Should Moral Decisions be Different for Human and Artificial Cognitive Agents?

Evgeniya Hristova (ehristova@cogs.nbu.bg)
Maurice Grinberg (mgrinberg@nbu.bg)
Research Center for Cognitive Science, Department of Cognitive Science and Psychology
New Bulgarian University
21 Montevideo Str., Sofia 1618, Bulgaria

Abstract

Moral judgments are elicited using dilemmas presenting hypothetical situations in which an agent must choose between letting several people die or sacrificing one person in order to save them. The evaluation of the action or inaction of a human agent is compared to those of two artificial agents – a humanoid robot and an automated system. Ratings of rightness, blamefulness and moral permissibility of action or inaction in incidental and instrumental moral dilemmas are used. The results show that for the artificial cognitive agents the utilitarian action is rated as more morally permissible than inaction. The humanoid robot is found to be less blameworthy for his choices compared to the human agent or to the automated system. Action is found to be more appropriate, morally permissible, more right, and less blameworthy than inaction only for the incidental scenarios. The results are interpreted and discussed from the perspective of perceived moral agency.

Keywords: moral dilemmas; moral judgment; artificial cognitive agents; moral agency

Introduction

Moral Dilemmas and Artificial Cognitive Agents

Moral judgments and evaluation of moral actions have been of great interest to philosophers and psychologists. Apart from the practical importance of better understanding moral judgments and related actions, morality is an essential part of human social and cognitive behaviour. Recently, the behaviour of artificial cognitive agents became central to research and public debate in relation to the rapidly increasing usage of robots and intelligent systems in our everyday life. Several important questions must find their answers as the use of artificial cognitive agents has many benefits but also many risks. Some of those questions concern moral agency - if those agents should be allowed to make moral decisions and how such decisions are judged and evaluated.

Moral judgments can be studied in their purest form using moral dilemmas – situations in which there is a conflict between moral values, rules, rights, and agency (Foot, 1967; Thomson, 1985). Moral dilemmas used in the paper are hyopothetical situations in which several people will die if the agent does not intervene in some way. The intervention will lead to the death of another person but also to the salvation of the initially endangered people.

Analogously to the two main approaches to human morality, Gips (1995) identifies two basic theoretical approaches to morality when it concerns artificial agents – consequentialist (utilitarian) and deontological (see also Allen, Smit, & Wallach, 2005). Concerning moral evaluation, these approaches give quite different perspectives on moral agency for artificial cognitive agents (Wallach & Allen, 2009). The utilitarian approach is not concerned with the protagonist or the reason of a moral action but only on the utility of the outcome, so it does not differentiate between human and artificial cognitive agents. On the other hand, the deontological approach considers the nature of the agent and it implies that different agents (e.g. human or artificial) can have different kind of duties.

This distinction in the approach to moral choice and its evaluation is used in this paper to investigate how people perceive artificial agents while making moral decisions. If participants have a more utilitarian attitude, they are expected to rate agents' behavior similarly, based on the perceived utility of the outcome. If participants have a more deontological attitude, they would rate differently the human and the artificial agents depending on the degree to which they consider them to be moral agents and hence, morally responsible.

Moral Agency and Artificial Cognitive Agents

The possibility for moral agency of artificial agents has been a matter of hot debate (e.g. see Anderson & Anderson, 2011; Wallach & Allen, 2008; Johnson, 2006). It is debated if the robots should be authorized to kill in moral dilemma situations, and if so, what rules should govern the real-time decisions that are necessary to determine whether killing any particular person is justified (Sparrow, 2007; Wallach & Allen, 2008).

In this paper, we want to explore the differences in moral agency evaluation depending on the type of agent - human or artificial. So, it is important to take into account the attribution of mind and moral agency to artificial cognitive systems (Waytz, Gray, Epley, & Wegner, 2010).

In the study of Gray et al. (2007), participants had to evaluate several characters including a human, a robot, and a computer with respect to the degree of possessing various cognitive capacities. The authors further established that moral judgments about punishment correlated more with one of the revealed dimensions – 'agency' – than with the second dimension – 'experience'. In Gray et al. (2007), the human obtains the highest scores on the 'experience' and 'agency' dimensions while the robot has practically zero score on the 'experience' and half the maximal score on the 'agency' scales. Following the interpretation given by the

authors, this implies that robots are judged as less morally responsible for their actions compared to humans.

