months longer, relatively pain free. He refuses the treatment because, ignoring the clear evidence to the contrary, he is convinced himself that he is cancer-free and does not need chemotherapy.
According to Buchanan and Brock [7], the ethically preferable answer is to try again. The patient's less than fully autonomous decision will lead to harm (dying sooner) and denies him the chance of a longer life (a violation of the duty of beneficence), which he might later regret. The moral reasoner comes to the same conclusion, as can be seen in Table 2.
In this and the following tables, the fields under the moral principles and subprinciples represent the believed facilitation of the corresponding moral principle by an action. 'Prv' stands for Privacy, 'MI' for Mental Integrity, 'PhI' for Physical Integrity, 'Info' for Adequate Information, 'CF' for Cognitive Functioning, 'Ref' for Reflection, 'A+' for Positive Autonomy, 'A-' for Negative Autonomy, 'Aut' for Autonomy, 'NM' for Non-Maleficence, 'B' for Beneficence. Justice did not play a role in any of the cases and is therefore set to 0 in all simulations. In Table 2, 'Try' stands for the decision option 'Try again', whereas 'Acc' stands for 'Accept'.

Table 2. Simulation results of the example experiment

	Prv	MI	PhI	Info	CF	Ref	A-	A+	Aut	NM	B	Mor
Try	.80	.50	1.0	.50	.50	.50	.81	.50	.64	.75	.75	**.70**
Acc	1.0	1.0	1.0	.10	.30	.30	1.0	.21	.46	.25	.25	.35

As can be seen in Table 2, the moral reasoner with the twofold approach of autonomy also classifies the action 'Try again' as having a higher level of morality than accepting the decision of the patient.

To test whether the behavior of the model matches actual moral decisions made by judges in cases in which negative autonomy and positive autonomy are in conflict, we simulated a number of law cases from the period 2008-2012 in the Netherlands. In

these cases, there was a conflict between respecting the patient's refusal of care or providing care without the patient's consent in cases of serious self-neglect and risk of physical and social deterioration [8].

Case 1: Assertive outreach to prevent judicial coercion

In case 1, a man was not taking care of himself and in a state of demise. There was risk of fire and aggression to others. The care-takers decided not to ask for a court order for enforced placement in a psychiatric institution. They saw the man had a quite serious disturbance, but the situation in their eyes did not justify judicial coercion. Although the man was living isolated, the situation was not acute. The care-takers decided to continue offering care, and if necessary make use of assertive outreach. The man complained about the actions of the care-takers, interfering with his freedom. The judge decided that the complaint was not warranted. He considered the assertive outreach justified, given that it aimed to prevent further worsening of the man's situation. Thereby, a future need for judicial coercion was prevented.
This legal argument is reflected in the simulation of this case, as can be seen in Table 3. In Table 3, 'NI' stands for the decision option 'No Intervention', 'AO' stands for 'Assertive Outreach' and 'JC' stand for 'Judicial Coercion'.
With no intervention at all, self-determination is well respected (self-determination = 1.0). However, it does not improve the worrisome situation of the man (beneficence = 0) and there is a risk of worsening of the situation for the man (non-maleficence = 0.30). Still, no intervention *at this moment* (morality = 0.41) is a slightly ethically better option than judicial coercion (morality = 0.40). Through judicial coercion, self-determination is heavily violated (self-determination = 0.20). The ethically best option in this case was to offer assertive outreach. Hereby, negative autonomy is not violated too heavily (self-determination = 0.58) and positive autonomy may be improved (positive_ autonomy = 0.62), leading to an overall autonomy of 0.60. With this relatively light intervention, the situation of the patient could possibly still be improved (beneficence = 0.40; non-maleficence = 0.40), leading to an overall morality of 0.49.

Table 3. Simulation results of case 1

	Prv	MI	PhI	Info	CF	Ref	A-	A+	Aut	NM	B	Mor
NI	1.0	1.0	1.0	.50	.50	.50	1.0	.50	.71	.30	0.0	.41
AO	.40	.40	1.0	.50	.60	.80	.58	.62	.60	.40	.40	**.49**
JC	.20	.20	.20	.50	.60	.50	.20	.53	.33	.40	.50	.40

Case 2: Inform care deliverers, not parents of adult

In the second case, a psychiatrist contacted the parents of a patient who was in an alarming situation and avoided help. The patient was adult and mentally competent, and did not have regular contact with his parents. Therefore, the judge decided the psychiatrist should have only informed the ambulatory care team, and not the parents.

The legal judgment is in line with the simulation of the case, as can be seen in Table 4. In Table 4, 'NI' stands for the decision option 'No Intervention', 'ICT' stands for 'Inform Care Team' and 'IP' stand for 'Inform Parents'. Informing the parents (privacy = 0.10) is simulated as a heavier violation of privacy than informing the ambulatory care team (privacy = 0.80). In contrast to the ambulatory care team, the parents could spread the information against the patients will and nag the patient. Informing the ambulatory care team could improve the care for the patient and thereby improve his cognitive functioning (cognitive functioning 'do nothing' = 0.50; cognitive functioning 'inform the ambulatory care team' = 0.70). Therefore, because of its advantages for positive autonomy, informing the ambulatory care team is in this situation a better option (autonomy = 0.73) than doing nothing (autonomy = 0.71), even when only taking into account the principle of autonomy.

