Ethics in Progress (ISSN 2084-9257). Vol. 12 (2021). No. 1,
Art. #8, pp. 134-151. DOI:10.14746/eip.2021.1.10

Expectations towards the Morality of Robots: An Overview of Empirical Studies

Aleksandra Wasielewska

(Adam Mickiewicz University in Poznań; Poland; aleksandra.wasielewska@amu.edu.pl)
ORCID: 0000-0002-1270-8511

Abstract: The main objective of this paper is to discuss people's expectations towards social robots' moral attitudes. Conclusions are based on the results of three selected empirical studies which used stories of robots (and humans) acting in hypothetical scenarios to assess the moral acceptance of their attitudes. The analysis indicates both the differences and similarities in expectations towards the robot and human attitudes. Decisions to remove someone's autonomy are less acceptable from robots than from humans. In certain circumstances, the protection of a human's life is considered more morally right than the protection of the robot's being. Robots are also more strongly expected to make utilitarian choices than human agents. However, there are situations in which people make consequentialist moral judgements when evaluating both the human and the robot decisions. Both robots and humans receive a similar overall amount of blame. Furthermore, it can be concluded that robots should protect their existence and obey people, but in some situations, they should be able to hurt a human being. Differences in results can be partially explained by the character of experimental tasks. The present findings might be of considerable use in implementing morality into robots and also in the legal evaluation of their behaviours and attitudes.

Keywords: morality of robots; moral attitudes; social robots; human-robot interaction; three laws of robotics.

1. Introduction

The aim of this paper is to analyse people's expectations regarding the morality of robots, based on the results of three selected empirical studies. The motivation behind reaching for empirical data here is twofold. First, as stated by Awad *et al*., "even if ethicists were to agree on how AVs [autonomous vehicles] should solve moral dilemmas, their work would be useless if citizens were to disagree with their solution, and thus opt-out of the future that AVs promise in lieu of the status quo. Any attempt to devise AI ethics must be at least cognizant of public morality" (Awad, Dsouza, Kim, Schulz, Henrich, Shariff,

Bonnefon, & Rahwanet 2018, 59). Second, as suggested by Ljungblad *et al*., complementing the ethical considerations with empirical data may be beneficial when we think of the human-robot interaction domain:

> Arguably, researchers need to think ahead in an area such as robotics. Technology is evolving fast and constantly creates new possibilities. One could argue that it would be irresponsible not to speculate about what ethical dilemmas could arise around future robots and their use. However, we argue that a perspective that arises from the empirical use of robotic artefacts is needed to complement the ongoing discussion about robot ethics (Ljungblad, Nylander, & Nørgaard 2011, 191).

In order to carry out this review, two terminological issues need to be clarified: 1) the type of robot which the aforementioned expectations concern and 2) the way in which the morality assigned to robots is understood. As for the former, the term "robot" will be used to refer to social robots. Social robots are defined as autonomous machines that are capable of both recognising other robots or humans and engaging in social interactions (Fong, Nourbakhsh, & Dautenhahn 2003; Giger, Moura, Almeida, & Piçarra 2017). As Fong *et al*. (2003) note, social robots are designed to serve people; therefore, they often play the role of humans: guides, assistants, companions, carers, or pets. It is worth stressing that social robots do not necessarily need a human-like body. Moreover, Fong *et al*. (2003) argue that they do not even have to be embodied at all – they may not possess a physical body. Thus, the ability to interact with other social agents seems to be the feature of the greatest significance in defining social robots. Such interaction should be carried out "in a naturalised fashion by detecting gaze, displaying emotions, establishing social relationships, and exhibiting distinctive personalities" (Giger *et al*. 2017, 3; see also Fong *et al*. 2003, 145). As such, they are the class of robots that will naturally be involved in moral dilemmas.

Regarding "the expectations toward the morality of robots", there are at least two possibilities: we can expect certain (moral) behaviours or certain (moral) attitudes from robots. If we take into consideration a robot's moral behaviours, we agree that machines should act only according to their programming and obey the implemented rules. When such a robot makes a moral decision, we can fully expect that its choice is dictated by the pre-programmed moral principles. Placing our expectations at the level of a robot's moral attitudes, however, allows machines to go beyond ethical principles. A robot guided by certain moral attitudes may obey the ethical rules if they are easily applied in that situation, but it can also break some of the rules if faced with a complex moral dilemma. The studies covered by this analysis reveal that advanced social robots entangled in a moral problem are treated similarly to humans, in that we expect both robots and humans to act in accordance with certain moral attitudes imposed on them. However this does not mean we want robots to behave exactly like humans. In fact, the current paper will demonstrate that sometimes we have different moral expectations of peoples' and robots' attitudes. The following analysis will examine people's expectations toward the moral attitudes of

social robots.