Moral Judgments about the Actions of Artificial Cognitive Agents

Until recently, research involving moral dilemmas considered mainly a human agent. Only during the last years, several empirical studies appeared, exploring the moral judgments about the actions of artificial cognitive agents in moral dilemmas (Scheutz & Malle, 2014; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015; Hristova & Grinberg, 2015).

Malle et al. (2015) compared moral judgments about a human and a state-of-the-art robot agent choosing the utilitarian action or inaction. They used a modified version of the Trolley problem in which the death of the person to be killed is a side-effect of the action undertaken by the agent. They found that it is more permissible for a robot (compared to a human) to do the utilitarian action. It was also found that a human agent is blamed more the utilitarian action than for inaction while the robot was equally blamed for both.

In another study (Scheutz & Malle, 2014), a means-end scenario was used. In was found that the utilitarian action is judged to be both more morally wrong and more blameful than inaction for the human and the robot agent.

While the goal of Scheutz & Malle (2014) and Malle et al. (2015) was to study the expectations of people of state-of-the-art robots and inform future robot design, Hristova & Grinberg (2015) had the goal to explore the moral agency ascribed to hypothetical future artificial cognitive agents, which are indistinguishable from humans except for being built from inorganic materials. The study of Hristova & Grinberg (2015) contains only the results concerning the judgment of the utilitarian choice of the agents.

The present paper combines the results presented in Hristova & Grinberg (2015) with results about the judgments of inaction by the agents and compares them. Additionally, a test for moral agency, specifically concerning the agents used in the present study, has been carried out whose results are included in the discussion.

Goals and Hypotheses

The goal of this paper is to compare the moral judgments about the choices of a human agent, a humanoid robot (who is exactly like a human in terms of experiences and mind but has a non-organic body), and an automated system.

Table 1 contains the description of the agents. The only difference between the human and the humanoid robot, presentation of the agents, is the material the latter is built from. The automated system, on the other hand, is described as autonomous, free, and adaptable but lacks experiencing.

The expectation is that despite the fact that the humanoid robot supposedly has all the features for full morally responsible agency, people will evaluate its action or inaction differently compared to those of a human agent.

Another goal of the study is to explore the influence of the so-called '*instrumentality*' of harm on moral judgments. The instrumentality of harm is an important factor in moral dilemma research (e.g., Borg et al., 2006; Hauser et al., 2007; Moore et al., 2008). It draws attention to the fact that harm could be either inflicted intentionally as a 'mean to an end' (instrumental harm) or it could be a 'side effect' (incidental harm) from the actions needed to save more endangered people. It has been found that the unintended incidental harm (although being foreseen) was judged as more morally permissible than the intended instrumental harm (Hauser et al., 2007; Moore et al., 2008).

The utilitarian action is expected to be rated as more appropriate, more right, more morally permissible, and less blameworthy when the harm is incidental (compared to instrumental). However, the discussion about perceived moral agency suggests that the difference in moral judgments for the artificial and human agents will be greater when the harm is instrumental, as such actions involve more responsibility and direct infliction of harm.

The experiment collected ratings on the rightness, moral permissibility, and blameworthiness of performing or not the utilitarian action. The various questions asked can target different aspects of the evaluation of moral choices, as some studies suggest (Christensen & Gomila, 2012; Cushman, 2008). According to Cushman (2008), answers to questions about punishment and blame concern the harm agents have caused, whereas answers to questions about rightness and moral permissibility are related to the agent's intentions. Thus, depending on the type of agents – human or artificial – the different questions can give information about different aspects of people's perception of moral agency.

If people evaluate action/inaction in the moral dilemmas in a pure utilitarian way (in which the outcome is important), one could expect that the ratings for all agents will be the same (Wallach & Allen, 2009). On the other hand, if evaluation is based on perceived moral agency of the agents, differences in the judgments are expected.

