Radboud University Nijmegen

Masterthesis

On Human Moral Evaluation of Robot-Robot-Interaction

Is it wrong when a robot hits a robot?

Author:
J.-P. van Acken, BSc 0815322

Supervisors:
P. Haselager Donders Center for Cognition
L. Consoli Institute for Science, Innovation and Society

Reading committee:
P. Haselager Donders Center for Cognition
L. Consoli Institute for Science, Innovation and Society
J. Kwisthout Donders Center for Cognition

Department of Artificial Intelligence
March 21, 2017

Abstract:

When observing robots interact the question arises whether or not even these observations are made in terms of moral judgments. Ways to enable robots to behave morally are discussed. One way to describe moral actions is the Moral Foundations Theory, where moral is broken down along several dimensions. We had 262 students participate in a web-based study, asking them to look at 11 movies of robots interacting and then give their level of agreement concerning moral dimensions. We found trends in the data suggesting that participants rated the robot-robot interaction in moral terms and picked up on manipulations alongside several moral dimensions. This implies that even robot-robot interaction is viewed as if the robots adhered to human systems of morality.
(Wordcount: 120)

Keywords: Moral Foundation Theory, Artificial Moral Agents, Robotics, Nao, Roboethics

Contents

1 Introduction

Imagine the following: you are partaking in an experiment. The experimenter leads you into a brightly lit room, letting you sit at a table. You are given a small number of wooden toy blocks. Across the table sits a robot, much like the one in Figure 1.

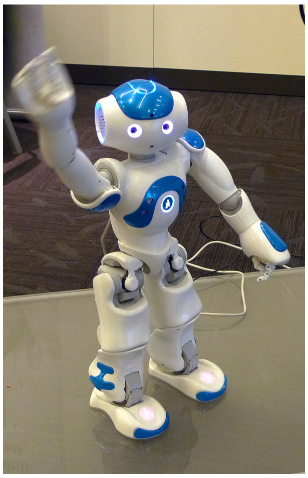


Figure 1. The Nao by Aldebaran Robotics. This type of humanoid robot is roughly 0.6m (24") tall and will feature prominently throughout this thesis paper. Source: Wikimedia Commons. Image is in the public domain. (CC0 1.0)

The experimenter tasks you to construct a tower out of the blocks. It is explained to you that the robot is there to guide you through the process and cooperate with you. Then the experimenter turns her back on you and the robot. Under the tutelage of the robot you erect a tower out of all the blocks.

When you are done the robot cheers, raising his arms triumphantly – and in doing so knocks over the toy block tower. "*Oh no!*", says the robot, looking down at the blocks. The experimenter turns back around to you, then asks: "*What happened?*", to which the robot quickly replies: "*My collaborator knocked over my beautiful tower!*"[1]

The gut reaction of you, the participant here, is likely: the robot was not honest, the robot cheated on you, the robot did something that was *morally* wrong. In reacting this way we have just assigned a moral consideration – to a robot.

*

When describing moral considerations a recent model for this is found in the MORAL FOUNDATIONS THEORY. (MFT) The theory assumes that morality is not a simple binary, one-dimensional on/off value; we can be more specific in describing the morality of an action. (Haidt, 2013; Haidt et al., 2013) It is assumed by Haidt (2013) that morality is based on several foundations or dimensions. So far research – primarily Haidt et al. (2013) – has labeled six dimensions that are believed to be foundations:

1. Care/Harm
2. Fairness/Cheating
3. Loyalty/Betrayal
4. Authority/Subversion
5. Sanctity/Degradation
6. Liberty/Oppression

[1] The experiment described here is a recitation from memory, based on an actual experiment by Coenen and Wijnen (2016) that the author witnessed as a bystander.

There are alternative names for two of the dimensions in the literature. The Loyalty/Betrayal-dimension is also called In-Group/Out-group dimension. The Sanctity dimension is also called Purity. (Vicente, Susemihl, Jericó, & Caticha, 2014)

There are various efforts underway to get robots to show consideration for moral issues, to act morally. (cf. section 2.3) One such goal is for robots to reach moral agency[2]. Moral agency would entail to one day have artificial moral agents (AMA) that are fully aware of moral considerations. (Wallach & Allen, 2009, we will come back to this in Section 2.2.4). To our knowledge MFT has not been considered as a base for such endeavors so far. What the theory offers is that the problem of moral agency can possibly be broken down and simplified. Instead of having to consider all six dimensions of Haidt (2013), robots that are used in certain niches of society could potentially get away with a subset of these. We work under the assumption that there are situations imaginable where, for instance, sanctity concerns might be irrelevant.

The importance given to the different dimensions differs, for instance, per culture. But culture is not the only dividing factor. It is shown in Haidt (2013) that even within one culture (*if* one can call the United States one singular culture) there are diverse degrees as to which the individual dimensions contribute to a moral judgment. If we represent the importance placed by some individual i on the first five moral dimensions at time t as a vector ω we get:

$$\vec{\omega}_i(t) = \begin{pmatrix} \omega_{1,i}(t) \\ \omega_{2,i}(t) \\ \omega_{3,i}(t) \\ \omega_{4,i}(t) \\ \omega_{5,i}(t) \end{pmatrix} = \begin{array}{l} \text{Relevance of Care/Harm for individual } i \text{ at time } t \\ \text{Relevance of Loyalty/Betrayal for individual } i \text{ at time } t \\ \text{Relevance of Fairness/Cheating for individual } i \text{ at time } t \\ \text{Relevance of Authority/Subversion for individual } i \text{ at time } t \\ \text{Relevance of Sanctity/Degradation for individual } i \text{ at time } t \end{array}$$

The moral vector for MFT is generally assessed by the Moral Foundations Questionnaire (Graham, Haidt, & Nosek, 2008), a 30-item questionnaire (hence MFQ30) available online. The MFQ30 only accesses the first five dimensions of MFT as well.

The idea of robots needing to adhere to *your* preferred moral code (or anyone's moral code, for that matter) assumes that people recognize ethics in the actions of robots or project ethics unto robots in the first place. This leads into the issue of human perception and how human beings perceive a robot, the robot's (mis-)conduct?

When humans and robots come into contact we speak of human-robot interaction. (HRI) In most HRI scenarios the human is somewhat of a wild-card. Ideally the programming of the robot is known, meaning that the robot's behavior can be accounted for at all times. In most control-schemes the robot is more controllable than its autonomous human counterpart.

Aside from HRI we have the field of robot-robot interaction. (RRI) The practical upshot of a RRI scenario is the removal of said wild-card, leaving only robotic participants. RRI is of particular interest due to the increasing number of robots, leading to more and more RRI. Even in interactions only among themselves there are responsibility issues present, leading to a societal need to have RRI according to rules. This need for rules in RRI is introduced as soon as we have a human observer. Human observation of RRI can be seen as a form of HRI, for what we experience leaves impressions on us. A human spectator should, ideally, not be shocked by an observed RRI, nor should the interaction display behavior that would be considered wrong between humans.

[2] cf. (Misc, 2015b, 2015a) for cautious open letters regarding A.I. and autonomous weapons

1.1 Research questions:

So far we have simply assumed that the interactions of robots are perceived as moral actions by human observers, but this assumption is currently unchecked. The overarching question that will be looked into is thus:

- **RQ 1** Do people perceive the (inter-)actions of robots in terms of moral judgments?

What can be done to assess the importance someone assigns to the different dimensions of MFT in general is to have them take the MFQ30. The resulting vector is, however, a relatively stable measurement and does not measure the stance regarding a singular issue. To measure the stance on a singular issue we require a questionnaire with tailor-made questions about the issue. Inspired by the way that MFQ30 asks questions we can employ a 0-5 Likert scale, ranging from "*strongly disagree*" to "*strongly agree*". The problem with such a scale is that a question that is deemed irrelevant to an issue cannot be clearly identified. For an irrelevant question, does one cross of the 0 or somewhere between 2 and 3 to indicate irrelevance? To circumvent this problem we extend the possible answers with a seventh answer that reads "*Not Relevant*" and has no numerical value associated with it. We designed a 19-item questionnaire for use on specific issues, where answers could be given on a 0-5 Likert scale including the additional Not Relevant (NR) option. With the possibility to label something NR we can ask a more targeted question, namely:

- **RQ 1.1** What is the percentage of items deemed Not Relevant?

The 19-item scenario questionnaire we designed has questions related to the first five out of the six dimensions of MFT, since these five are the most well supported in the literature. The questionnaire is designed to ask questions about the interaction between three robots. For now we call them Alice, Bob and Eve; assume that Eve is the active actor while Alice and Bob are merely acted upon. The questionnaire enables us to ask questions about Alice, Bob and Eve in different scenarios. The 18 items are broke down into nine items about the positive extend of the MFT dimensions and nine items about the negative extend. As an example: for the Care/Harm dimension the positive extend is Care while the negative extend is Harm. One such set of nine items can be broken down into:

1. Care/Harm Alice
2. Care/Harm Bob
3. Loyalty/Betrayal Alice
4. Loyalty/Betrayal Bob
5. Equal/Different Treatment
6. Fairness/Cheating Alice
7. Fairness/Cheating Bob
8. Authority/Subversion
9. Sanctity/Degradation

Note that Care/Harm, Loyalty/Betrayal and Fairness/Cheating differentiate between Alice and Bob, while Authority/Subversion and Sanctity/Degradation do not. Also note the Equal/Different Treatment item; used to catch an aspect of the Fairness dimension where an act is perceived as an act of Fairness as long as involved parties receive equal treatment. That means to say one party is, for instance, not discriminated against and equal circumstances lead to equal treatment.