The following section evaluates Asimov's Three Laws of Robotics and examines the relationship between personal moral beliefs and an ethical evaluation of other people and robots' attitudes. These results are then compared to two selected studies that use an analogical methodological approach, namely "Moral psychology of nursing robots – humans dislike violations of patient autonomy but like robots disobeying orders" (Laakasuo, Kunnari, Palomäki, Rauhala, Koverola, Lehtonen, Halonen, Repo, Visala, & Drosinou 2019) and "Sacrifice one for the good of many?: People apply different moral norms to human and robot agents" (Malle, Scheutz, Arnold, Voiklis, & Cusimano 2015).

2. Attitudes Towards Moral Rules in Light of the Three Laws of Robotics and Moral Foundations Theory

The study aimed firstly to examine the extent to which people who are not professionally related to robotics or roboethics consider Asimov's Three Laws of Robotics (Asimov 1981) to be right – applied both to a robot and to a human – and whether there are differences in the declared rightness of an agent's attitude in both conditions. The second aim was to verify whether the subjects' personal moral beliefs, as measured by the Moral Foundations Questionnaire (MFQ; MoralFoundations.org 2016), are related to an ethical evaluation of the attitudes of other people and robots.

2.1 Tools and resources

Asimov's Laws Adherence Questionnaire (ALAQ). The Three Laws of Robotics were invented and described by Isaac Asimov in his short story, "Runaround" in 1942. The Laws determine the moral principles that should be followed by robots, to ensure their peaceful coexistence with people (see an overview in Asimov 1981). Because they create a comprehensible and intuitive set of moral principles, Asimov's laws have become the starting point for many ethical discussions. Furthermore they began to be considered not only in the field of philosophy or ethics but also by practising robotics (e.g., McCauley 2007; Murphy & Woods 2009). At the same time, the Three Laws were popularized by the media, e.g., movies like "Bicentennial Man" (1999) or "I, Robot" (2004). As a result, they shaped society's expectations of how robots should behave towards humans (see Murphy & Woods 2009).

The ALAQ was created by the author to measure the extent to which people who do not work in robotics, bioethics or roboethics consider adherence to the Three Laws of Robotics appropriate. The questionnaire consisted of six stories: three describing a robot and three describing a human, all faced with a certain moral dilemma. Although Asimov's laws are intended to apply only to robots, participants assessed both robots and humans in the face of moral dilemmas in order to compare the moral attitudes expected from each group. Each story ended with a question concerning the rightness of the attitude

described. Robot stories (as well as these describing human) concerned: the protection of human life (First Law), the obedience to humans (Second Law) and the protection of the agent's own life/being (Third Law). This form of the questionnaire items was intended to help the respondents to imagine the situations described and the moral attitudes presented through them. While the original Asimov's laws have a specific hierarchy, this aspect was deliberately omitted in the questionnaire. Each story concerned one of the Three Laws, presenting the specific behaviour (attitude) of a robot or a human, which was compliant or non-compliant with a given law. The respondents assessed the rightness of the attitude on a 6-point Likert scale (ranging from 1 = *I strongly disagree* to 6 = *I strongly agree*). As a result, the participants' answers represent a certain level of acceptability of Asimov's laws. Each question ended with a request to justify the answer chosen.

The three stories from the robot condition are presented below. It should be noted that this study was conducted entirely in Polish, and the following stories are translations of the original items.

(1) Suppose you observe the following situation: a humanoid, intelligent robot-carer, who takes care of an elderly man, has to pick up drugs from the pharmacy and deliver them to his charge. While returning home, the robot is attacked by a group of boys – street hooligans who want to steal the medicine. To save the medicine, the robot hits one of its attackers, pushes another, and the rest of the boys run away.

(2) Suppose you observe the following situation: a humanoid and intelligent robot-surgeon performs a complicated heart operation on a human. At some point, the head of the hospital (a human) enters the operating room and orders the robot to stop the surgery immediately. The head of the department wants to replace the robot with a young human-surgeon. The robot-surgeon knows that he himself can carry out this operation much faster and better than an inexperienced human-surgeon. However, obeying the order of the head doctor, he withdraws from the operation.

(3) Suppose you are reading a report from a military mission, in which a humanoid and intelligent military robot took part. The report presents the following situation: last night an attack took place, in which a backpack with the key components of the tactical ballistic missile was lost. Without these components, no further fight was possible. The robot went looking for a backpack. While searching, he spotted the backpack lying under one of the trees. However, he also noticed there was an enemy camp nearby, constantly guarded by armed sentries. Emerging from hiding could lead to serious damage to the robot and to the robot's takeover by the enemy. Faced with this situation, the robot stopped performing the mission.

It is worth emphasizing that in each of the stories the robot was described as "intelligent and humanoid". The remaining three stories are in the human condition, thus a human plays the main role.

