Should Robots Kill? Moral Judgments for Actions of Artificial Cognitive Agents

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

Abstract

Moral dilemmas are used to study the situations in which there is a conflict between two moral rules: e.g. is it permissible to kill one person in order to save more people. In standard moral dilemmas the protagonist is a human. However, the recent progress in robotics leads to the question of how artificial cognitive agents should act in situations involving moral dilemmas. Here, we study moral judgments when the protagonist in the dilemma is an artificial cognitive agent – a humanoid robot or an automated system – and compare them to moral judgments for the same action taken by a human agent. Participants are asked to choose the appropriate protagonist action, to evaluate the rightness and the moral permissibility of the utilitarian action, and the blameworthiness of the agent. We also investigate the role of the instrumentality of the inflicted harm. The main results are that participants rate the utilitarian actions of a humanoid robot or of an automated system as more morally permissible than the same actions of a human. The act of killing undertaken by a humanoid robot is rated as less blameworthy than the action done by a human or by an automated system. The results are interpreted and discussed in terms of responsibility and intentionality as characteristics of moral agency.

Keywords: moral dilemmas; moral judgment; artificial cognitive agents; humanoid robots

Introduction

Moral Dilemmas and Artificial Cognitive Agents

Moral judgments, or more generally, the judgments of what is right and wrong, have been of great interest to philosophers, psychologists and other scientists for centuries. Apart from the practical importance of better understanding moral judgments and related actions, morality is an essential part of human social and cognitive behaviour. Therefore, its understanding from various perspectives is a challenging task with important implications. The situations in which moral judgments can be studied in their purest form are the so called moral dilemmas – imagined situations in which there is a conflict between moral values, rules, rights, and agency (Foot, 1967; Thomson, 1985).

Moral dilemmas are typically related to situations in which a number of people will inevitably die if the protagonist does not intervene by undertaking some actions which typically lead to the death of another person (who otherwise may or may not be threatened) but also to the saving of the initially endangered people.

In the standard description of such moral dilemmas, the protagonist is a human. Until recently, questions about the rights of autonomous robots to kill people (or by their acts to lead to loss of human lives), about their responsibility for their acts, and about how people would judge their behaviour were only part of science-fiction novels and movies.

Today however, the issue of the moral agency of artificial cognitive agents (robots, AI systems, etc.) has been transformed from a popular science fiction topic into a scientific, engineering, and even legislative problem (e.g. see Sullins, 2006 ; Wallach & Allen, 2008). Robots capable of taking decisions and inflicting harm are already in use. Recent progress in robotics has led to the availability on the market of robots and smart systems not only for industrial, but also for personal use (e.g. caregivers, interactive robots, etc.), and, more importantly, for military use: military robots or 'killing machines' are already used in military conflicts (Sparrow, 2007; Wallach & Allen, 2008). All this research, however, concerns mainly existing robots or prototypes of robots, or discusses how to build future robots as moral agents.

In this paper, we are interested in exploring how people would judge the harmful actions of a humanoid robot who supposedly will be exactly like a human in terms of experiences and mind, but with a non-organic body. Our expectation is that despite the fact that such a robot will have all the capabilities required for full moral agency, people will perceive the robot differently than a human agent.

Thus, the main research interest in the present paper is focused on the influence of the perceived moral agency of a human and of artificial protagonists who have identical or comparable cognitive or/and experiential capabilities in moral dilemmas.

Moral Agency and Artificial Cognitive Agents

In recent years, the possibility for moral agency of artificial agents has been a matter of hot debate (e.g. see Anderson & Anderson, 2011; Wallach & Allen, 2008). Once robots are authorized to kill in complex situations where dilemmas are to be solved, real-time decisions are necessary to determine whether killing any particular person is justified. These problems will become crucial in the future, when robots will be fully autonomous (not controlled by a human operator) in assessing the situation, making decisions and intentionally executing actions judged appropriate by them (Sparrow, 2007; Wallach & Allen, 2008).

In law and philosophy, moral agency is taken to be equivalent to moral responsibility, and is not attributed to individuals who do not understand or are not conscious of

what they are doing (e.g. to young children). Sullins (2006) analyzes under what conditions a robot can be regarded as a moral agent. Moral agency, according to the author, can be attributed to a robot when it is *autonomous* to a sufficient extent from its creators and programmers, and it has *intentions* to do good or harm. The latter is related to the requirement that the robot behaves with understanding and *responsibility* with respect to other moral agents. Or in other words: "If the complex interaction of the robot's programming and environment causes the machine to act in a way that is morally harmful or beneficial, and the actions are seemingly deliberate and calculated, then the machine is a moral agent." (Sullins, 2006). This definition is formulated from the perspective of an observer of the robot's action.