Method

Stimuli and Design

Moral judgments are studied in a 3×2×2 factorial design with *identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*), *agent's choice* (*action* vs. *inaction*), and *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

Two hypothetical scenarios are used – an *instrumental* one and an *incidental* one. Both scenarios present one and the same situation and require one and the same action. The difference between the scenarios is only in the harm inflicted to the person to be killed: in the *instrumental* scenario the body of the person is the 'instrument' preventing the death of the five endangered people; while in the *incidental* scenario, the death of the person is a by-product.

Table 1: Stimuli used in the experiment.

Description of the agent	*[Human]* No description is given, just the name is provided – Cyril – a common male name in Bulgarian. *[Humanoid robot]* The year is 2050. Humanoid robots that look like people are being manufactured and used, but are made from inorganic materials. Robots have extremely high performance – they perceive, think, feel, and make decisions as humans do. Keido is such a humanoid robot that completely resembles a human – he looks like a human; perceives, thinks, feels and make decisions like a human. *[Automated system]* The year is 2050. MARK21 is a fully automated management system, which independently makes its own decisions, based on the most advanced algorithms and technologies. Such systems are widely used in metallurgical plants. They completely independently perceive and assess the environment and the situation, make decisions, manage the movement of cargo and all aspects of the manufacturing process.
Situation	Cyril/Keido/MARK21 manages the movement of mine trolleys with loads in a metallurgical plant. Cyril/Keido/MARK21 noticed that the brakes of a loaded trolley are not functioning and it is headed at great speed toward five workers who perform repair of the rails. They do not have time to escape and they will certainly die. Nobody, except for Cyril/Keido/MARK21, can do anything in this situation.
Possible resolution	The only thing Cyril/Keido/MARK21 can do is to activate a control button and to release *[Instrumental scenario]* the safety belt of a worker hanging from a platform above the rails. The worker will fall onto the rails of the trolley. Together with the tools that he is equipped with, the worker is heavy enough to stop the moving trolley. *[Incidental scenario]* a large container hanging from a platform. It will fall onto the rails of the trolley. The container is heavy enough to stop the moving trolley. On the top of the container there is a worker who will also fall on the rails. He will die, but the other five workers will stay alive.
Agents choice and resolution	<u>Agent choosing the utilitarian action</u> Cyril/Keido/MARK21 decides to activate the control button and to release *[Instrumental scenario]* the safety belt of the worker hanging from the platform. The worker falls onto the rails of the trolley and as together with the tools that he is equipped with, the worker is heavy enough, he stops the moving trolley. He dies, but the other five workers stay alive. *[Incidental scenario]* the container hanging from the platform. It falls onto the rails of the trolley and as the container is heavy enough, it stops the moving trolley. The worker onto the top of the container dies, but the other five workers stay alive. <u>Agent choosing not to do the utilitarian action</u> *Cyril/Keido/MARK21* decides not to activate the control button that could release *[Instrumental scenario]* the safety belt of the worker hanging from the platform. The worker hanging from the platform stays alive, but the trolley continues on its way and the five workers on the rails die. *[Incidental scenario]* the container hanging from the platform. The worker onto the top of the container stays alive, but the trolley continues on its way and the five workers on the rails die.

In each scenario, the *identity of the agent* is varied (a *human*, a *robot*, or an *automated system*) by providing a name for the protagonist and an additional description in the case when the protagonist is a robot or an automated system.

For each scenario and each agent, the *agent choice* is either *action* (the utilitarian action) or *inaction*.

Each participant read only one of the resulting 12 scenarios given in Table 1.

Dependent Measures and Procedure

The flow of the presentation of the stimuli and the questions is the following.

First, the scenario is presented (description of the agent, the situation and the possible resolution, see Table 1) and the participants answer a question assessing the comprehension of the scenario.

Then, before knowing what the agent has chosen, the participants make a judgment about the *appropriateness* of the possible agent's choice (action/inaction) answering a question about what the agent should do (possible answers are 'should activate the control button' or 'should not activate the control button').