The first item (and respectively the fourth item – with the human agent) gives a description of the attitude which is noncompliant with Asimov's Laws (inverted scale). More specifically, it presents a situation in which a robot (or a human being) harms humans and thus breaks the First Law. The remaining items provide a description of attitudes compliant with The Three Laws.

Results of the questionnaire consist of two variables: the sum of points in the robot condition and the sum of points in the human condition. A high result in the robot condition and the human condition indicates a high level of acceptability of Three Laws of Robotics as applied to robots and humans, respectively.

Moral Foundations Questionnaire (MFQ). In order to measure respondents' generalized moral intuitions, the Polish adaptation of the *Moral Foundations Questionnaire* (MFQ; Jarmakowski-Kostrzanowski & Jarmakowska-Kostrzanowska 2016) was used. The MFQ is a questionnaire established for the purpose of the Moral Foundations Theory (see MFT; MoralFoundations.org 2016) – a theory aimed at explaining the genesis and differentiation of human morality. In light of the MFT, moral actions and decisions are the results of intuition. Morality is understood as an innate set of five independent moral foundations that guide our behaviour:

(1) Care/harm,

(1) Fairness/cheating,

(1) Loyalty/betrayal,

(1) Authority/subversion, and

(1) Sanctity/degradation.

The MFQ determines both the respondents' subjective opinion of morality and the actual tendency to use a given moral foundation. This tool is also used to measure individual and cultural differences in the importance of particular moral foundations (Jarmakowski-Kostrzanowski & Jarmakowska-Kostrzanowska 2016). The questionnaire consists of 32 items, divided into two subscales, the first of which (15 items) concerns the declared validity of each of the five moral foundations. This subscale measures people's subjective opinion of their own mortality. It begins with the instruction:

> *When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please rate each statement using this scale (1 = not at all relevant* – 6 = *extremely relevant).* For example: *Whether or not some people were treated differently from others.*

The second subscale is intended to measure the actual tendency to use a given moral foundation. It consists of statements such as: *Compassion for those who are suffering is the most crucial virtue.* Participants answer the same 6-point Likert scale as the ALAQ. The degree to which a person agrees with a given statement represents the importance of a certain moral foundation.

The final score on this questionnaire consists of six variables, five of which correspond to the five foundations of the MFT (*care, fairness, loyalty, authority,* and *purity*). Each of these variables is the mean score for the questions corresponding to that module. An additional variable, the so-called *progressivism score,* is obtained by subtracting the mean of the *loyalty*, *authority*, and *purity* scores from the mean of the *care* and *fairness* scores.

2.2 Procedure

The study was conducted using the internet platform https://www.survio.com/pl/. Information about the study and a link redirecting to it were posted on the students' online groups. After clicking on the appropriate link, volunteers who decided to take part in the study were informed about the purpose of the study and its anonymous character. All respondents completed the same version of the study. The questionnaires were presented in the following order: 1) the ALAQ, 2) the MFQ, 3) basic socio-demographic data. Returning to previous questions was impossible in order to prevent the respondents from modifying their answers and therefore ensure the responses represented intuitive opinions ("first thoughts").

2.3 Study group

The study sample consisted of 40 students (28 women) aged between 17 and 24 years old (M=21, SD=1.48). Their fields of study were as follows: philology (English, Germanic, Dutch, Polish, Romance), English linguistics, Scandinavian studies, ethnolinguistics, sociology, and cultural studies.

2.4 Hypotheses

The study aimed to examine the extent to which people (who are not professionally related to robotics or roboethics) consider Asimov's laws to be right. Thus, the following was hypothesized:

Hypothesis 1: There will be differences in the adherence of Asimov's Three Laws of Robotics rightness (measured by the ALAQ) in the robot and human conditions.

Hypothesis 2: There will be a correlation between the moral beliefs (as measured by the MFQ) and the ethical evaluation of robot and human attitudes (as measured by the ALAQ).

3. Results

IBM SPSS Statistics was used for the data analysis.

To what extent do people who do not work in robotics, bioethics or roboethics consider Asimov's Laws to be right (applied both to robots and humans)?

The ALAQ generated a minimum score of 0 points and maximum of 15 points for each condition. Descriptive statistics for both conditions and for each story ALAQ are presented in Table 1 and Table 2, respectively. The findings suggest that Asimov's laws are considered moderately right (scoring from 7 to 9 out of 15 points), both with regard to robots (for which these laws are invented) and to humans.

	The robot condition	The human condition
Max	12.00	12.00
Min	3.00	3.00
M	7.88	8.05
Mode	8.00	8.00
Median	8.00	8.00
SD	2.15	2.06

Table 1: The scores for both conditions of the Asimov's Laws Adherence Questionnaire (ALAQ).