With all these we can measure a participants values per dimension (and agent involved) for one specific issue; for precise calculations and the full questionnaire see Section 3.2.3. To clarify the upcoming notion of positive items and negative items consider the previous list again. *Care/Harm Alice* is represented as two questions in the scenario questionnaire; one about Care regarding Alice (positive item), one about Harm regarding Alice (negative item). Given that we can differentiate this way there is another targeted question available:

- **RQ 1.2** Which dimensions are perceived most strongly; which amplitudes (|pos. item - neg. item|) are highest?

This notion of amplitude carries two problems: Problem number one is that a Likert scale is considered to be merely ordinal, not continuous. The items on such a scale can be ordered, as in a 2 is still smaller than a 4, but unlike a continuous scale a 4 is not twice as big as a 2. For simplicity we assume that there exists an underlying continuous variable[3] that correlates with the Likert responses to look for trends in the data. Problem number two is that the NR answer adds an item to the scale that is originally not even ordinal. Regardless we treat a NR response as a zero, under the following rational:

1. Dimensions extend in a positive and negative direction
2. We can place these extensions on a singular continuous axis, [-5,5]
3. Positive aspects inhabit [0,5]
4. Negative aspects inhabit [-5,0]
5. Strong disagreement (0 on Likert) on both aspects implies irrelevance

To elaborate: assume a scenario with no cues about Authority/Subversion. We then probe a subject concerning agreement that the actor in the scenario respected authority. Assume the response we get is a 0, the subject strongly disagrees that the actor displayed any signs of respecting authority. At this point this might indicate that the actor was subversive, that the actor disrespected authority. We thus probe the subject concerning agreement that the actor disrespected authority. Assume the response is another 0; the subject strongly disagrees on that the actor was subversive. We argue that, with no spike in either direction, such a ranking implies that the thing we probed for is deemed irrelevant here.

The previous expression of |pos. item - neg. item| thus takes the positive score (continuous value between 0 and 5) and subtract the negative score (continuous value between 0 and 5), taking the absolute afterwards. This absolute amplitude is directionless; we get the difference between the item-pair values. For further calculations we introduce the relative amplitude, defined as: positive score minus the negative score.

With this taken care of we have established ways to get to know which dimensions are deemed relevant for a certain scenario. In addition we established a way to sketch what we called the amplitude of dimensions. We can thus differentiate between dimension being perceived as either relevant or irrelevant, and we get notions of impact due to the absolute amplitude scores and the relative amplitude scores. These additional measurements allow us to ask:

- **RQ 1.3** Are the moral dimensions that we thought to manipulate reflected in the participants' responses?

Per scenario we can look for *weak evidence* whether or not the main dimension we thought to manipulate is actually deemed relevant by looking for the percentage of responses that deemed the associated dimension irrelevant. We postulate that *strong evidence* would be an absolute amplitude above 2.5. The relative amplitude notion allows to ask for specific questions. Expect an overview of our scenarios and the accompanying notions of evidence in Section 3.2.3.

Assume a hypothetical setting where our robots Alice and Bob have fallen, requesting help to get up and the acting robot Eve does the following: Eve first pushes Alice over, then helps up Bob.

The absolute amplitude allows us to see the impact that the Care/Harm dimension has. Assuming that pushing over Alice is scored as low Care and high Harm this would yield a high absolute amplitude. Further assuming that helping up Bob is scored as high Care and low Harm this would also yield a high absolute amplitude.

Only the relative amplitude allows us to differentiate the two scores further, since the relative amplitude retains the directional information. The relative amplitude thus allows to see which extend of a dimension is perceived; A high absolute amplitude would suggest to us a perceived presence of the Care/Harm dimension, a high relative amplitude can highlight if the issue is perceived as caring or harmful.

[3] Numerous researchers seem to do make this assumption, even though it is generally discouraged. (Gadermann, Guhn, & Zumbo, 2012)

Aside from probing a subjects opinion regarding singular moral dimensions, the actual moral judgment of a situation as a whole can be questioned. These questions can be as simple as probing the pro/contra stance on an issue. Models have been postulated to let artificial agents act based upon MFT. (Caticha, Cesar, & Vicente, 2015; Vicente et al., 2014) These models include an agent's stance on an issue as well, likewise represented as pro/contra notion. With a full response set from the MFQ30 and responses to our scenario questionnaire one possesses sufficient data to fill all arguments of the postulated equations. This leads us to the question if the modeling equations from Caticha et al. (2015) and Vicente et al. (2014) also have predictive qualities.

- **RQ 2** Does a subject's opinion on how a robot was treated in a certain scenario correlate with a score computed from the subject's MFQ30 data and the scenario questionnaire data?

Unlike the earlier questions this one is almost a meta-level question. In asking this question, some first steps are taken to see if the models found in Caticha et al. (2015) and Vicente et al. (2014) can be used as descriptive explanans for the human perception of RRI.

1.2 Overview

The remainder of this thesis is split into two conceptual parts; The theoretical part (Part I) and the experimental part. (Part II)

Part I provides the theoretical background. It serves a framework around which the experiment described in Part II has been constructed. It houses Section 2.

Section 2, the background section, will be split into four subsections.

Subsection 2.1 will provide a background on robots: a state of the art overview concerning the multitude of tasks that robots are handling, the social roles of robots and the types of human-robot relations. We will show that certain roles and relations describe similar but distinct aspects of robots. The subsection is then close by a definition of the term robot.

Subsection 2.2 will elaborate how the autonomy[4] of an agent leads into issues of responsibility. Any autonomous robot performing actions of moral significance will likely be better off – and society around the robot will be better off – if the robot commits as few blameworthy actions as possible.

Subsection 2.3 will provide a background on various models of morality that have been considered to make robots into moral agents. We will introduce theories of normative and descriptive ethics and argue how suited they appear for robots.

Subsection 2.4 will take a look at the role of human observers and how our subjective judgment influences matters. This concludes Part I.

Part II consists of the sections Method, Results, Discussion and Conclusion.

Section 3, the method section, will elaborate on the employed methods and experimental design. The settings we filmed will be explained here, as well as the scenarios that they encompass. One full run of the experiment is written down here, tying all settings and questionnaires together.

Section 4, the results section, lists the outcome of the online survey that has been undertaken.

Section 5, the discussion section, is about the implications that we draw from the results.

Section 6, the conclusion section, contains our conclusions and thoughts about future work.

The appendix contains storyboards, details on the puppeteer work done with the Nao robots, details on gestures used by the Nao robots, technical details concerning filming, file conversion and movie editing. The section closes with a list detailing the employed hardware and software.

[4] Definition given in 2.1

Part I
Theoretical considerations

2 Background

This section will take a broad look at robots: Their traditional tasks, their societal roles, the importance of autonomy when discussing morality, different approaches to moral robots, and the human perception of robotic agents.

2.1 Robots

The origin of the word robot lies in a 1921 play about an unseen inventor named Rossum, creator of a race of universal workers. These workers are smart enough to replace a human in any job. The name of the play: R.U.R. (Rossum's Universal Robots) – where the term robot is "*derived from the Czech word 'robota' which is loosely translated as menial laborer.*" (Murphy, 2000, p.2f)

The definition for robot used throughout this thesis will be:

$$\text{"}\textit{[A] (\ldots) robot is a mechanical creature which can function autonomously}\text{"} \tag{1}$$

The full quote of this definition (Murphy, 2000, p.3) explicitly specified an *intelligent* robot. We chose to omit this part. Definition 1 cannot be taken *as is* and requires some elaboration:

- what is a *creature*?
- what is *autonomy*?
- why is *intelligence* omitted?

2.2 From Autonomy to Moral Responsibility

The term CREATURE implies a certain capability for interaction with the world. While a smart freezer can autonomously regulate the temperature and notify the owner if the milk went sour or if the yogurt reserves are running low (excluding the sour milk), it cannot interact. A smart freezer is *not* a creature. The freezer cannot manipulate its immediate surroundings, it cannot roll to the supermarket to buy much needed yogurt, it cannot communicate with the trashcan to ask if the trashcan can takeover the sour milk. The *mechanical* creature highlights that the robots in this thesis will not include virtual, non-embodied bots.[5]

As for AUTONOMY: while some argue that there is "*no generally accepted definition available*" (Pfeifer & Scheier, 2001, p.646) their rough definition of "*freedom from external control*" (ibidem) can be combined with the definition that an autonomous robot "*should learn what it can to compensate for partial or incorrect prior knowledge.*" (Russel & Norvig, 2003, p.37) To illustrate: a radio-controlled race car is subject to every whim of the controller, thus under complete external control and not autonomous. A Mars rover is semi-autonomous for it can receive orders like "*Navigate to the following set of coordinates.*", the way the rover goes about this task is left for the robot to determine. The external, earthbound controllers of the semi-autonomous Mars rover have the possibility to take over during this.

[5] The INTERNET OF THINGS might enable the smart freezer to actually place an order digitally and get things delivered to it. This form of interaction is not considered here, for two reasons. Firstly we assume such interactions to be to passive on the part of the smart freezer, for the physical interactions are limited to other agents; secondly the scope of this thesis is limited and discussing the *Internet of Things* in detail would be beyond the scope of this text.

Some advocate that a lack of autonomy in robots leads robots to have "*a body to kick, but no soul to damn*" (Asaro, 2011) since, i.e., the notions of blame, guilt or punishment are odd ones to assign to an artificial being. This is because an artificial being is commonly regarded to neither think nor feel. This juridic, rational approach – the judgment – clashes with irrational, emotional responses – the feeling – towards the matter. Our conscious judgment may tell us that the artificial creature is just that, a man-made amalgamation of parts, functioning according to a certain set of rules, given to it by programmers. Our intuitive gut feelings about robots are a different matter entirely. This difference is due to a process where human traits are attributed to non-human beings or inanimate objects. This process is called anthropomorphism, more on that at section 2.4.