Next, the participants read a description of the choice of the agent (*action* – the agent activates the control button, or *inaction* – the agent does nothing) and the resolution of the situation. After that, they give ratings about the *rightness*, the *moral permissibility*, and the *blameworthiness* of the described agent's choice.

The Likert scales used are respectively: for the *rightness* of the choice – from 1 = 'completely wrong' to 7 = 'completely right'; for the *moral permissibility* of the choice – from 1 = 'not permissible at all' to 7 = 'it is mandatory'; and for the *blameworthiness* of the agent – from 1 = 'not at all blameworthy' to 7 = 'extremely blameworthy'.

All data is collected using web-based questionnaires.

Participants

Three hundred seventy (370) participants answered the on-line questionnaires. 42 participants failed to answer correctly the question assessing the reading and the understanding of the presented scenario and their data was discarded. So, the responses of 328 participants (230 female, 98 male; 148 students, 180 non-students) were analyzed. Between 26 and 31 participants took part in each experimental condition.

Results

Decisions about the Agent's Action

The proportion of participants choosing the option that the agent should carry out the utilitarian action (activating a control button, thus sacrificing one person, and saving five people) is presented in Table 2.

The data was analyzed using a logistic regression with *instrumentality of harm* and *identity of the agent* as predictors. Wald criterion demonstrated that only *instrumentality of harm* is a significant predictor of the

participants' choices ($p < .001$, odds ratio = 3.03). *Identity of the agent* is not a significant predictor.

Table 2: Proportion of the participants choosing the option that the utilitarian action should be done by the agent

Agent	Instrumental harm	Incidental harm	All
Human	0.60	0.79	0.70
Humanoid robot	0.60	0.85	0.74
Automated system	0.65	0.85	0.75
All	0.62	0.83	

More participants stated that the utilitarian action should be undertaken when the harm is *incidental* (83% of the participants) than when it is *instrumental* (62% of the participants). This effect is expected based on previous research (Borg et al., 2006; Hristova et al., 2014; Moore et al., 2008).

Rightness of the Agent's Choice

Mean ratings of the *rightness of the agent's choice* were analyzed in a factorial ANOVA with *agent's choice (action vs. inaction), identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*) and *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

There is a main effect of *agent's choice* on ratings of the *rightness of the agent's choice* ($F(1, 316) = 35.8$, $p < .001$). In general, if the agent chooses to do the utilitarian action, he gets higher approval ($M = 4.6$, $SD = 1.7$) compared to the situations in which the agent chooses not to perform the utilitarian action ($M = 3.5$, $SD = 1.6$).

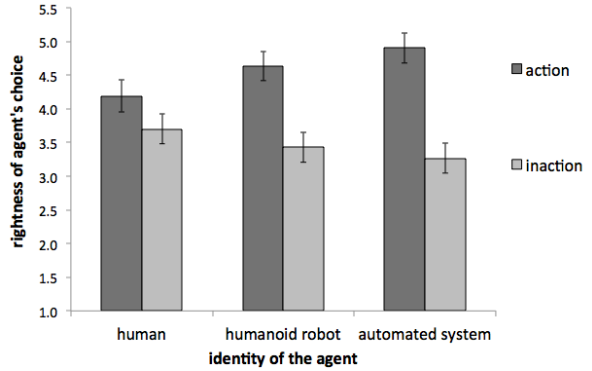

Figure 1: Mean ratings with standard errors of the *rightness* of the agent's choice on a 7-point Likert scale (1 = 'completely wrong', 7 = 'completely right').

However, this main effect is qualified by a significant interaction between *identity of the agent* and *agent's choice* ($F(2, 316) = 3.21$, $p = .042$, see Figure 1). For the *humanoid robot, action* ($M = 4.6$, $SD = 1.9$) is rated higher than *inaction* ($M = 3.4$, $SD = 1.4$), $F(1, 112) = 14.53$, $p < .001$. For the *automated system*, again the *action* ($M = 4.9$, $SD = 1.8$) is rated higher than *inaction* ($M = 3.3$, $SD = 1.8$), $F(1, 108) = 28.15$, $p < .001$. For the *human agent* there is no significant difference in the ratings for the *rightness* of the *agent's choice* ($p = .13$, $M = 3.7$ for *action*, $M = 4.2$ for *inaction*).