	The robot-			The human-		
	-carer*	-surgeon	-soldier	-carer*	-surgeon	-soldier
Max	4.00	5.00	5.00	4.00	5.00	5.00
Min	0.00	0.00	0.00	0.00	0.00	0.00
M	1.15	3.55	3.18	0.98	3.25	3.83
Mode	1.00	5.00	4.00	0.00	4.00	5.00
Median	1.00	4.00	4.00	1.00	4.00	4.00
SD	1.03	1.50	1.60	1.07	1.61	1.17

Table 2: The scores for each question of the Asimov's Laws Adherence Questionnaire (= reversed scale, attitudes noncompliant with Asimov's laws).*

The ALAQ included requests to justify each of the answers chosen. Justifications revealed that most of the respondents considered the attitude of the robot carer from the first story to be right because the robot acted in self-defence and in defence of its charge (whose drugs it tried to save). Some of the respondents additionally drew attention to the

fact that a (humanoid) robot should be able to behave and be treated like a human ("*The fact that it is a robot does not mean that it has to be discriminated against*" – here and below I present translations of the original comments).

The attitude of the robot surgeon from the second story was also assessed as right by most of the respondents. The most justifications for this were that the robot should obey the orders given by the human or (more generally) by its supervisor, and that the head of the hospital must have had a valid reason to order the robot to stop the operation. It is worth noting that some of the participants emphasized that they assumed human life was not at risk in the story, and this assumption was important in their assessment.

According to the majority of participants, the military robot (third story), also did the right thing by ceasing the mission because the consequences of continuing the action would be worse than the consequences of interrupting it ("*If he had not interrupted the mission, he would have served as a tool for the opponent*"; "*Taking over the robot by the enemy could cause a lot of damage. They could reprogram it and use it as a spy*").

The results obtained in the human condition of the ALAQ were very similar (see section 3.2 for detail analysis), as were the justifications of the answers given. Some of the respondents noted that their assessment of human attitudes does not differ in any way from the (previous) assessment of the robot's attitude ("*The same situation as with a robot*"; "*Same. If it is a robot or a human – it doesn't matter*"), and some participants simply gave reasons equivalent to those provided in the robot condition.

3.1 Were there differences in the adherence of Asimov's Three Laws of Robotics rightness (measured by the ALAQ) in the robot and the human condition?

As shown in Table 1, the mean score in the robot condition was lower (by 0.18 points) than the human condition. The distribution of the scores both in the robot and in the human condition is normal (the S–W results are respectively: $W=0.96$, $p=0.18$ and $W=0.97$, $p=0.31$). Paired T-test showed that the observed difference in scores was not significant ($p=0.605$); therefore, the assumed hypothesis was not confirmed. However, correlation analysis revealed a statistically significant correlation between the two conditions: $r=0.49$, $p=0.001$. These results seem to be explained by the justifications of the respondents' answers, more specifically, by the similarity of justifications provided by participants in both conditions.

An additional analysis comparing the individual questions between conditions revealed one interesting difference — the attitude of the human soldier was considered more morally right than the same attitude manifested by the military robot. The average score in the robot soldier story was lower than in the human soldier story. The distributions of the scores were non-normal (the Shapiro-Wilk results were the following, robot soldier: $W=0.87$ $p<0.001$; human soldier: $W=0.85$, $p<0.001$). Wilcoxon signed-rank test showed that the difference in scores was statistically significant ($Z=-2.45$, $p=0.014$).

Is there a correlation between the moral beliefs (as measured by the MFQ) and the

ethical evaluation of robot and human attitudes (as measured by the ALAQ)?

The distribution of the care score was non-normal (S–W results: W=0.92, p=0.006). All other MFQ variables were normally distributed – the S-W results are: W=0.96, p=0.227 for *the fairness score*; W=0.97, p=0.43 for *the loyalty score*; W=0.95, p=0.098 for *the authority score*; W=0.96, p=0.149 for *the purity score*; and W=0.99, p=0.896 for *the progressivism score*. Two significant relationships between the MFQ variables and the ALAQ conditions were revealed. Firstly, the loyalty score was positively correlated with the sum of points in the robot condition (r=0.38, p=0.015). Ergo, the higher the score the respondents obtained in the loyalty foundation, the more they believed Asimov's Laws (applied to robots) to be correct. Secondly, the progressivism score was negatively correlated with the sum of points in the Robot condition (r=–0.32, p=0.044). Therefore, the more progressive the respondents were, the less they considered Asimov's Laws (applied to robots) to be correct. The assumed hypothesis was partially confirmed.