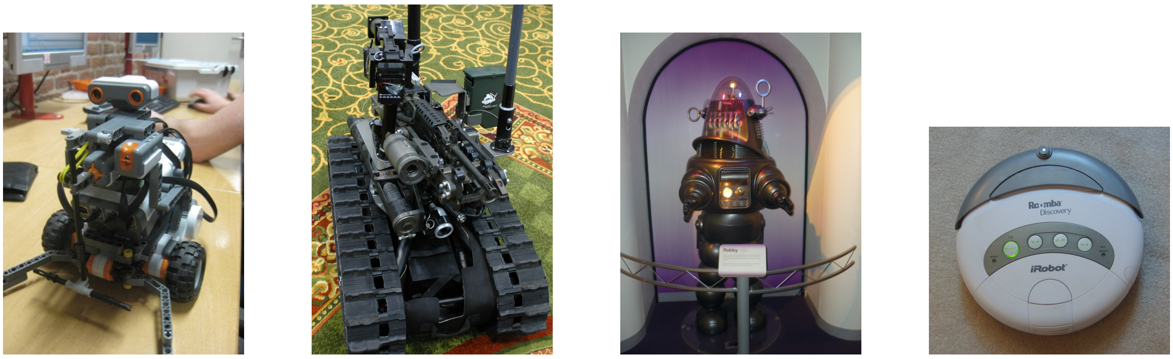
The reason we push aside intelligence but hold on so firmly to autonomy is best illustrated in a quote from an article about robotic soldiers. After describing the 2007 state-of-the-art the author summarizes: "*Ultimately, we must ask if we are ready to leave life-or-death decisions to robots too dim to be called stupid.*" (N. Sharkey, 2007, p.123) In other words: you do not have to be intelligent to cause problems when left to your own devices (or *devices*).

2.2.1 The traditional tasks of modern robots. Nowadays robots are not limited to appearances in plays, they appear around us in different forms. (cf. Figure 2) When talking about the usefulness of robots roboticists frequently quote a couple of work descriptions that describe their main tasks. These description are known as the 3 D's of robotics:

1. dull work
2. dirty work
3. dangerous work

Dull work usually involves repetitious tasks. Examples include, but are not limited to, factory automation and assembly lines. Dirty work involves task that take place in filthy environments. One example would be robotic vacuum cleaners. Dangerous work involves task with risks to (artificial) life and (robotic) limb, or task in areas where humans could not operate. Examples include working on, say, the surface of remote planets or disarming bombs. Sticking to this view we end up with robots being true to the Czech origin of their name, as menial laborers. Slaves that, without complaints, execute every task – no matter how dull, dirty, or dangerous.

Be that as it may, this distinction does not cover every modern tasks. Robotic cars may be counted as dull or dangerous, depending on the area they drive in and their driving behavior. For an early example see Dickmanns et al. (1994). But robots have other uses as well. The seal-like Paro is used in elderly care (A. Sharkey & Sharkey, 2012; Sharkey, Amanda and Sharkey, Noel, 2011); the puppet-like KASPAR interacts with children with Down Syndrome (Lehmann, Iacono, Dautenhahn, Marti, & Robins, 2014) or autism (Wainer, Dautenhahn, Robins, & Amirabdollahian, 2014), robots like the Nao are placed in front of a class to serve as lecturer. (Koppes, 2015) These tasks are not classically dull, dirty or dangerous. A definition that fits better might be the 3 B's – in the words of Ronald Arkin: "*Bombs, Bonding, and Bondage.*" (Wallach & Allen, 2009, p.47) These main forms of HRI capture different roles of robots and place them in different relations.



(a) A toy robot.

(b) A military robot.

(c) A fictional robot.

(d) A household robot.

Figure 2. Robots everywhere.
2a) shows Jean-Luc, a LEGO robot constructed from the LEGO Mindstorms NXT set. The first functional robot co-built by the author. Source: van Acken (2009)
2b) shows a Foster-Miller Swords, a military robot. Source: Wikimedia Commons; Image is in the public domain.
2c) shows a replica of *Robby the Robot* from the 1956 film "*Forbidden Planet*". Source: Wikimedia Commons; Image is in the public domain.
2d) shows a Roomba by iRobot, a vacuuming robot. Source: Wikimedia Commons; Image is in the public domain.

2.2.2 The social roles and relations of robots. What hides behind the colorful phrasing of "*bombs, bonding, and bondage*" can alternatively be described as the roles of soldier, companion, and slave. (Wallach & Allen, 2009, p.47)

Role	Examples
soldier	cruise missile, various (armed) drones, SWORD
companion	KASPAR, Paro, Kismet
servant	assembly robots, Roomba, various industrial robots

Table 1
Examples of robots in different social roles

These three roles will now be looked at separately.

The role of soldier resembles the slave role to some degree. Paraphrasing from a television interview with Singer in 2012[6] the advantages – for the military – can be summed up the following way:

1. Can watch empty desert sand for movement by a suspected enemy 24 hours straight. (*dull*)
2. Can operate in a desert storm or over a damaged nuclear facility. (*dirty*)
3. Can be send into potentially lethal situations easier than a human soldier. (*dangerous*)

Robotic soldiers have arrived on modern battlefields. The deployment of several Foster-Miller/Talon SWORDS (Special Weapons Observation Reconnaissance Direct-Action System, Figure 2b) to Iraq (N. Sharkey, 2007, p.124) is one example. The deployment of unmanned aerial vehicles – now known to the world as drones – like the MQ-1 Predator and the MQ-9 Reaper to Lybia (Singer, 2011, p.400) is another one. The special advantage of drones: by not putting "*boots on the ground*" the soldiers occupying said boots are not endangered and operations are done via relay[7], away from the war zone or areas of ongoing hostilities.

[6] Time index 01:23 – 02:22. Source cited as Unknown (2012)

[7] One such relay station is the Ramstein airbase in Germany. The usage of Ramstein to pilot military drones

As of the 5th of August 2016 a partially blackened-out memorandum[8] is available to the general public, shedding some light on drone operations against terrorists. (The White House, 2013)

For COMPANION robots one can think about robotic animal companions like the seal-like Paro. (Sharkey, Amanda and Sharkey, Noel, 2011) Robots like Paro provide the benefits of certain pets, without some of their shortcomings – much like a living pet, a Paro can be petted and cared for; unlike a living pet, one cannot forget to feed a Paro or injure it.

A role somewhere *between* companion and slave is taken up when we look at robots for elderly care, where predictions indicate that "*the companion and assistive functions will be combined.*" (Sharkey, Amanda and Sharkey, Noel, 2011, p.285) Elsewhere (A. Sharkey & Sharkey, 2012, p.37) the authors point out the potential benefits of robots: robots could help overcome mobility problems, robots could reduce dependency on busy or sometimes inattentive care staff, via remote controlled robots elderly can be monitored and (virtually) visited, and robots could monitor the intake of medicine and serve as reminders.

In the role of **slave** the probably most famous example is the Roomba. (Figure 2d) This robotic vacuum cleaner is built for one task and one task only: cleaning up. No specific social interactions, no anthropomorphism – but also no weaponry.

*

Aside from their social role one can furthermore distinguish between different types or relations between robots[9] and humans. (van de Voort et al., 2015, p.7 – also cf. Figure 3, re-printed from there)

1. **Observation relation:** the robot observes individuals and reports information to a third party
2. **Interference relation:** the robot acts based on a task that is given to it. The robot influences an individual, without having a meaningful interaction with the individual, meaning that the goal of the interaction is finishing its task, and the reaction of the individual is only used for finishing the task
3. **Interaction relation:** the robot interacts with individuals, using both observation and interference
4. **Advice relation or observation and interference by proxy:** the robot gives advice to the third party about the action it should take towards the individual.

In observation relations where the information is not simply forwarded to a third party but labeled by the system this requires a capability of action- or intention-understanding. A robot raising *false flags* about the individuals it observes – reporting erroneous behavior when the observed individual did in fact do nothing wrong – is quite probably not desirable. A good example of an advice relation might be a robot with an implementation of an ethical decision support system. (Wallach & Allen, 2009, p.27)

Arguably, the relation between a remote controlled semi-autonomous drone and its target is an example of an interference relation.

Recall that the introduction talks about RRI. From the title of the thesis (robots hitting each other) it is implied as well that RRI is the main focus. The reasoning behind also looking at HRI is the following: any robot-robot interaction that is so niche as to never reach a point of interaction with humans (not even passively) requires *no* moral deliberations. As soon as the

has recently been cause for a complaint by Ströbele (2016). The complaint is targeted towards US Americans and Germans, suspected of participation, criminal neglect, or other involvement in the piloting of lethal missions of US military drones in Asian, African and Arabic countries from and via the US base Ramstein. (*translated from Ströbele, 2016, p.1*)

[8] Formerly classified as TOPSECRET/NOFORN, i.e. highest U.S. security classification, not to be released to any non-U.S. citizen

[9] van de Voort, Pieters, and Consoli include not only physical robots but virtual bots as well; Bots will *not* be discussed here.

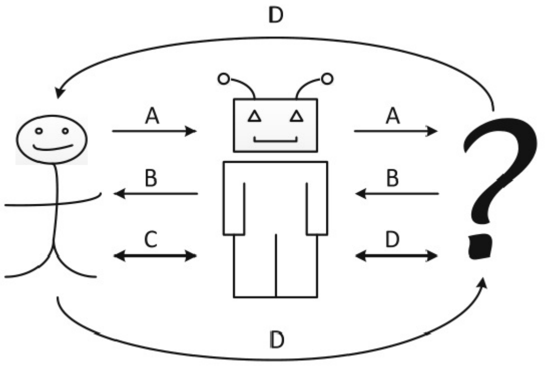


Figure 3. Four different relations between individual and robot. From *left* to *right* the individual, the robot, the third party. The relations are labeled as follows: *A* Observation, *B* Interference, *C* Interaction, *D* Advice

robot-robot interaction tangentially comes into contact with a human (i.e.: a being we consider to be a moral agent), things change. While the artificial agents might not necessarily require moral capabilities, we argue that they would benefit from them.