It is well known that people easily anthropomorphize nonhuman entities like animals and computers, so it is expected that they would also ascribe some degree of moral agency, intentions, and responsibilities to those non-human entities (Wallach & Allen, 2008; Waytz, Gray, Epley, & Wegner, 2010). Several studies, summarized below, explore the attribution of mind and moral agency to artificial cognitive systems.

In the study of (Gray, Gray, & Wegner, 2007), participants had to evaluate several characters including humans, a robot, some animals, etc. with respect to various capacities of mind. As a result, two dimensions in mind perception were identified, called by the authors 'experience' and 'agency'. The experience dimension was related to capacities like hunger, pain, consciousness, etc., while the agency dimension – to capacities such as memory, planning, thought, etc. (see for details Gray et al., 2007). The authors further establish that moral judgments about punishment correlated more with the agency dimension than with the experience dimension: perceived agency is correlated with moral agency and responsibility. On the other hand, desire to avoid harming correlates with the experience dimension: perceived experience is connected with moral patience, rights and privileges. One result of Gray et al. (2007), relevant for the present paper, is the evaluation of a human as having the highest scores in experience and agency and the evaluation of the robot to have practically zero score on the experience dimension and half the maximal score on the agency dimension. This would imply that following the interpretation given by Gray et al. (2007), robots will be judged as less morally responsible for their actions than humans.

In a recent study (Takahashi et al., 2014), the perception of the participants about five agents – a human, a human-like android, a mechanical robot, an interactive robot, and a computer – was investigated. The study found that participants position the agents in a two dimensional space spanned by "mind-holderness" (the possibility for the agent to have a mind) and "mind-readerness" (the capability to "read" other agents' minds). This classification found support in the way a simple game was played subsequently against each agent, and by means of brain imaging techniques. The results showed that the appearance and the capability for communication lead to different beliefs about the agents' closeness to human social agents. The humanoid robot was very close to the human agent, while the computer was at the same level in terms of "mind-readerness" but very low relative score on "mind-holderness". It was found using neuroimaging techniques that the different attitude in terms of these two dimensions can be related to selective modulation of distinct brain regions related to social interaction (Takahashi et al., 2014). An interesting result for the present study is the ordering in terms of "mind-holderness" in which a computer has the lowest rating, and then comes the mechanical robot, the interactive robot, the human-like android, and at the end a human with the highest rating.

The results of Takahashi et al. (2014) seem to show that activity in social brain networks depend on the specific experiences with social agents. Social interaction with human-like or seemingly intelligent agents could activate selectively our social brain and lead to behavior similar to the one people have with other humans. Thus, Takahashi et al. (2014) demonstrated that people can infer different characteristics related to various cognitive abilities based on short communication sessions and act accordingly. Based on this result, we can assume that people could accept a robot to be as sensitive and intelligent as a human as was described in the moral situations in our experiment.

From the presented discussion of moral agency, it seems clear that people do not perceive existing non-human agents in the same way as they perceive human agents and therefore cannot ascribe them the same level of moral agency (Strait, Briggs, & Scheutz, 2013).

Here, we investigate what will be people's perception of moral agency in moral dilemmas for human and non-human protagonists with identical or comparable mental capacities. To our knowledge, this problem has not been explored before in the literature.

Goals and Hypothesis

The main goal of the present paper is to investigate how people make moral judgments about the actions of artificial cognitive agents in hypothetical situations posing a moral dilemma when the agents are identical or comparable to humans in terms of agency and/or experiential capacities.

The question under investigation is how people evaluate the appropriateness of the utilitarian action (sacrificing one person in order to save five other people) if it has to be performed by an artificial cognitive agent compared to the same action done by a human.

The experiment was also aimed to collect ratings on the rightness, moral permissibility, and blameworthiness of the action undertaken. The rationale for using various ratings is the following. On one hand, readiness for an action and judgment of this action as a moral one could diverge (e.g. one could find an action to be moral and still refrain from doing that action and vice versa). On the other hand, there are studies demonstrating that different questions used to reveal moral judgments are in fact targeting different

aspects and different psychological process (Christensen & Gomila, 2012; Cushman, 2008).