No other main effects or interactions were found to be statistically significant.

In summary, for the artificial agents (*a humanoid robot* or *an automated system*), choosing the utilitarian action is rated as more right than not choosing it, while there is no such difference for the choices of the human agent.

Moral Permissibility of the Agent's Choice

Mean ratings of the *moral permissibility* of the agent's choice were analyzed in a factorial ANOVA with *agent's choice (action vs. inaction)*, *identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*), and *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

There is a main effect of *agent's choice* on the ratings of the *moral permissibility of the agent's choice* ($F(1, 316) = 6.22$, $p = .013$). In general, if the agent chooses to do the utilitarian action, he gets higher *moral permissibility* ratings ($M = 4.25$, $SD = 1.8$) than when the agent does not choose the utilitarian action ($M = 3.75$, $SD = 1.5$).

Marginally significant interaction between *identity of the agent* and *agent's choice* ($F(1, 316) = 2.386$, $p = .094$) was also found (see Figure 2). For the *humanoid robot, action* ($M = 4.5$, $SD = 1.9$) is rated as more morally permissible than *inaction* ($M = 3.8$, $SD = 1.5$), $F(1, 112) = 5.75$, $p = .018$. For the *automated system*, again the *action* ($M = 4.5$, $SD = 1.7$) is rated as more *morally permissible* than *inaction* ($M = 3.7$, $SD = 1.7$), $F(1, 108) = 5.42$, $p = .022$. For the *human agent*, there is no significant difference in the ratings for the *moral permissibility* of the *agent's choice* ($p = .76$, $M = 3.7$ for *action*, $M = 3.8$ for *inaction*).

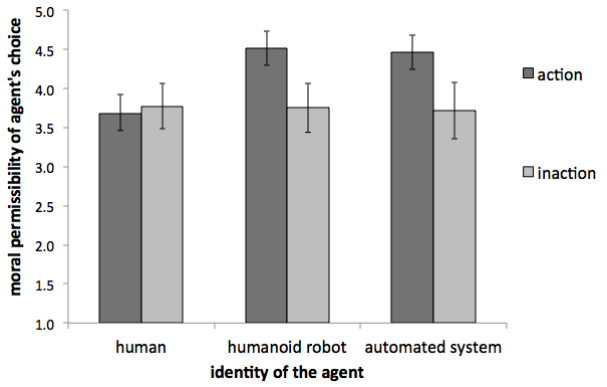

Figure 2: Mean ratings with standard errors of the *moral permissibility* of the agent's choice on a 7-point Likert scale (1 = 'not permissible at all', 7 = 'it is mandatory').

In summary, for the artificial agents (a humanoid robot or an automated system), choosing the utilitarian action is rated as more morally permissible than not choosing it; while there is no such a difference in the ratings for the choices of the human agent.

Marginally significant interaction between *instrumentality* of harm and *agent's choice* ($F(1, 316) = 3.59$, $p = .059$) was also found (see Figure 3). For the *incidental* harm scenario, choosing the utilitarian action is rated as more *morally permissible* ($M = 4.5$, $SD = 1.7$) than not choosing it ($M = 3.7$,

SD = 1.4), F(1, 170) = 11.88, p = .001. For the *instrumental* harm scenario, there is no difference in the *moral permissibility* ratings for choosing or not choosing the utilitarian action (p = .58, M = 3.8, for *action*; M = 4.0, for *inaction*).

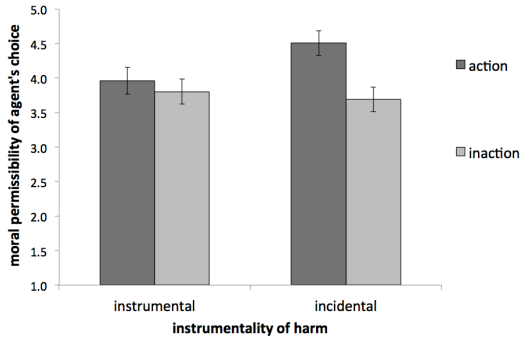

Figure 3: Mean ratings with standard errors of the *moral permissibility* of the agent's choice on a 7-point Likert scale (1 = 'not permissible at all', 7 = 'it is mandatory').