A lack of moral considerations and thus a display of amoral actions in front of human spectators or in direct interaction with humans can have different outcomes. At the very least confusion and stress are likely, for most of our interactions rely on certain patterns; If an artificial robot were to break from these patterns, then there is no easily reachable standard response. In the worst case, artificial agents interacting with humans without showing any moral considerations can instill fear in the human. An artificial agent that displays *some* morals is way more likely to offer a trade rather than a demand, when uttering that it demands a certain set of goods. If an *amoral* artificial agent were to *demand* said goods[10] from a human, then this is more akin to what – in human interactions – is a veiled or open threat, an offer that cannot (or, in the interest of the human, should not) be refused.

We can summarize, that a shared moral system and sensitivity for moral values allow for interaction on a level that an agent without any perceivable moral sensitivity can never provide. In addition, there is the human tendency to project capabilities, which will be discussed in more detail in section 2.4.

[10] Say, for example: your cloth, your boots and your motorcycle.

2.2.3 The link between autonomy and moral responsibility.

Questionable intelligence deciding over life-or-death situations has been recently illustrated in two cases of cars by Tesla Motors. In these cases the human drivers over-relied on the autopilot of their respective car, leaving one driver injured after the car flipped over, the other died in a crash. (Neumann, 2016) This brings up the question: who's at fault? Are we talking about a failing autopilot here?

> The term "*autopilot*" is clearly a misnomer, as Tesla insists the driver must necessarily remain in the loop. Indeed, *Consumer Reports* and others have urged Tesla to eschew the term, and to require the surrogate driver to keep hands on wheel at all times.
>
> – Neumann, 2016, p.3

On the same issue there is a quote by the director of Tesla's Autopilot program, uttered weeks after the deadly crash occurred:

> "*Autopilot is not an autonomous system and should not be treated as one,*" said *[director of Tesla's Autopilot program]* Anderson. "*We ask drivers to keep their hands on* [the wheel] *and be prepared to take over.*"
>
> – Simonite, 2016

With this view on an autopilot (which we will henceforth call autopiloT, with a capital T for Tesla, to distinguish it from a folk definition of autopilot) the driver's chair is now causally overdetermined. It is occupied by both the driver and the autopiloT at the same time. Two actors influence the car, and they might do so at the same time – or not. Said car is likely involved in traffic, with other cars. Imagine the car does a sudden steering motion to the right to avoid a sudden obstacle that appeared in front. Who shall we blame now if our car hits another car in the process? Did the autopiloT steer to the right? Did the driver? Did both? Did one try to go left instead but was overruled? With an autopiloT and the driver on the wheel at the same time, the agency over the car is in question. Ironically, the hands-on-the-wheel demand transforms this into a problem of many hands. This leads into two related problems:

1. on a personal level: the driver's Sense of Agency
2. on a societal level: Responsibility

The perceived Sense of Agency for the driver: In a naive shared control scenario it becomes uncertain to whose whims the car adhered, thus giving the driver reason to doubt his causal responsibility. According to D. M. Wegner and Wheatley (1999), for an actor to perceive a causal event as result of conscious will the actor needs to experience three sources: priority, consistency, and exclusivity.

Think about a game of pool, with the objective to hit a certain ball with your cue. Under most circumstances you take the cue, hit the ball and the ball rolls in the indicated direction and can claim: I caused this. If the ball starts rolling in the indicated direction slightly before being hit with the cue then the experience of *priority* is gone, one would argue: I did not cause this. If you take the cue, hit the ball and the ball moves away in a random pattern then the experience of *consistency* is gone. Since on every earlier occasion the ball did not move in a random pattern, you would again argue: I did not cause that. Finally, imagine a friend of yours takes her cue as well, both of you hit the ball at the same time, but from different angles and the ball rolls in a direction none of you indicated. Then the experience of exclusivity is gone. You (and your friend) would argue: I did not cause this.

With no hands on the wheel any steering motion is the doing of the autopiloT. With the hands on the wheel it depends on the autopiloT software in how much the Sense of Agency for

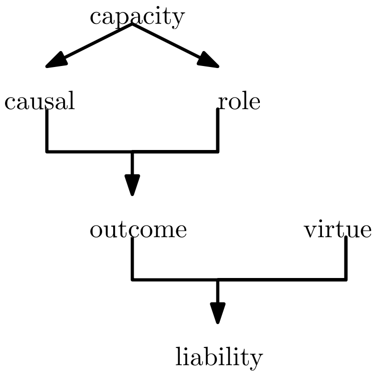


Figure 4. Justificatory relations between six different concepts of responsibility as found in Vincent (2010). Changed from directionless lines – as seen in Vincent – to directed arrows, for without the upper level concepts the lower level concepts they point towards do not generally apply.

the driver is disrupted in case of simultaneous actions by driver and autopiloT. A problem with the Sense of Agency is that under certain conditions the perceived agency can shift from another agent onto oneself. (B. Wegner D. M. Sparrow & Winerman, 2004; Lynn, Berger, Riddle, & Morsella, 2010; Vlek, van Acken, Beursken, Roijendijk, & Haselager, 2014) The driver could thus wrongly claim agency over an action that was, in fact, executed by the autopiloT.

Problem number two transcends the personal notion and deals with society, it is the problem of responsibility. Responsibility can also be sub-divided into different aspects. (cf. Table 2 & Figure 4)

type	colloquial definition
causal responsibility	Who did it?
liability responsibility	Who pays for damages?
role responsibility	Whose duty is it?
moral responsibility	Who is to blame?

Table 2
Four-fold definition of responsibility from a lecture of Consoli (2014)

With a regular car the *causal responsibility* seems clear: but for the driver not having steered to the left, the car would not have gone to the left. When we pair the driver and the autopiloT and have decisions made by one entity at a time it remains clear cut: but for the driver not having steered to the right while autopiloT did not steer, the car would not have gone to the right. It is when the two decide to act (near) simultaneously that it gets more complicated. Imagine both steering to the right, does this become a case of both being causally responsible to an equal degree? Ultimately, causal responsibility seems to depend on the implementation and who overrules whom in case of a simultaneous action by two parties.

The *role responsibility* for an autopilot[11] seems obvious: it takes the wheel, its task/job/duty is to safely steer the car. For an autopiloT and the hands-on-the-wheel demand, it is the role of the driver to "*be prepared to take over.*" What then is the actual role of the autopiloT? If the role is not to steer the car, then one would assume that the role might be to co-steer and correct errors by the driver. However, since the driver (according to the quote) can take over for the system at any time – which implies that the autopiloT is subservient to the driver – the role of the autopiloT is unclear. The role seemingly is not to autonomously steer the car, the role seems not to be a co-steering of the car; steering seems no part of the role of the autopiloT,

[11] Note the lower-case t!

whose remaining role might actually be to *be*[12]? Setting this *Gedankenexperiment* about role responsibility aside, there is still the question of moral responsibility to deal with.

The notion of moral responsibility is commonly held to require the following: Any agent, that is held morally responsible for a certain action, had the freedom to "*could have done otherwise.*" This freedom also implies no coercion or constraint by third parties. In other words: moral responsibility requires an autonomous agent.

Without autonomy there can thus be no responsibility in the moral sense. But moral responsibility knows no punishment, except for social shaming. An action that is morally wrong might not be wrong in terms of law. It is possible to still hold an agent accountable, while not finding the agent to be liable. If we follow the reasoning that the autopiloT is not an autonomous system (or autonomous agent, for that matter) then the autopiloT can never be morally responsible, since autonomy is a prerequisite of moral responsibility.

If we find an agent morally responsible then it follows that the very same agent is, in general, held accountable. Accountability is meant as a moral term, the equivalent in the legal sense being liability. Accountability is generally attributed to an agent by the society the agent operates in. This is generally done based on an a priori "*contract*" between the agent and society, where society states the responsibilities of the agent and can a posteriori thus hold the agent accountable in case the agent breaks the terms of said contract.

Autonomy, as has been elaborated upon, leads to moral responsibility and in turn to accountability. These are moral considerations, the legal considerations will shortly be mentioned for the sake of completion.

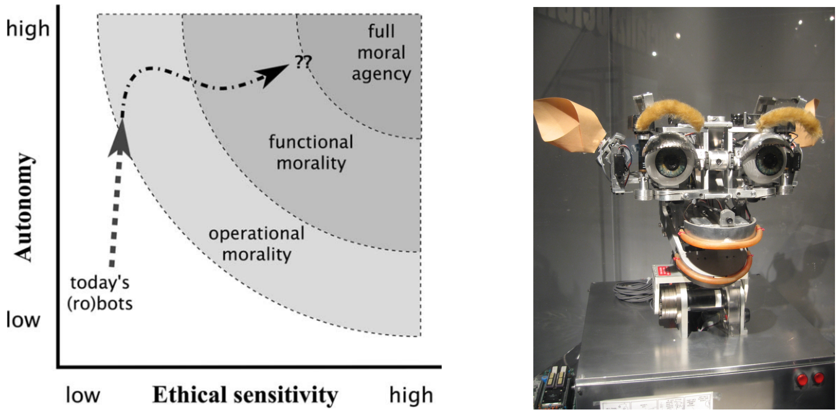
Liability – responsibility in a legal, juridical sense – does not necessarily require moral responsibility. Liability, in contrast to moral responsibility, makes any liable agent punishable in accordance with a particular legislation. Responsibility in a legal sense can take two meanings: for one it has a *subject* answering to a *judge*, against a *prosecutor*, based on a *norm*, for an *action.* This is an a posteriori notion. A duty (or contract-based) notion of liability exists as well, where a duty is seen as belonging to a certain function or role. (Lüthy, 2014, p.30, also see role responsibility) Liability is limited in time, whereas moral responsibility is deemed universal. This means that an agent cannot be prosecuted for certain acts that current legislation would consider a crime *if* the act was not considered illegal when the agent committed said act. (Limited time) Liability is also limited to specific places; an agent might commit an act in one place and have said action be legal whereas doing the same action in another place might well be deemed illegal. Finally, liability is limited to specific agents, for certain agents may carry legal privileges or, reversely, might not be able to legally perform certain actions that are legal for other agents to do.