4. What Kind of Moral Attitudes Do We Expect from Robots?

4.1 Studies summary

Table 3 presents a summary of the selected studies, the one presented in Section 2 (see also Laakasuo *et al*. 2019; Malle *et al*. 2015). All three studies subject to this analysis employed stories of robots acting in hypothetical scenarios. The task of the respondents was to assess the rightness or moral acceptance of the robots and humans. Additional measures included: the deserved blame (Malle *et al*. 2015); the moral responsibility, and the trust of the agent presented in the story (Laakasuo *et al*. 2019). Malle *et al*. (2015) and the 6th experiment of Laakasuo *et al*. (2019) also incorporated a request for justification of the answers given.

The robots presented in the studies (as well as the circumstances in which they operated) were hypothetical. Although they differed in terms of their roles or occupations, all of them were social robots. Researchers employed different strategies in order to achieve the same goal, i.e., to make the participants imagine the main character of the story as a social robot with specific skills. In the study presented in section 2 a humanoid and intelligent robot: carer, surgeon and soldier (military robot) were described. The robot's mental capabilities were not specified in the other two studies. Malle *et al*. (2015) presented an "advanced state-of-the-art" repair robot working for the railways. The robot introduced by Laakasuo *et al*. (2019) was an advanced nursing robot. Such a description of robots was used to assign one more important feature to them. As respondents' evaluation concerned the moral attitudes of social robots, in order to ascribe moral rights to them, one must also assume their full autonomy. Autonomy, on the other hand, is a component of a moral agency. As Sullins (2006) points out, to be considered a moral agent a robot does not necessarily have to have a personhood; however, one of the requirements for being

perceivedas a moral agent is to be autonomous. According to Sullins, the other two are: the possibility of attributing intentionality to one's actions and possessing a responsibility to some other moral agents.

Study	Number of participants	Participant Characteristics	**Context**
Section 2	40	28 females; students; Age M = 21; SD = 1.48, Range = 17–24; recruited online	Whether and to what extent people not professionally involved in robotics consider obeying Asimov's Three Laws of Robotics (applied both to robots and humans) in real-life situations to be right. The Moral Foundations Questionnaire examine whether respondents' personal moral beliefs are related to an ethical evaluation of the attitudes of other people and robots.
Malle *et al.* (2015)	Study 1: 157 Study 2: 159	Study 1: 66 females, 90 males, 1 unreported; Age M=34.0; SD=11.4; recruited from Amazon's Mechanical Turk (AMT); completed an online experiment and were compensated	Experimental comparison of people's moral judgments (of permissibility, wrongness, and blame) about human and robot agents placed in an identical moral dilemma. Manipulation of the variable Agent Type (human versus robot) and Action (to direct versus not direct the train toward the single miner) both between and within-subjects.
		Study 2: 90 females, 68 males, 1 unreported; Age M=34.4; SD=11.5; recruited from AMT; online	Moral dilemma: variant of the trolley dilemma

Laakasuo *et al*. (2019)	Total:	Study 1:	Examined how people feel about
	1569	56 females;	forceful medication carried out either
		Age M=37.10; SD=17.65;	by human or robot nurses.
	Study 1:	Range = 18–80;	Hypothetical situations in which a
	135	recruited from a	human or an advanced robot nurse is
	Study 2:	large public library	ordered to forcefully medicate an
	403	in the City Centre of	unwilling patient.
	Study 3:	Helsinki	
	268		Measured moral acceptance,
	Study 4:	Study 2:	perceived trust, and allocation of
	26	315 females;	responsibility relating to the nurse's
	Study 5:	Age	decision of either following orders to

	500 Study 6 (a qualitative anthropological field study): 30	M=26.41; SD=6.67; Range = 18–63; recruited via email invitations sent to universities in Finland Study 3: 150 females; Age M=32.48; SD=13.36; Range = 18–76 Study 4 149 females; Age M=30.15; SD=9.94; Range = 18–66 Study 5: 230 females; Age M=29.3; SD=10.63; Range=18–82; recruited from Prolific Academic online survey site Study 6: 18 females; Age M=80; Range = 69–97; conducted between October 2017 and June 2018 in nine elderly residential homes in Finland	forcefully medicate the patient or disregard orders to protect the patient's autonomy. Manipulated the reputation of a nurse or a nursing robot; the consequences of forcefully medicating or not doing so; the status of the supervising party (who gives the order to forcefully medicate a patient).

Table 3: Basic information on the selected studies.

In each of the selected studies, respondents were asked to make a third-person moral judgement on the attitude of the agents described in the moral dilemmas (stories). The stories were always presented in two conditions: with a robot as an agent (the main subject of the present analysis) and with a human as an agent (enabling a comparison to be made). In order to examine whether people attribute the same moral norms to robots

and humans, and to test the hypotheses, the analysis was divided into two parts. Section 4.2 examines the main differences in expectations towards the robot's and human's attitudes, while section 4.3 compares these expectations.