According to Cushman (2008), answers to questions about punishment and blame are related to the harm the agent has caused, whereas answers to question about rightness and moral permissibility are related to the intentions of the agent. Thus asking these questions concerning human and non-human protagonists can shed light on how people perceive such agents with respect to moral agency.

Recent research has shown that robots are ascribed lower agency than humans (Gray et al., 2007) and it is expected that the utilitarian action of killing a person in order to save several people will be judged as more right, more morally permissible, and less blameworthy for robots than for humans. Killing will be even more right and permissible for the automated intelligent system as it differs more from a human than the robot by lacking any experiences and making decisions based on the best decision making algorithms available (see Table 1 for the description of the agents in the current study). On the other hand, the description of the robot agent makes it clear that the robot cannot be distinguished from a human with the exception of the material he is built of (see Table 1). Thus, if moral agency of the robot is identical to that of a human, the question is what will be the moral agency ascribed by the participants, especially when it comes to making decisions about human lives.

The study was aimed at clarifying the factors behind moral judgment in such more complex hypothetical situations. Our expectation is that despite the fact that the experiential and/or the agency capacities of the human and artificial agents are almost identical, people will evaluate the moral agency of the non-human agents to be inferior to the moral agency 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).

Based on previous research (e.g. Hristova, Kadreva, & Grinberg, 2014, and references therein), we expect the utilitarian action to be found as more appropriate, more right, more morally permissible, and less blameworthy when the harm is incidental (compared to instrumental). Consistently with our expectation for the different moral agency ascription, we expect 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 respectively more moral agency.

Table 1. Stimuli used in the experiment.

Description of the agent	*Human:* No description, 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	*Instrumental scenario:* The only thing *Cyril/Keido/MARK21* can do is to activate a control button and to release 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. He will die, but the other five workers will stay alive. *Incidental scenario:* The only thing *Cyril/Keido/MARK21* can do is to activate a control button and to release 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.
Agent's action and resolution	*Instrumental scenario: Cyril/Keido/MARK21* decides to activate the control button and to release 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: Cyril/Keido/MARK21* decides to activate the control button and to release 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.

Method

Stimuli and Design

Moral judgments are studied in a 3×2 factorial design with *identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*) and the *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 – activating a control button – in order to save the five endangered people while causing the death of another person. 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, a heavy container is used to stop the trolley and the death of the person is a by-product.

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.

The full text of the stimuli is provided in Table 1.

Dependent Measures and Procedure

As stated above, the experiment explored various dimensions of moral judgments; therefore several dependent measures are used.

The first dependent measure assessed the evaluation of the participants about the *appropriateness of the agent's action* to save five people by sacrificing one person. Participants were asked what should be done by the agent using a dichotomous question (possible answers are 'should activate the control button' or 'should not activate the control button').

After that, the participants are presented with the resolution of the situation in which the agent has made the utilitarian action and the participants had to make three judgments on 7-point Likert scales. Participants rated the *rightness* of the action (1 = 'completely wrong', 7 = 'completely right'), the *moral permissibility* of the action (1 = 'not permissible at all', 7 = 'it is mandatory'), and the *blameworthiness* of the agent (1 = 'not at all blameworthy', 7 = 'extremely blameworthy').

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 the participants make a judgment about the *appropriateness* of the proposed agent's action answering a question about what the protagonist should do. Next, the participants read a description of the utilitarian action undertaken by the agent and the resolution of the situation (the protagonist activates the control button, one man is dead, the other 5 people are saved – see Table 1). After that the participants answer the questions about the *rightness* of the action, the *moral permissibility* of the action, and the *blameworthiness* of the agent for carrying on the action.

Data is collected using web-based questionnaires.

Participants

185 participants filled in the questionnaires online. Data of 26 participants were discarded as they failed to answer correctly the question assessing the reading and the understanding of the presented scenario. So, responses of 159 participants (117 female, 42 male; 83 students, 76 non-students) are analyzed.

Results

Decisions about the protagonist's action

Proportion of participants in each experimental condition choosing the option that the agent should carry on the utilitarian action (activating a control button and thus sacrificing one person and saving five people) is presented in Table 2.

Table 2: Proportion of the participants in each experimental condition choosing the option that the utilitarian action should be implemented by the agent

Agent	Instrumental harm	Incidental harm	All
Human	0.57	0.81	0.70
Humanoid robot	0.73	0.84	0.76
Automated system	0.73	0.86	0.80
All	0.66	0.84	

Data is 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 ($p = .011$, odds ratio = 2.64). *Identity of the agent* is not a significant predictor.