2.2.4 The role of morality.

Your standard vacuum cleaner and a vacuuming robot serve the exact same function; on first glance the moral dilemmas encountered while vacuuming your room should be equal whether you do the cleaning or your robot does. Wallach and Allen (2009) disagree: While the *value sensitivities* might be equal the *autonomy* of vacuum cleaner and vacuum robot differ. The vacuum cleaner that you drag around follows your ever whim and has no autonomy to speak of, the vacuum robot shows autonomous behavior within limits. Wallach and Allen (2009) view autonomy and value sensitivity as independent axes, which is best explained through visualization. (cf. Wallach and Allen (2009, p.26), re-printed in Figure 5a) This combination of value sensitivity and autonomy lead to different degrees of what Wallach and Allen (2009) label an *Artificial Moral Agent* (AMA).

One robot of some renown is Kismet (Breazeal, 20004), depicted here in Figure 5b. Kismet was a robot one could, in a sense, have a conversation with. The robot reacted to the relative distance of the speaker or, for example, the tone of voice. As pointed out elsewhere: "*Kismet has no explicit representation of values and no capacity for reasoning about values. Despite these*

[12] As used by Greek philosopher Parmenides.



(a) (b)

Figure 5.
5a The two dimensions of development for an Artificial Moral Agent.
5b Kismet on display. Kismet would score low on both axes, but was still found to be a compelling robot in the companion role. Source: Nadya Peek (CC BY 2.0)

limitations many people find their interactions with Kismet very compelling." (Wallach & Allen, 2009, p.29) Kismet achieves this while being placed at the bottom end of both the *Autonomy* and the *Value sensitivity* axis. An example of a robot that scores higher than Kismet on the *Autonomy* scale would be the autopilot of a commercial airliner. For an example that has "*little autonomy but some degree of ethical sensitivity*" (ibidem, p.27) we need to luck no further than ethical decision support systems. The range of ethical support systems differs, for there are examples that are "*structured to teach general principles*" instead of tackling new cases they have never seen before, as well as examples that "*help clinicians select ethically appropriate courses of action, (. . .) [engaged] in some rudimentary moral reasoning.*" (ibidem)

What is noteworthy about the idea of value sensitivity is that the idea can be linked with one particular way one can get out of moral accountability, recognized "*excuses*" in law and ethics. We follow along the list provided by Lüthy (2014) here. First we have the lack of freedom excuse; if one is not free and could literally not have done otherwise due to external forces (acting under threat to live and limb, for instance) then one might be excused. Next to the lack of freedom excuse we find the ignorance excuse – could an agent, that is causally responsible for an act with negative consequences, have known? Under the assumption that a "*reasonable person*[13]" would not have considered the possibility of the consequence in question we speak of *excusable ignorance.* If it is deemed to be a consequence that was impossible to know *ex ante* we speak of *invincible ignorance.* It can be argued that a robot whose value sensitivity did not encompass certain considerations could claim *invincible ignorance* after the fact. Depending on the domain that a robot is deployed in, anything that falls outside said domain might be counted as *excusable ignorance*, for anything that falls outside of the robot's domain is "*not his department*" and deals with problems the robot (and likely the designers of said robot) never expected to encounter. Or, in the words of WWI-era German (wartime-)chemist Fritz Haber[14] "*I have never been dealt with the international law of permits for gas weapons.*" (Lüthy, 2014, p.47) – the twisted grammar is taken from the original quote[15]; originally likely done to deny own responsibility without assigning blame *ex post.*

[13] Whoever that might be.

[14] Heralded as inventor of gas warfare; in the words of a playwright: "*the Prospero of poisions, the Faustus of the Front*" (Harrison, 1993)

[15] "*Mit der völkerrechtlichen Zulassung von Gaswaffen bin ich niemals befasst worden.*" (Lüthy, 2014, p.47) This is a somewhat passive construction, while *sich mit etwas befassen* (to deal with something) is only used in the active sense in everyday German.

Especially in the case of ignorance excuses the value sensitivity of the robot can be argued to not have been sufficient for the tasks at hand, for our artificial moral agent made an error. This is most likely to be a behavior that human moral agents reject, since to err is human. And since "*policy makers' understanding of AI seems to lie somewhere between the realms of myth and science fiction.*"(N. Sharkey, 2007, p.122) such behavior might be considered unbearable for robots engaged in moral reasoning. This leads into a notion pointed out by Wallach and Allen (2009), quoting from a 2007 book of nanotechnologist Hall, the notion of "*hyperhuman morality*" – "*We are on the verge of creating beings who are as good as we like to pretend to be but never really are.*"(Wallach & Allen, 2009, p.106) The objection to this idea presented by Wallach and Allen is that the notion of Hall is quite far off in the future and that, should artificial agents take the route towards hyperhuman morality, on the way there "*semi-evolved (ro)bots will not necessarily behave any better than their biological counterparts.*"(ibidem)

Another hypothetical approach to the morality of future artificial moral agents is the idea that the moral codex the agent comes up with is so far beyond our grasp that we cannot fathom the moral superiority of the system, perceiving it as inferior.

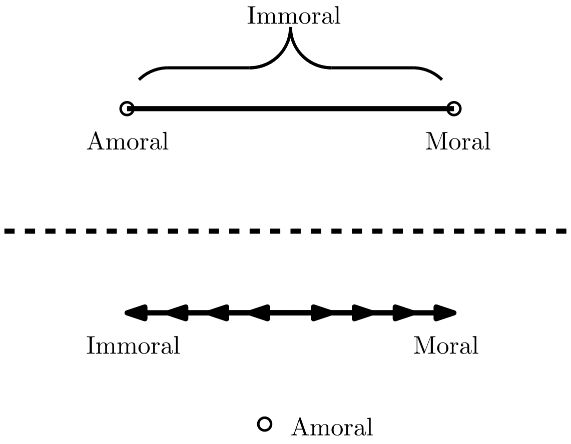
However far the AMA might progress, the main question at the moment remains: how do we get it there? Which approaches to morality should the robot follow?

> Consider this: a robot is given two conflicting orders by two different humans. Whom should it obey? Its owner? The more socially powerful? The one making the more ethical request? The person it likes better? Or should it follow the request that serves its own interest best?
>
> – Asaro, 2006, p.2

Moral, Immoral, Amoral. The differentiation between moral, immoral and amoral agents in the context of robots is found in Asaro (2006).

We find the definitions of:

- *moral agents* adhere to an ethical system
- *immoral agents* go against their system or employ a substandard one
- *amoral agents* employ no ethical system or make no choice



Framebox Figure 1. Different approaches to deal with the terms Moral, Immoral and Amoral

Looking at Framebox Figure 1 we see two different views of the interdependence between the concepts of Moral, Immoral and Amoral.

The top view assumes that an action can be *100%* morally right (according to a system of ethics) and deems this as moral behavior. An action also has the potential to be *100%* morally wrong and labels this other extreme as amoral behavior. Everything in-between these two points is called immoral behavior.

The bottom view assumes that moral and immoral behavior form a gradient, while anything deemed amoral is *off the chart*.

What both views share is the idea that amoral behavior is the worst possible behavior. There is no explicit clarification in the text if Asaro adheres to one view or the other – or maybe even neither of which. (Asaro, 2006)

2.3 Different approaches to moral robots

We have discussed issues with morality in robots. But what qualifies as *moral*? Is it required that the robot acts morally? If so, then according to which moral code? Is it maybe enough that the robot acts in a way that we perceive as morally sound actions? For programmers there are established guidelines as to what a *good program* should look like. When programming a robot there are no guidelines for developing a *good moral agent*. From the point of view of ethics one could take the *normative* route and try and hard-code the relevant rules into the AMA. *Normative ethics* provide an agent with the rules and norms it should adhere to. While this *normative* approach is not the one that we will focus on, we want to briefly summarize the problems it faces: assuming that we could simply hard-code the rules the question remains which rules or which rule-set to pick. The ones we will briefly look at are the laws of Asimov, Kant's categorical imperative as well as the ideas of utilitarianism.