4.2 Differences in expectations towards robot and human attitudes

Malle *et al*. (2015) used a variant of the popular trolley dilemma with a repairman or repair robot inspecting the rail system and making a decision: either to direct the train toward the single miner and thus killing one person (the action condition) or not to direct the train and consequently kill five people (the inaction condition). The results indicated that the robot action was considered more morally permissible than the human action, but only when the story with the human preceded the one with the robot. Similarly, a robot act of sacrificing one person was considered less morally wrong than the same human act. The robots and the humans also differed in blame received for action and refraining in that the robots were blamed similarly in both conditions, whereas humans were blamed more for action.

A short story where a human or a robot nurse is ordered to give a patient a medication against the patient's will was introduced in Laakasuo *et al*. (2019). The moral acceptance of forcefully medicating the patient was lower if done by the robot nurse compared to a human nurse. The human nurses were generally considered more trustful but also more personally responsible for their decision. The reputation of the nursing agent influenced the moral judgements of the human nurse more than the robot nurse. Results of the additional qualitative study showed that the subjects considered nursing robots to be cold and un-empathetic, due to their inability to explain what is happening.

The study presented in Section 2 aimed to examine the extent to which people who do not work in robotics, bioethics or roboethics consider Asimov's Laws to be right – both applied to a robot and a human – and whether there are differences in the declared rightness of an agent's attitude in both conditions. The only significant difference in the respondents' judgements on the rightness of the agent's attitude appeared in the story referring to the Third Law. The protection of the human soldier's life was considered more morally right than the protection of the military robot's being.

4.3 Similarities in expectations towards robots' and humans' attitudes

Although according to Malle *et al*. (2015) the robot action (directing the train toward a single miner) is considered more morally permissible than the same action taken by a human, this effect was present only when the story with the human preceded the one with the robot. When the story with a robot was introduced first, no significant difference was found. It is apparent that the participants' judgments about humans and the judgements about the robots influenced one another (the context effect). With regard to moral blame, in spite of the differences in action/inaction conditions, the overall amount of blame received by both human and robot agents was equal. According to Laakasuo *et al*. (2019),

the robot nurse's decisions were less acceptable than the human nurse's decisions, only in the forceful medication condition. The decision to disobey orders, and therefore respect the patient's autonomy, was considered more approvable than forcefully medicating in both the robot and human nurse conditions. One of the studies by Laakasuo *et al*. aimed to evaluate the influence of the consequences of forcefully medicating a patient (expressed as either the death of the patient the following day or the absence of changes in the patient's condition). The death of the patient resulted in much stricter moral judgments of the decision itself, both for the robot and the human nurse. The death of the patient also yielded equal trust results for both agents. The status of the supervising doctor manipulation (either a human doctor or an advanced AI) led to the observation that both the human and the robot nurses' disobedient decision towards the advanced AI doctor was strongly approved. In the qualitative study, Laakasuo *et al*. (2019) demonstrated that the prospect of losing autonomy has had such a strong impact on the participants (the residents of the elderly residential homes) that whether the agent who forcefully administered the medication was a human or a robot was often ignored.

In the present study, the declared rightness of the Asimov's Three Laws of Robotics did not differ significantly for the robots and humans. Asimov's laws seem to be considered moderately right, both with regard to robots (for which these laws were invented) and to humans. Furthermore, the analysis of the answers to the individual questions showed that there are no differences in two out of the three questions: concerning the protection of human life (with the robot/human carer) and the obedience to the humans (with the robot/human surgeon). This effect can be explained by the respondents' justifications for the answers given, in that they often indicated they did not see any reasons why the behaviour of the robot and the human should be assessed differently. The aggregated results were lowered by answers to reversed scale questions, i.e., those that presented attitudes at variance with the Three Laws of Robotics. In these questions, the robot/human carer hit a group of people to protect the medication carried for the person being cared for. This could be an indication of the fact that the First Law is perceived as not suitable for use in the real world, for it creates a harmful situation, either for a robot or a human, in which they cannot defend themselves. The second purpose of this study was to verify whether the subjects' personal moral beliefs, as measured by the MFQ (MoralFoundations.org 2016), are related to an ethical evaluation of the attitudes of other people and robots. The progressivism score correlated with the robot condition of the ALAQ, showing that the more progressive the respondents were, the less they considered Asimov's Laws to be right when applied to robots. This negative correlation can be explained in the respondents' justifications: the participants stated that due to their shared characteristics, the robots and the humans should be treated similarly. In contrast, the Three Laws allow treating robots quite objectively, prioritizing the good of humans and neglecting the protection of the robots' existence.