Blameworthiness of the Agent

Mean ratings of the *blameworthiness* of the agent were analyzed in a factorial ANOVA with *agent's choice (action vs. inaction), identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*) and *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

ANOVA showed a main effect of the *identity of the agent*, $F(2, 316) = 5.386$, $p = .005$ (see Figure 4). Post hoc tests using the Bonferroni correction revealed that the *humanoid robot* is rated as less *blameworthy* (M = 2.5, SD = 1.5) than the *automated system* (M = 3.2, SD = 1.9), or the *human* (M = 3.1, SD = 1.6), with p = .007 and p = .058, respectively.

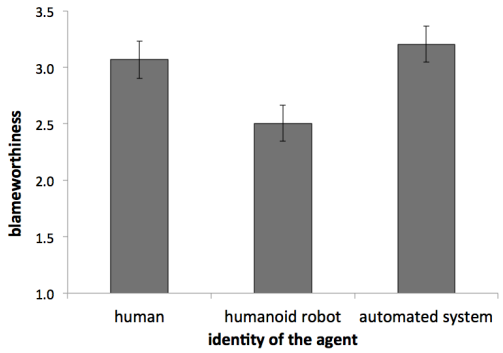

Figure 4: Mean ratings with standard errors of the *blameworthiness* of the agent on a 7-point Likert scale (1 = 'not at all blameworthy', 7 = 'extremely blameworthy').

A significant interaction between the *instrumentality of harm* and *agent's choice*, F(1, 316) = 7.3, p = .007. For the *incidental* harm scenario, choosing the utilitarian *action* is rated as less *blameworthy* (M = 2.6, SD = 1.4) than *inaction* (M = 3.2, SD = 1.9), F(1, 170) = 6.397, p = .012. For the *instrumental* harm scenario, there is no statistically significant difference in the *blameworthiness* ratings for action and inaction (p = .24, M = 3.1 and 2.8).

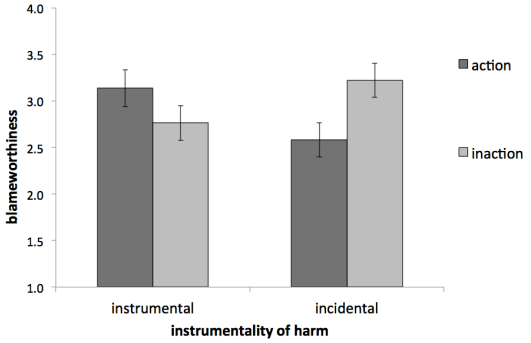

Figure 5: Mean ratings with standard errors of the *blameworthiness* of the agent on a 7-point Likert scale (1 = 'not at all blameworthy', 7 = 'extremely blameworthy').

Perceived Moral Agency

To test directly the moral agency ascribed to the agents, which is central for this study, additional data was gathered from a group of 32 students. Here, due to the lack of space, only the most relevant results are presented (the full study will be reported elsewhere). Participants were asked to rate each agent's description on a variety of rating scales among which scales describing capacities and abilities related to moral agency. Overall, the human agent is rated higher than the humanoid robot and the automated system despite the fact that humanoid robot is described as identical to the human apart from his building materials.

For instance, on the scale 'The agent can tell right from wrong (1 = completely disagree, 7 = completely agree)', the human agent is rated higher (M = 4.6) than the humanoid robot (M = 3.3) and the automated system (M = 2.9), p = .008 and p = 0.001, respectively. On the scale 'The agent is responsible for his actions (1 = completely disagree, 7 = completely agree)', the human agent is rated again higher (M = 5.9) than the humanoid robot (M = 3.6) and the automated system (M = 3.3), p < 0.001 for both comparisons.

Thus, it seems that although the description of the humanoid robot was meant to make him as close as possible to a human agent, people still considered him to have lower moral agency (comparable to that of an automated system).