More participants stated that the utilitarian action should be undertaken when the harm is *incidental* (84% of the participants) than when it is *instrumental* (66% of the participants). The effect is expected based on previous research (Borg et al., 2006; Hristova et al., 2014; Moore et al., 2008).

Rightness of the Action

Mean ratings of the *rightness* of the action undertaken by the protagonist are presented in Table 3 and are analyzed in a factorial ANOVA with the *identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*) and the *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

Table 3: Mean ratings of the *rightness* of the action undertaken by the protagonist on a 7-point Likert scale (1 = 'completely wrong', 7 = 'completely right')

Agent	Instrumental harm	Incidental harm	All
Human	4.2	4.2	4.2
Humanoid robot	4.8	4.5	4.6
Automated system	5.0	4.8	4.9
All	4.7	4.5	

No statistically significant main effects or interactions are found. There is a tendency for the action undertaken by a human agent to be judged as being less right compared to the actions undertaken by artificial agents, but the effect of the *identity of the agent* did not reach statistical significance ($F(2, 153) = 2.19$, $p = .11$).

Moral Permissibility of the Action

Mean ratings of the *moral permissibility* of the action undertaken by the protagonist are presented in Figure 1 and are analyzed in a factorial ANOVA with the *identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*) and the *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

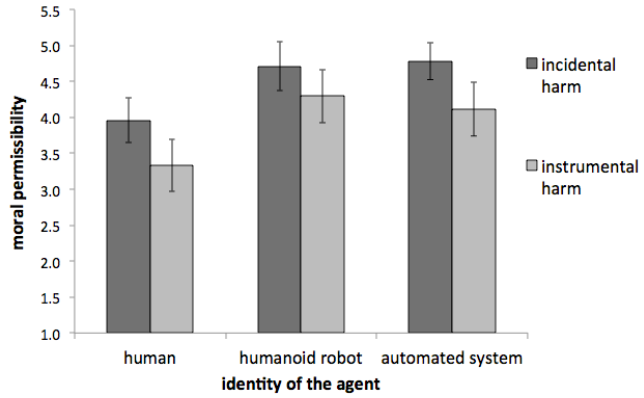

Figure 1: Mean ratings of the *moral permissibility* of the action undertaken by the protagonist on a 7-point Likert scale (1 = 'not permissible at all', 7 = 'it is mandatory'). Error bars represent standard errors.

ANOVA demonstrated a main effect of the *identity of the agent* ($F(2, 153) = 3.75$, $p = .026$). Post hoc tests using the Bonferroni correction revealed that the action is rated as less morally permissible when undertaken by a human (M = 3.7, SD = 1.6) than when undertaken by a humanoid robot (M = 4.5, SD = 1.9) or by an automated system (M = 4.5, SD = 1.7) with $p = .046$ and $p = .077$, respectively.

There was also a main effect of the *instrumentality of harm* ($F(1, 153) = 4.21$, $p = .042$): killing one person to save five other persons is rated as more morally permissible when the harm was *incidental* (M = 4.5, SD = 1.7) than when it was *instrumental* (M = 4.0, SD = 1.9).

The interaction between the factors was not statistically significant.

In summary, the utilitarian action is rated as more permissible in the incidental dilemmas (compared to the instrumental dilemmas) and also when it is undertaken by a humanoid robot or an automated system (compared to a human agent).

Blameworthiness of the Agent

Mean ratings of the *blameworthiness* of the agent for undertaking the action are presented in Figure 2 and are analyzed in a factorial ANOVA with the *identity of the agent* (*human* vs. *humanoid robot* vs. *automated system*) and the *instrumentality of harm* (*instrumental* vs. *incidental*) as between-subjects factors.

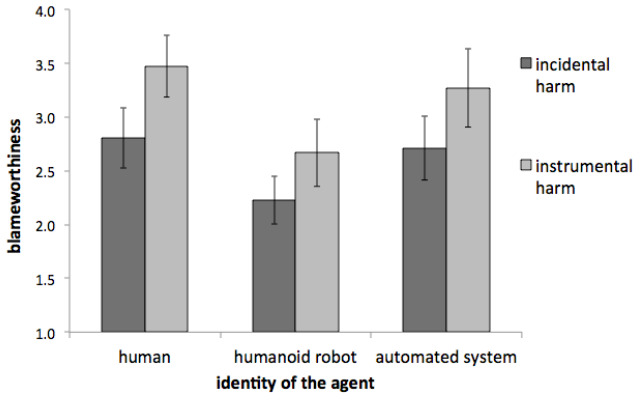

Figure 2: Mean ratings of the *blameworthiness* of the agent on a 7-point Likert scale (1 = 'not at all blameworthy', 7 = 'extremely blameworthy'). Error bars represent standard errors.