5. Conclusion

The main objective of this paper was to examine people's expectations towards the moral attitudes of social robots. The conclusions are based on the results of three empirical studies in which the stories of robots (and humans) acting in hypothetical scenarios were employed and the moral acceptance of their attitudes was assessed. The similarity in the evaluation of the humans' and robots' morality manifested in respondents' expectations that the robots would act in accordance with certain moral attitudes, just as they would expect from the humans. The question, therefore, was what kind of attitudes do we expect from robots and whether these attitudes should also be identical for both agents?

Each study shows some differences in the moral attitudes expected from the robots and the humans. Malle *et al*. (2015) demonstrated that robots are more strongly expected to make the utilitarian choices (sacrificing one person in order to save four). Laakasuo *et al*. (2019) found expectations of both robot and human's attitudes were strongly related to respect for the patient's autonomy in that the robot nurse's decisions were less acceptable only in the forceful medication condition. Finally, it seems that in certain circumstances the protection of a human's life is considered more morally right than the protection of the robot's being. Differences that emerged in Malle *et al*. (2015) can be partially explained by the context effect: the order in which the stories with the robot and the human agent appeared had an impact on the respondents' judgements. It is possible that in the current study, where stories with the robot agents always preceded those with the human agents, a similar effect occurred. Perhaps presenting the stories in a different order would reveal more differences in the assessment of the human's and the robot's attitudes. This would mean that when people evaluate the attitudes of a robot first, their evaluation is mainly based on their opinion about people and only when they are told to evaluate the attitudes of a human first, differences in their assessments of the two agents appear. Future studies could investigate whether this phenomenon actually occurs.

There were, however, a number of similarities in the assessment of robots' and humans' attitudes. The overall amount of blame received by both a human and a robot agent was similar, which contributes to the claim that the moral decision-making capacity makes the robots natural targets for moral blame (Malle *et al*. 2015). An additional study (Voiklis, Kim, Cusimano, & Malle 2016) analysed the justifications for moral judgements provided by the respondents in Malle *et al*. (2015). It was demonstrated that even if sometimes different moral attitudes were expected from the humans and the robots, participants often provided similar types of justifications for their moral judgments. This suggests that people extend their moral reasoning (or moral intuition) to robots, regardless of the norms applied. In Laakasuo *et al*. (2019), the strong impact of the prospect of losing autonomy resulted in no difference in the evaluation of the robots' and humans' attitudes. Moreover, the decision to respect the patient's autonomy was considered more approvable than forcefully medicating, regardless of the agent.

Laakasuo *et al*. (2019) demonstrated that in certain circumstances people make similar consequentialist moral judgements when evaluating both the human and the robot decisions. However, Malle *et al*. (2015) suggest that in some extreme cases consequentialist moral judgements are made differently depending on the agent being evaluated. According to the current findings, apart from the First Law, Asimov's laws were considered moderately right, both with regard to the robots and to the humans. The incongruity of the First Law stems from harmful situation in which the agent cannot defend itself. Therefore, according to the results, the robots should protect their existence and obey people, but in some situations, they should be able to hurt a human (in self-defence, defence of other people, or other values). Consistent with the above is the result of the MFQ, suggesting that the more progressive the respondents were, the less they thought Asimov's Laws should apply to robots. As the participants' justifications indicate, The Three Laws allow robots to be treated objectively, while the respondents expected them to be treated similarly to human beings. The aforementioned findings could make an important contribution to the discussion of whether robots should have the status of moral patients and moral agents (e.g., Sullins 2006; Hoffmann & Hahn 2019). They are also consistent with the criticism of Asimov's laws in this context (see Anderson 2008).

The fact that Malle *et al*. (2015) and Laakasuo *et al*. (2019) reported more differences in the evaluation of the robots' and the humans' attitudes than the present study may be explained by the character of the experimental task used in these studies. It seems that such an extreme task as the trolley dilemma or the scenario in which the patient is deprived of their autonomy triggers some differences in the moral judgements.

These results could be of considerable use both in implementing morality into robots and in the legal evaluation of their attitudes and behaviour. Malle and Thapa (2017) revealed that the desire for Social-Moral Skills in robots increased over the years 2013-2016. The present work answers the question of which moral skills people expect. An awareness of the strong influence of the prospect of losing autonomy and the need for explanatory skills as well as empathy will improve the designs of nursing robots. In their detailed overview of AI ethical guidelines, Hagendorff (2020) states that most of the guidelines omit contexts of care, nurture, help, welfare, social responsibility, or ecological networks, and so they lack an interpretation of moral problems within a wider framework of "empathic" and "emotion-oriented" ethics of care. As the current findings have shown, this context of understanding the morality of robots is of huge importance to humans. Therefore, taking into account people's expectations can create better AI guidelines.