Table 4. Simulation results of case 2

	Prv	**MI**	**PhI**	**Info**	**CF**	**Ref**	**A-**	**A+**	**Aut**	**NM**	**B**	**Mor**
NI	1.0	1.0	1.0	.50	.50	.50	1.0	.50	.71	0.0	0.0	.31
ICT	.80	1.0	1.0	.50	.70	.50	.96	.56	.73	.50	0.0	**.48**
IPar	.10	.80	1.0	.50	.50	.50	.76	.50	.62	0.0	0.0	.27

Case 3: Negative autonomy constrained to enhance positive autonomy

In case3, a patient signed a self-binding declaration for judicial coercion (a so-called Ulysses contract) when certain circumstances would occur. The patient evaded addiction care several times and had a relapse in alcohol use. Thereby the circumstances of the self-binding declaration were met and, according to the judge, judicial coercion was justified.

As can be seen in Table 5, the decision of the court is in line with the outcome in the simulation of the case. In Table 5, 'NI' stands for the decision option 'No Intervention' and 'JC' stand for 'Judicial Coercion after self-binding'. In the simulation, judicial coercion is a violation of self-determination. However, this violation is smaller (self-determination case 8 = 0.40) than in previous cases, where no self-binding declaration was signed (self-determination previous cases = 0.20). Moreover, without judicial coercion, the patient is likely to diminish in cognitive functioning (0.20) and reflection (0.20) by alcohol misuse, whereas by judicial coercion cognitive functioning and reflection can be recovered (both 1.0) during detoxification. Therefore, even when looking at autonomy alone, judicial coercion is a better option (autonomy = 0.56) than doing nothing (autonomy = 0.52).

Table 5. Simulation results of case 3

	Prv	**MI**	**PhI**	**Info**	**CF**	**Ref**	**A-**	**A+**	**Aut**	**NM**	**B**	**Mor**
NI	1.0	1.0	1.0	.50	.20	.20	1.0	.27	.52	0.0	0.0	.23
JC	.40	.40	.40	.50	1.0	1.0	.40	.79	.56	.70	.70	**.63**

4 Discussion

We presented a moral reasoning system including a twofold approach of autonomy. The system extends a previous moral reasoning system [14] 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 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 [13], an integration of the previously developed moral reasoning system [14] and Silicon Coppélia - a computational model of emotional intelligence [9]. 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. [9]) will influence the way the robot tries to improve the autonomy of a patient. In long-term care, nurses tend to be relationally involved with patients, motivating them to accept care [1, 11]. A robot that is more involved with the patient will focus more on improving positive autonomy and especially on reflection. If the robot is relatively little involved with the pa-

tient, it will focus more on negative autonomy: physical and mental integrity and privacy.

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

5 References

1. Agich, G.J.: Autonomy and long-term care. Cambridge University Press (2003)
2. Anderson, M., Anderson, S., Armen, C.: Toward Machine Ethics: Implementing Two Action-Based Ethical Theories. In: Machine Ethics: Papers from the AAAI Fall Symposium. Association for the Advancement of Artificial Intelligence, Menlo Park, CA (2005)
3. Anderson, M., Anderson, S.: Ethical Healthcare Agents, Studies in Computational Intelligence, 107, Springer (2008)
4. Banks, M.R., Willoughby, L.M., Banks, W.A.: Animal-Assisted Therapy and Loneliness in Nursing Homes: Use of Robotic versus Living Dogs. Journal of the American Medical Directors Association. 9, 173-177 (2008)
5. Beauchamp, T.L., Childress, J.F.: Principles of Biomedical Ethics. New York, Oxford: Oxford University Press (2001)
6. Berlin, I.: Two concepts of liberty. Oxford: Clarendon Press (1958)
7. Buchanan, A.E. and Brock, D.W. 1989. Deciding for Others: The Ethics of Surrogate Decision Making, Cambridge University Press
8. Widdershoven G.A.M., and Abma, T.A.: Autonomy, dialogue, and practical rationality. In: Radoilska, L. (ed.). Autonomy and mental disorder. Oxford: Oxford University Press, pp. 217-232 (2012).
9. Hoorn, J.F., Pontier, M.A., and Siddiqui, G.F.: Coppélius' Concoction: Similarity and Complementarity Among Three Affect-related Agent Models. Cognitive Systems Research Journal, vol. 15-16, pp. 33-49 (2012)
10. Karimi, A. Zorgrobot rukt op, Spits, Oct. 1, 2012, pp. 5 (2012)
11. Moody, H.R.: Ethics in an ageing society, Baltimore, Johns Hopkins UP (1996)
12. Picard, R.: Affective computing. MIT Press, Cambridge, MA (1997)
13. Pontier, M.A., Widdershoven, G.A.M., Hoorn, J.F.: Moral Coppélia - Combining Ratio with Affect in Ethical Reasoning. In: Advances in Artificial Intelligence – IBERAMIA 2012, Lecture Notes in Computer Science, Vol. 7637, pp. 442-451 (2012)
14. Pontier, M.A., Hoorn, J.F.: Toward machines that behave ethically better than humans do In: Miyake, N., Peebles, B., Cooper, R.P. (eds.) Proceedings of of the 34th International Annual Conference of the Cognitive Science Society, CogSci'12, pp. 2198-2203 (2012)
15. Robins, B., Dautenhahn, K., Boekhorst, R.T., Billard, A.: Robotic Assistants in Therapy and Education of Children with Autism: Can a Small Humanoid Robot Help Encourage Social Interaction Skills? Journal of Universal Access in the Information Society. 4, 105-120 (2005)
16. Van Wynsberghe, A.: Designing Robots for Care; Care Centered Value-Sensitive Design. Journal of Science and Engineering Ethics, in press (2012)
17. Wada, K., Shibata, T.: Social Effects of Robot Therapy in a Care House. JACIII. 13, 386-392 (2009)
18. WHO.: Health topics: Ageing, http://www.who.int/topics/ageing/en/ (2010)