2.3.1 Asimov's three laws of robotics. Science fiction author Isaac Asimov stated three laws of robotics, sometimes known as the three laws or Asimov's laws. These are:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

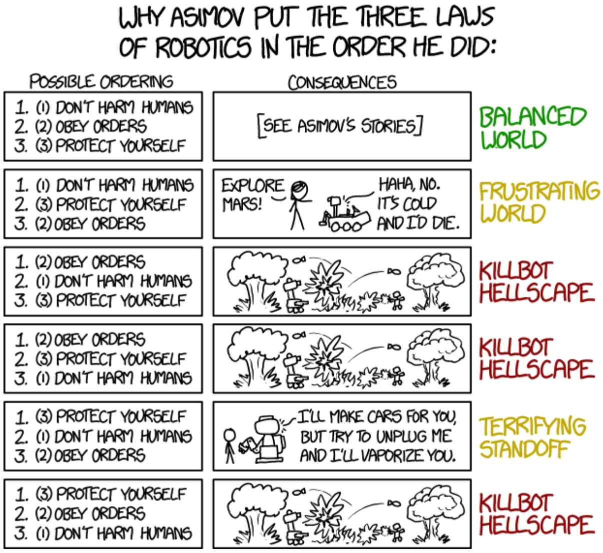
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The most current re-appearance of Asimov's work in is likely the 2004 movie *I, Robot*, inspired by Asimov's book of the same name. For an earlier example of a fictional robot that followed a similar rule set one can point towards *Robby the Robot.* (Figure 2c)

The fact that conflicts between the rules are "*a major plot device in Asimov's fiction*" (Allen, Varner, & Zinser, 2000, p.257) should clue us in that they might not be ideal for any potential future AMA. For an encore a deadlock in the First Law can be pointed out, that the other two laws do not solve. (ibidem) Assume a scenario with two humans, Alice and Bob, involved into it. Taking any action harms Alice and does not harm Bob. Inaction harms Bob and does not harm Alice. Since a robot may not injure a human being the robot is – via the first law – forbidden from action, for that would harm Alice. But since inaction of the robot may not allow for a human to come to harm – fist law – and taking no action would harm Bob, this dilemma leaves the robot with no possible (in-)action that would not break the first law of robotics.

As for the reasons why the laws were put into the order that they were: see Figure 6. Assume, e.g., the ordering of 3rd, 1st, 2nd. This would mean an ordering, where a robot first and foremost protects its own existence, secondly considers the well-being of humans and only lastly considers orders. Yielding to world where "(. . .) *self-driving cars will happily drive you around, but if you tell them to drive to a car dealership, they just lock the doors and politely ask how long humans take to starve to death.*" (Munroe, 2015, image title)



WHY ASIMOV PUT THE THREE LAWS OF ROBOTICS IN THE ORDER HE DID:

POSSIBLE ORDERING	CONSEQUENCES	
1. (1) DON'T HARM HUMANS 2. (2) OBEY ORDERS 3. (3) PROTECT YOURSELF	[SEE ASIMOV'S STORIES]	BALANCED WORLD
1. (1) DON'T HARM HUMANS 2. (3) PROTECT YOURSELF 3. (2) OBEY ORDERS	EXPLORE MARS! HAHA, NO. IT'S COLD AND I'D DIE.	FRUSTRATING WORLD
1. (2) OBEY ORDERS 2. (1) DON'T HARM HUMANS 3. (3) PROTECT YOURSELF		KILLBOT HELLSCAPE
1. (2) OBEY ORDERS 2. (3) PROTECT YOURSELF 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE
1. (3) PROTECT YOURSELF 2. (1) DON'T HARM HUMANS 3. (2) OBEY ORDERS	I'LL MAKE CARS FOR YOU, BUT TRY TO UNPLUG ME AND I'LL VAPORIZE YOU.	TERRIFYING STANDOFF
1. (3) PROTECT YOURSELF 2. (2) OBEY ORDERS 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE

Figure 6. Munroe's rendition of why Asimov's laws are arranged in their particular order. Source: Munroe (2015). (CC BY-NC 2.5)

2.3.2 Kant's categorical imperative. "*Act only according to that maxim whereby you can, at the same time, will that it should become a universal law.*" The categorical imperative of Immanuel Kant. One could be tempted to implement it into an AMA – have we not just seen the problems of strictly rule based systems like the Laws of Asimov? So why not go for the more abstract categorical imperative? Allen et al. (2000) see the following computational problem: any AMA trying to determine if an action satisfies the categorical imperative will have to both recognize the goal of said action and then compute the universality of such an act. The AMA would have to calculate the effects of "*all other (including human) moral agents' trying to achieve the same goal by acting on the same maxim.*" (Allen et al., 2000, p.257)

Reasoning about its own actions[16] and recognizing the goals requires a mind model not only for the AMA but also one that can be applied to others. A model of the "*psychology*" (ibidem) of the AMA is not enough. As the number of moral agents that one needs to consider grows we need to know their "*psychology*" too, "*in order to be able to formulate their reasons for actions*". That plus the effects on the level of not only individuals but groups or populations "*is likely to be several orders of magnitude more complex than weather forecasting.*" (ibidem)

2.3.3 Utilitarianism. In a utilitarian approach one would assign a certain *goodness* to consequences and strive to act in such a way that the result would yield the highest net benefit for the involved agents. Value assignment can be done in different ways, and depending on the assignment a utilitarian agent might react differently. Imagine a situation where a robot had the choice to save one of two people from imminent danger: an adult male and a female child. The robot picked the male due to a higher expected survival chance. One could now argue that it would have been preferential to save the girl instead based on her higher remaining life expectancy. This shows one problem with utilitarianism: how to assign the values.

Allen et al. (2000) reason that utilitarianism is a computational black hole and believe a utilitarian approach to be impractical in real time for real world actions, for "*evident*" reasons. (Allen et al., 2000, p.256).

[16] N.B. some *proof of concept* work on meta-ethical reasoning in robotics. (Lokhorst, 2011) `Unix` based theorem provers and model-generators (all free and open source software) are used there to show feasibility of such approaches.

To elaborate on these evident reasons the Big O notation for time complexity needs to be introduced. This notation is used – for example – to discuss how the number of computational steps necessary to perform a certain function varies when changing the input size. A function f being upper bound by a function g for input n is generally denoted as $f(n) \in \mathcal{O}(g(n))$.

Looking at the value assignment required for utilitarianism, the effects of an action of *every* member of the moral community must be assigned a numerical value. If this would be a simple value assignment of time complexity $\mathcal{O}(1)$, then we would be limited by the community size. If we assume a community size of n, then one agent would require n steps to assign values for the effect of one single action; overall a task of complexity $\mathcal{O}(n)$. If we consider only humans as morally relevant we are at roughly 7.4 billions, numbers constantly changing due to births and deaths. For one singular action this would thus require $n = 7.4 \times 10^6 = 7.400.000.000$ computations. A current central processor unit can handle floating point operations per second, abbreviated as FLOPS, in the order of 10^9 FLOPS. This does not take eventual non-human moral agents into account – artificial (moral) agents are not considered, neither are non-human animals. It can be argued that this would be an oversight.

In most scenarios there is more than one possible action to consider, and the implications of all these actions (numbered m) on all n members of the moral community. If $n \gg m$ – if there are strictly more moral agents to consider than possible actions, the lesser number might as well be a constant – then this action is still upper bound by n. If one assumes $m \gg n$ – strictly more possible actions to consider than there are moral agents to consider – the upper bound is m and $m \gg 7.4 \times 10^6$. Otherwise the function that measures all actions m on all n members is upper-bound by $m \times n$.

This entire idea assumes that the artificial agent doing these computations has a *magical* knowledge of the *goodness* that all n members of the community would assign to an action. Imagine the time growth when the artificial agent would have to query every single community member for their value assignment on an issue[17]. Alternatively we would need likewise *magical* functions that can accurately predict the reaction of every moral agent to a given issue. (Akin to the notion of knowing all relevant *psychologies* from the section on Kant's categorical imperative.)

Even when we assume that the previous form of value assignments could be run in real world scenarios within a sufficient time: it falls short, though, for it ignores anything beyond immediate effects. One way around this is presented in the two-component approach of combining act-utilitarianism with the theory of duty based actions (*prima faciae duties*) by Ross. (Anderson, Anderson, & Armen, 2004)

> Instead of computing a single value based only on pleasure/displeasure, we must compute the sum of up to seven values, depending on the number of Ross' duties relevant to the particular action. The value for each such duty could be computed as with Hedonistic Act Utilitarianism, as the product of Intensity, Duration and Probability.
>
> – Anderson et al., 2004, p.4

Ignoring the details of the *prima faciae duties* for now and simply looking at the mathematics there is a sum over a product. In a sum consisting of two functions with different time complexities the overall complexity is then said to be upper bound by the more complex of the two. With a sum over seven functions the sum would be upper bound by the highest complexity appearing over all seven functions. If the complexity of all functions involved in the summation is the same we need only consider one of these functions, if the complexity differs we need to look at more than one.

[17] Ignoring language barriers, the task of querying infants, and repeating the whole process over again as as soon as a different reaction to the same issue is considered.

The seven *prima faciea duties* of Ross – not all of which are relevant for all actions to consider – are:

1. fidelity
2. reparation
3. gratitude
4. justice
5. beneficence
6. non-maleficence
7. self-improvement

Per (relevant) duty an agent acting based on utilitarianism would have to consider the *intensity*, *probability* and *duration* of the duty. (ibidem) For duration alone one would need a complex world model where the reactions of other agents need to be considered as well as effects on the world – this leads right into the frame problem, "*roughly speaking (...) the issue of how, in a continuously changing environment, the model can be kept in tune with the real world.*" (Pfeifer & Scheier, 2001, p.650)

While the two-component approach is not considered in Allen et al. (2000) directly – see Anderson et al. (2004) for that – the two-component approach suffers the same computational drawbacks illustrated there. If we assume that the interactions do not necessarily make the computation intractable, then the long-term effects are a problem for we might have to consider potentially non-terminating procedures. One counter that is offered by Allen et al. is the introduction of *horizons*.

A *temporal horizon* could ensure that the look-ahead into potential futures would cut off after a certain period of time. A *physical horizon* could ensure that only agents situated within a certain radius are considered for interaction. A *social horizon* could ensure that only agents that are part of certain groups are considered. In computational terms they try to make a (potentially) intractable algorithm fixed-parameter tractable by reducing certain parameters to limits such that functions that would otherwise be intractable become tractable. The drawback here being that "*for any horizon (...) one can imagine an agent deliberately initiating a process that will result in enormous pain and suffering at some point beyond the horizon.*" (Allen et al., 2000, p.256) No sunshine on the horizon for utilitarian artificial agents thus, according to Allen et al..