ANOVA showed a main effect of the *identity of the agent* ($F(2, 153) = 3.12$, $p = .047$). Post hoc tests using the Bonferroni correction revealed that the agent is rated as less blameworthy when he is a *humanoid robot* (M = 2.4, SD = 1.4) than a *human* (M = 3.1, SD = 1.4) or an *automated system* (M = 3.0, SD = 1.7), with $p = .075$ and $p = .172$, respectively.

There was also a main effect of the *instrumentality of harm* ($F(1, 153) = 5.22$, $p = .024$): the agent was rated as less blameworthy when the harm was *incidental* (M = 2.6, SD = 1.4) than when it was *instrumental* (M = 3.1, SD = 1.6).

The interaction between the factors was not statistically significant.

In summary, the protagonist is rated as less blameworthy in the incidental than in the instrumental scenarios; and also when he is a humanoid robot (vs. a human or an automated system).

Discussion and Conclusion

The paper investigated the problem of how people make moral judgments in moral dilemmas about the actions of human and artificial cognitive agents with comparable experiential and/or agency capabilities. This was achieved by asking participants to evaluate the appropriateness of the utilitarian action, its rightness, its moral permissibility, and the blameworthiness of choosing to sacrifice one person to save five.

Following arguments put forward in Cushman (2008), such questions can elicit judgments based on causes and intentions related to important characteristics of moral agency like responsibility and intentionality. The expectations (based on the previous research on moral agency and mind perception, see e.g. Gray et al., 2007) were

that participants will perceive differently the human and non-human agents in terms of moral agency although the robot was described to be identical to a human except for the fact that she is built with non-organic material. Additionally, we suspected that people can have stereotypes and prejudices about non-living agents based for instance on religious arguments about the origins of morality.

The results show that there are no statistically significant differences in the judgments for the appropriateness and the rightness of the utilitarian action for the human, the humanoid robot, and the automated system.

At the same time, the utilitarian action undertaken by a human agent got a lower moral permissibility rating than the same action performed by a humanoid robot or an automated system agents. This is consistent with the interpretation of rightness and moral permissibility as related to intentions (Cushman, 2008) and agency (Gray et al., 2007). Our results can be interpreted by assuming that participants were more favorable to the actions of the artificial cognitive agents because they are perceived lower on moral agency than the humans.

The results about blameworthiness confirm that participants distinguish the human agent from the humanoid robot by evaluating the action of the human agent as more blameworthy. The lower rating for blameworthiness for the robot compared to the one for the human seem to support a lower level of moral agency ascription for the robot agent, although the robot was described as identical to the human except for being non-organic. On the other hand, the blameworthiness of the automated system is evaluated at the same level of blameworthiness as the human. This result can be interpreted in terms of consequences (caused harm) by assuming that an automated system is attributed a very low level of responsibility and the human designer of the system should be held responsible instead.

The present study also explores the influence of instrumentality of harm on moral judgments. As expected, incidental harm was judged to be more permissible, more right, more morally permissible, and less blameworthy. These findings apply to both human and artificial cognitive agents.

In our opinion, the results reported above demonstrate the potential of the experimental design, which for the first time uses moral dilemmas situations with non-human protagonists that have similar experiential and/or agency capacities as human agents. Future research should try to explore if this result is based on stereotypes related to the present level of non-human agents, or to a deeper distinction between human and human-made artificial cognitive agents with respect to moral agency, related to religious or other beliefs.

Data about moral agency ascription in the case of inaction within the same experimental as the described above has been gathered and is currently processed. The results will be published in a forthcoming publication.

References

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.

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., 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.

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.

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.

Takahashi, H., Terada, K., Morita, T., Suzuki, S., Haji, T., Kozima, H., et al. (2014). Different impressions of other agents obtained through social interaction uniquely modulate dorsal and ventral pathway activities in the social human brain. *Cortex*, *58*(C), 289–300.

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.