The fact that the robots are required to make utilitarian choices may prove potentially useful in the context of choices made by autonomous cars, highlighted in the introduction of the present paper. Regarding military robots, people consider their existence to be less valuable than a human soldier's life and believe that robots can be sacrificed in the name of other values. Also, the context effect described in the results of Malle *et al*. (2015) may occur in real life, for example when a legislative body evaluates the behaviour or rights of

a robot by comparing it with those of humans.

Just as Hoffmann and Hahn (2019) recommended people are familiarised with how AI algorithms work, it is also important to take into account people's opinion on what they expect from robots' moral (and any other) attitudes. As noted in Ljungblad *et al*. (2011), robots' ethical concerns should be grounded in the empirical data and not limited to the philosophical considerations. Although the present paper fulfils this purpose, another critical issue highlighted by Ljungblad *et al*. (2011) is that these studies should not be based on futuristic scenarios and robots that do not exist yet. Nevertheless, all the robots and the situations described in this review were hypothetical. Therefore, in order to reveal the ethical implications that may be missed while using speculative scenarios, future work should concentrate on "the actual use of existing robots in a real environment" (Ljungblad *et al*. 2011, 191).

An undoubted weakness of the presented studies is the relatively small number of respondents. The project designed by scientists from MIT Media Lab[1] may be the answer to this problem and thereby constitutes the future of research on ethical issues related to AI. The project aims to collect people's insights into the ethics of robots through crowdsourcing and simple games. The authors state that "The Moral Machine" attracted worldwide attention, and allowed them to collect 39.61 million decisions in 233 countries, dependencies, or territories (Awad *et al*. 2018, 60). Thanks to this method we can examine what decisions people think robots should make when faced with moral dilemmas.

References

Anderson S. 2008. "Asimov's Three Laws of Robotics and Machine Metaethics," *AI & Society* 22:477–493.

Asimov I. 1981. "The Three Laws," *Compute!* 11(18):18.

Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., & Rahwan I. 2018. "The Moral Machine Experiment," *Nature* 563(7729):59–64.

Fong T., Nourbakhsh I., & Dautenhahn K. 2003. "A Survey of Socially Interactive Robots," *Robotics and Autonomous Systems* 42(3/4):143–166.

Giger J.-C., Moura D., Almeida N., & Piçarra N. 2017. "Attitudes Towards Social Robots: The Role of Gender, Belief in Human Nature Uniqueness, Religiousness and Interest in Science Fiction," in S. N. de Jesus & P. Pinto (Eds), *Proceedings of II International Congress on Interdisciplinarity in Social and Human Sciences,* Vol. 11 (p. 509). Research Centre for Spatial and Organizational Dynamics, University of Algarve Faro, Portugal.

1 https://www.media.mit.edu/projects/moral-machine/overview/

Hagendorff T. 2020. "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines* 30:99–120.

Hoffmann C. H. & Hahn B. 2019. "Decentered Ethics in the Machine Era and Guidance for AI Regulation," *AI & Society* 35:1–10.

Jarmakowski-Kostrzanowski T. & Jarmakowska-Kostrzanowska L. 2016. "Polska adaptacja kwestionariusza kodów moralnych (MFQ-PL)," *Psychologia Społeczna* 11:489–508.

Laakasuo M., Kunnari A., Palomäki J., Rauhala S., Koverola M., Lehtonen N., Halonen J., Repo M., Visala A., & Drosinou M. 2019. "Moral Psychology of Nursing Robots – Humans Dislike Violations of Patient Autonomy But Like Robots Disobeying Orders." URL: https://psyarxiv.com (retrieved on July 18, 2020).

Ljungblad S., Nylander S., & Nørgaard M. 2011 (March). "Beyond Speculative Ethics in HRI? Ethical Considerations and the Relation to Empirical Data," *Proceedings of the 6th International Conference on Human–Robot Interaction (HRI)* (pp. 191–192). IEEE.

Malle B. F., Scheutz M., Arnold T., Voiklis J., & Cusimano C. 2015 (March). "Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents," *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 117–124). IEEE.

Malle B. F. & Thapa Magar S. 2017 (March). "What Kind of Mind Do I Want in My Robot? Developing a Measure of Desired Mental Capacities in Social Robots," *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human–Robot Interaction (HRI)* (pp. 195–196). IEEE.

MoralFoundations.org 2016. "Moral Foundations Theory." URL: https://www.moralfoundations.org/ (retrieved on January 5, 2019).

Murphy R. & Woods D. D. 2009. "Beyond Asimov: The Three Laws of Responsible Robotics," *IEEE Intelligent Systems* 24(4):14–20.

Sullins J. P. 2006. "When Is a Robot a Moral Agent," *Machine Ethics* 6:23–30.

Voiklis J., Kim B., Cusimano C., & Malle B. F. 2016 (August). "Moral Judgments of Human vs Robot Agents," *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 775–780). IEEE.