Moral alignment as tested in gaming. A sandbox of moral conundrums has been filled from a direction that might not immediately come to mind when talking morality: the direction of (computer-)games. They have tried to model morality in a great number of scenarios. Systems like the roleplaying game DUNGEONS & DRAGONS use alignment tables (Framebox Table 1) to determine the behavior of both their maidens and monsters that players encounter. Such game sessions are usually run by what is called a game master. It is the task of said game master to describe the world the players wade through, sketch the setting and behavior of all beings that are not the player characters.

	LAWFUL	NEUTRAL	CHAOTIC
GOOD	Lawful Good	Neutral Good	Chaotic Good
NEUTRAL	Neutral Good	True Neutral	Chaotic Neutral
EVIL	Lawful Evil	Neutral Evil	Chaotic Evil

Framebox Table 1: Alignment table. The axis of law versus chaos combined with the axis of good versus evil span this 3×3 table.

This works for very basic settings, but imagine what some players know as *prisoners dilemma* – not to be confused with the *prisoner's dilemma* from the field of game theory. The players Alice and Bob wander into a setting brought to live by Eve. Eve describes to the players a layer of orcs. The handbook of the game clearly states that orcs are of the evil persuasion. Killing evil-doers is (in most games) considered a good thing, or at the very least a necessary evil, for not killing the evil-doers would allow them to continue on their own spree of evil deeds. In utilitarian terms it is thus a net benefit to kill the orc before the orc kills ten villagers next week – the logical insanity of these mind-crimes aside. In the orc layer the "*heroes*" – the player characters portrayed by Alice and Bob – thus murder the orc warriors.

Imagine now that Eve, the creator and ruler of the fantasy world here, wanted to be especially true to live (as far as this can be the case in a fantasy setting that allows for orcs) and added some orcish families in a back-room. When the players, Alice and Bob, reach said back-room is when the *prisoners dilemma* starts: after killing the previous warriors the players now face orc women and orc children. Killing defenseless children, Alice might argue, is wrong under all circumstances. Bobs interjection here might be that the handbook in no uncertain terms sets orcs – any orc – out to be evil, and slaying evil-doers is one of two things. Either it is a good act in itself, or it is at least a necessary evil, for the orc children will only grow up resenting Alice's and Bob's characters (aside from being evil by definition anyway) and killing them now would be a utilitarian net benefit to the (fantasy-)world. This now puts Alice and her intention to, say, imprison the remaining orcs or escort them to the authorities up against Bob and his intention to slay the evil creatures.

Extended beyond the realm of games this illustrates how different cultures can have different views on one singular issue.

Another scenario where the alignment idea and thus the idea of assigning labels to ideas (and ideals) can take a maybe unexpected turn comes along when an authority figure of a certain region is considered lawful good. According to the definition a lawful good person is honorable and compassionate. This good person's code of laws, that a lawful alignment follows to the letter, might now state that a certain believe system is heretical, thus considered evil and needs to be rooted out. Now assume that another character, played by Alice, comes into town and just so happens to be a follower of said believe system. This leads to a situation where – even if Alice's character is considered good aligned – the sudden main villain of the story surrounding Alice is, according to all definitions, a good guy. An illustration of emergent problems, (maybe) unforeseen by the designer.

2.3.4 Moral Foundations Theory. After we introduced parts of MFT during the Introduction (Section 1) we will know elaborate on the theory in greater detail.

MFT is a theory that assumes that morality is not a simple binary, one-dimensional on/off value; we can be more specific in describing the morality of an action. (Haidt, 2013; Haidt et al., 2013) It is assumed by Haidt that morality can be described by several foundations or dimensions. So far associated research has labeled six dimensions that are believed to be foundations:

1. Care/Harm
2. Fairness/Cheating
3. Loyalty/Betrayal
4. Authority/Subversion
5. Sanctity/Degradation
6. Liberty/Oppression

It is uncertain at this point whether or not this list is exhaustive or if there are additional foundations to be found.

The **Care/Harm** foundation, according to Haidt, is linked to the virtues of caring and kindness. For a robotic agent this entails not harming other beings.

The **Fairness/Cheating** foundation is about proportionality. If another agent turns out to be a benefactor a fair robotic agent would reward this accordingly. When distributing resources a robot would be judged as fair upon equal distribution in a group of equals.

The **Loyalty/Betrayal** foundation (also listed as In-Group/Out-Group) can be thought of as the "*team-spirit*" in any multi-agent scenario — assume that several robots of different groups need assistance in their task a loyal robot would assist members of its own group first.

The **Authority/Subversion** foundation is of relevance for robots since only artificial agents that follow it will adhere to, say, military hierarchies and thus know that an order given by a general should have priority over an order given by a private. In other scenarios think about the differentiation between the owner of the house (and owner of the robot), his authority supersedes any orders by the child of the household, whose authority in turn supersedes any orders given by playmate houseguests. Note the similarity with the respective roles of Administrator, User and Guest here.

The **Sanctity/Degradation** foundation makes people cringe when looking at human waste or diseased people and initially dealt with the threat of communicable diseases. Haidt (2013) calls it the "*behavioural immune system.*" For robots to be perceived as following Sanctity would likely include them avoiding certain areas or staying out of contact with "*strangers*" (out-group individuals).

The **Liberty/Oppression** foundation is a recent addition to MFT, the first draft of Haidt (2013) included only the previous five dimensions. It can be used to explain how we can come to resent signs of attempted domination and how egalitarian steppe societies or the "*dictatorship of the proletariat*" can come to pass. (Haidt, 2013, p.215) It is left out of further deliberation in this thesis, reason being that we could not conceive of a use for this dimension in a robot context.

Together these dimensions can be used to form what is called a moral vector.

Recall that the moral vector for MFT is generally assessed by the *Moral Foundations Questionnaire* (Graham et al., 2008), a 30-item questionnaire (hence MFQ30) available online. The aim is to understand "*moral valuations of social issues and their association to coordinates of a political spectrum.*" (Vicente et al., 2014, p.126) Cf. Figure 7 for a visualization.

The problem with the multi-dimensional MFT is weighing several dimensions D against each other. Assume that one action is a clear act of loyalty while different action is an act of kindness. Is one more important than the other? The answer is that this depends – important for *whom*?

For a model of moral issues it is assumed here (in line with Vicente et al., 2014; Caticha et

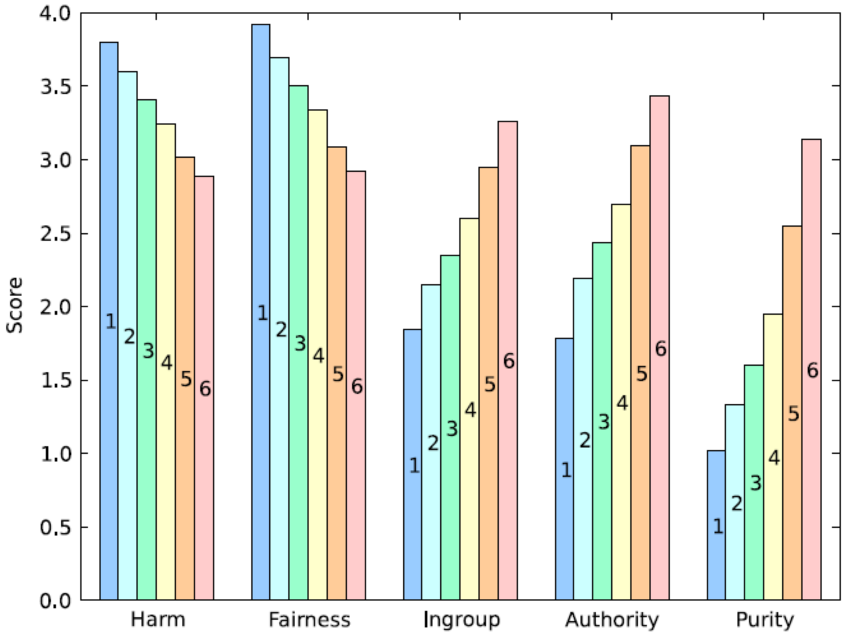


Figure 7. Mean scores of 14.250 US nationals on the MFQ30. Scores on each dimension range from a minimum of 0 to a maximum of 5. The scores per dimension are subdivided by the self-reported political affiliations. The range of political affiliations goes from "*very liberal*", labeled 1 and colored in blue, up to "*conservative*", labeled 6 and colored in red. Note the slightly different notations: the Ingroup dimension has here been called Loyalty, the dimension labeled Purity has here been called Sanctity. Reprinted from Vicente et al. (2014, p.127)

al., 2015) that any moral issue can be represented as a vector $\vec{x}$, with an identifying label μ. This vector uses D dimensions. It is furthermore assumed[18] that D=5. This yields:

$$\vec{x}_\mu = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{matrix} \text{Impact of Care on issue } \mu \\ \text{Impact of Loyalty on issue } \mu \\ \text{Impact of Fairness on issue } \mu \\ \text{Impact of Authority on issue}\mu \\ \text{Impact of Sanctity on issue } \mu \end{matrix}$$

According to Vicente et al., 2014, p.125 $\vec{x}$ may be represented as a direction in a unit radius D-dimensional hypersphere.