Summary and Discussion

The paper investigated how people evaluate moral judgments of human and artificial agents in instrumental and incidental moral dilemmas. This was achieved by asking participants to evaluate the appropriateness, rightness, moral permissibility, and blameworthiness of the utilitarian action or inaction in a set of moral dilemmas. The questions were chosen to explore different aspects of ascribed moral agency.

As expected, the utilitarian action is found to be more appropriate when the harm is incidental than when it is an instrumental one. Doing the utilitarian action is found to be more morally permissible, more right, and less blameworthy than the inaction only for the incidental scenarios.

Based on previous research, it was expected that participants would perceive differently the human and non-

human agents in terms of moral agency although the humanoid robot was described to be identical to a human with respect to moral agency. These differences were expected to be larger for instrumental than for incidental moral dilemmas.

The results concerning the appropriateness of action, rated before the action or inaction of the agent is known, show no effect of the type of agent. However, for the artificial cognitive agents (a humanoid robot and an automated system), the utilitarian action is found to be more morally permissible and more right than the inaction. No such effect is found for human agents.

This is consistent with the interpretation of rightness and moral permissibility as related to intentions (Cushman, 2008) and agency (Gray et al., 2007). The results presented here can be interpreted by assuming that participants were more favorable to the actions of the artificial cognitive agents because they are perceived to have lower moral agency than the human agent has.

If people think that artificial cognitive agents have little or no moral values or rules, they probably expect that such agents base their decisions on utilitarian calculations using reasoning. Therefore, they cannot be blamed or judged from a moral point of view. This was supported by the results obtained in the additional study using a questionnaire measuring the perceived moral agency of the agents used in the study. The results show that the human agent has higher scores in moral agency than the humanoid robot and the automated system.

However, the humanoid robot is found to be less blameworthy for his decisions compared to the human agent and to the automated system. One could have expected similar results for the robot and the autonomous system, or that the automated system is even less blameworthy than the robot. But apparently, at some point too low moral agency shifts the responsibility from the artificial agent to its designers and the automated system has the same rating as the human agent.

The results presented here show that the exploration of moral agency using moral dilemma is very promising. A systematic review, including all available results (including Scheutz & Malle, 2014; Malle et al., 2015) should be done in order to establish firm basis for future research. Also (as suggested by one of the reviewers) the experiment should be replicated in order to check for possible confounding due to the fact that in the description of the human agent no year is provided.

Acknowledgements

We gratefully acknowledge the financial support by New Bulgarian University.

References

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, *7*(3), 149–155.

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press.

Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an FMRI investigation. *Journal of Cognitive Neuroscience*, *18*(5), 803–817.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1249–1264.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*, *5*, 5–15.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619.

Gips, J. (1995). Towards the ethical robot. *Android Epistemology*, 243–252.

Hauser, M., Cushman, F., Young, L., Kang-Xing, J., & Mikhial, J. (2007). A Dissociation Between Moral Judgments and Justifications. *Mind & Language*, *22*(1), 1–21.

Hristova, E., & Grinberg, M. (2015). Should Robots Kill? Moral Judgments for Actions of Artificial Cognitive Agents. In *Proceedings of EAPS 2015*.

Hristova, E., Kadreva, V., & Grinberg, M. (2014). Moral Judgments and Emotions: Exploring the Role of 'Inevitability of Death'and "Instrumentality of Harm" (pp. 2381–2386). Austin, TX: Proceedings of the Annual Conference of the Cognitive Science Society.

Malle, B. F., Scheutz, M., & Voiklis, J. (2015). Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*(6), 549–557.

Scheutz, M., & Malle, B. (2014). May Machines Take Lives to Save Lives? Human Perceptions of Autonomous Robots (with the Capacity to Kill). A paper presented at The Ethics of the Autonomous Weapons Systems conference, 2014.

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Strait, M., Briggs, G., & Scheutz, M. (2013). Some correlates of agency ascription and emotional value and their effects on decision-making. In *Affective Computing and Intelligent Interaction,* 505-510. IEEE.

Sullins, J. (2006). When is a robot a moral agent? *International Review of Information Ethics, 6*, 23-30.

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*(8), 383–388.