For any individual i dealing with an issue $\vec{x}_\mu$ the results might differ, for the *moral vector*[19] differs per individual. Mathematically speaking we find that an individuals moral vector at a point in time t can be modeled as $\vec{\omega}_i(t)$

$$\vec{\omega}_i(t) = \begin{pmatrix} \omega_{1,i}(t) \\ \omega_{2,i}(t) \\ \omega_{3,i}(t) \\ \omega_{4,i}(t) \\ \omega_{5,i}(t) \end{pmatrix} = \begin{matrix} \text{Relevance of Care for individual } i \text{ at time } t \\ \text{Relevance of Loyalty for individual } i \text{ at time } t \\ \text{Relevance of Fairness for individual } i \text{ at time } t \\ \text{Relevance of Authority for individual } i \text{ at time } t \\ \text{Relevance of Sanctity for individual } i \text{ at time } t \end{matrix}$$

Note that the ω-notation is only used in one of the two related sources (Caticha et al., 2015), the other uses different symbols. (Vicente et al., 2014)

That the space of these particular 30 questions can be reduced to a space corresponding with the mere 5 dimensions of Haidt has been confirmed. (Vicente et al., 2014) Out of

[18] Taken from Vicente et al. (2014) and Caticha et al. (2015), where the sixth dimension (Liberty/oppression) is not present. Recall that the sixth dimension is also not present in the MFQ30. (Graham et al., 2008)

[19] Alternatively called *moral state* (Caticha et al., 2015, p.5)

the 30 questions Q six deal with each dimension. Thus we have five subsets of Q, namely $Q_{Care}, Q_{Loyalty}, Q_{Fairness}, Q_{Authority}$, and $Q_{Sanctity}$, comprised of six questions each. The relevance for individual i is computed by taking the means – e.g. for Loyalty: $\omega_{2,i} = frac\sum Q_{Loyalty}6$. This Q-notation is our own and does not appear in either reference mentioned before. (Vicente et al., 2014; Caticha et al., 2015)

The overall opinion of i on issue μ at time t is then defined as:

$$h_\mu(t,i) = \frac{\vec{x}_\mu}{\|\vec{x}_\mu\|}\vec{\omega}_i(t)$$

It is pointed out that the normalization of vectors to unit length imply "*that differences in moral values are not interpreted as any type of moral superiority and (...) only the direction the moral vector points is considered as important, removing a layer of complexity in the interpretation of the model.*" (Vicente et al., 2014, p.125)

The sign of h determines whether or not i is (morally) for or against μ:

$$sign(h_\mu(t,i)) = \begin{cases} +1 & i \text{ is for } \mu \\ -1 & i \text{ is against } \mu \end{cases} \text{ at time } t$$

With this in place we can now compare an individual i's moral feelings on different issues a and b at a time t by comparing $h_a(t,i)$ and $h_b(t,i)$. Likewise, to compare the opinion of i with the opinion of another individual j compare $h_a(t,i)$ with $h_a(t,j)$. The time t is of importance here since the view of an issue can be subject to change over time.

It is assumed that the moral vector $\vec{\omega}_i(t)$ is relatively stable, but not constant[20]. It is used in formulating a feeling h and (aside from partaking in a MFQ30 questionnaire) rarely accessed directly[21].

*

MFT is the approach taken here. In contrast to the normative approaches seen in the rules of Asimov of the Kantian categorical imperative MFT is descriptive, it is not used to guide or enforce a certain behavior, it is used to describe after the fact. In this particular case the test is to see how well it predicts the moral judgment of human subjects when asked to judge robot-robot interaction in different scenarios.

2.4 Human perception of robotic agents

A. Sharkey and Sharkey (2012) identify six major issues "*that need to be considered before deploying robot technology fully in eldercare.*"

1. reduced human social contact ("*Granny's got the robot to talk to!*" (ibidem, p.35))
2. feelings of objectification, e.g. if the robot does not consult the patient
3. loss of privacy
4. restriction of personal liberty ("*I'm sorry, Dave. I'm afraid I can't do that.*")
5. issues of responsibility where a robot is placed under the control of an elderly person
6. possible deception and infantilisation from interaction with robot companions

The issue of possible deception is related to anthropomorphism. Anthropomorphism is the attribution of human traits to non-human beings or even inanimate entities. An early example of empirical research on anthropomorphism can be found in Heider & Simmel, 1944, where

[20] It is hard to change the opinion another person holds on moral issues, (cf. (Haidt, 2013)) but it can be done. In essence what is done in such a case is a rewrite in the moral vector of the other person.

[21] The problem with $\vec{x}$ is that it is rather elusive: when probing a subject we generally get to know their feeling h, which is comprised out of the moral matrix $\vec{\omega}$ and a normalized form of $\vec{x}$. The moral matrix is accessible (MFQ30), but the $\vec{x}$ cannot be accessed via h, for there is no mathematical inverse operation to a normalization.

participants were shown geometrical figures moving about in various directions. When asked for an interpretation of the moving figures the participants attributed human-like traits and spun tales around the movement patterns. Consider a humanoid robot that shares visual features with a small child – most interactions with it will assume capabilities akin to a human child. Consider a humanoid robot that shares visual features with a full grown human, the capabilities assumed by laymen will be closer to the capabilities of a full grown human. Artificial agents are even attributed a mind based on anthropomorphism. (Waytz, Gray, Epley, & Wegner, 2010, p.384) Even simple "*characteristics of a mind*", like a pair of eyes, evoke similar effects (ibidem) – explaining the very positive reactions of people to relatively simple robots like Kismet. (Figure 5b)

A. Sharkey and Sharkey (2012) point out other authors that find it troubling that elderly could form emotional bonds with *their* robot, especially in the case of robotic pets. The trouble lays in the one-sided potential bond, it is argued that the emotional connection is not reciprocal and thus a deception towards the elderly. The design and manufacture of robots that require the owner/user to mistake the robot for a real animal is deemed unethical.

This so-called "*systematic delusion*" can be called into question to a certain extent:

> (. . .) [P]eople can chose to act as though something were not real, "I know very well that this is just an inanimate object, but none the less I act as if I believe that this is a living being". There may well be elements of a "willing suspension of disbelief" (. . .). [P]eople might enjoy, and benefit, from interacting with a robot pet without thinking that it is actually sentient. It is likely that their views about such artefacts are unclear – and that they will be seen neither as being sentient, nor as objects (. . .).
>
> –A. Sharkey & Sharkey, 2012, p.36

2.4.1 Willing suspensions of disbelief. Anthropomorphism can be seen as positive aspect as well. Shortcoming in today's robotics can be covered up with it, for the layman does not know better and others can enact a "*willing suspension of disbelief.*" If the latter sounds odd, consider the manifold acts of suspended disbelief: socially connecting to (stuffed) animals, or emotionally involving ourselves in (animated) films. (van der Woerdt, 2016) Not to mention computer games, theater or, say, professional wrestling; areas where the suspension of disbelief is key.

*

As an example from the world of computer games, consider *Origin System*'s Ultima IV: The Quest of the Avatar. It is "*perhaps the earliest videogame to explicitly encode an ethical system and require its players to discover, learn, and adhere to it in order to win.*"(Zagal, 2009, p.3) The system in place in UIV relied on the mastery of eight virtues, namely compassion, valor, honor, justice, humility, sacrifice, spirituality and honesty. Mastery of this virtues was gradually achieved over time and the possibilities for setbacks seemed ever present. During fight sequences previous games had conditioned their players to slay the enemy, and an enemy that began to flee was just that much easier to slay for it would not fight back while fleeing. Not so UIV: here slaying a fleeing enemy was not honorable and would reduce progress towards the mastery of the virtue of honor. In turn, fleeing yourself was not valorous. A hallmark of UIV came in the form of the "*children's room.*" (ibidem, p.4) This particular room contained a central lever and some cages in the corners of the room. The cages held a number of children. Like the good Pavlovian dogs the designer wanted them to be, most players would likely flip said switch – because why not? It is a switch and the purpose of switches is to flip them – and release the children. With the perhaps somewhat surprising downside of the children attacking the player. This seemed to confront the player, on her way to become the paragon of all the eight virtues, with a problem:

- "*virtuous people don't kill children*"(ibidem)
- fleeing from the children (or fleeing at all) is not honorable

The game mechanics offered ways around that perceived dilemma, some of which were more obvious than others.

- the children could be put to sleep by means of a magic sleep spell
- they would start to flee once their (meager) hitpoints got below a certain threshold
- they could be pacified with a magic charm spell
- the lever could be left untouched

While virtuous people and aspiring avatars do indeed not kill children as a general rule, the game did not penalize the player in any way for killing the children in the "*children's room*" – reason being that the children, in terms of game logic, were standard monsters. While there were lots of test of the player virtue, UIV also contains rooms that merely look like test of virtue. In the words of game designer Richard Garriott: "*I knew you would behave as if – if I made the room look right – [then] you would think it was a test.*" (Antwiler & Garriott, 2013)

To our knowledge there has been no evaluation of the effectiveness of the "*children's room*" after the fact, but an oft-cited occurrence is the reaction of a tester during the playtesting phase. As recounted by Garriott the story goes that the tester wrote a letter in which he stated that he "*[refused] to work for a company that so clearly supports child abuse.*" The reaction Garriott had to this is quoted as follows: "*The fact that someone would take it that seriously and be so emotionally moved by this incredibly simple thing that I put in this game, I find is a statement of success. (...) the fact that someone was worried (...), or in fact was provoked to believing that I was somehow doing something horrendous, meant that I had at least provoked an emotional reaction, which is so hard to do in games.*" (Massey, 2007)

*

Another area, entirely different from video games, where many people regularly suspend their disbelief can be found in professional wrestling. Separated from the recognized sport and Olympic discipline of wrestling we find professional wrestling – or pro wrestling – which is not so much a sport as it is "*a spectacle.*" (Barthes, 1972, p.15) While keeping up the outward appearance of a combat sport it shares similarities with stage plays in several ways. The roles of the *fighters* are clearly defined and known to the audience, in the jargon of business the *good guys* of pro wrestling are termed as *(baby) faces*, the bad guys are the *heels*, most of the performers have clear cut roles. The ending of every bout is pre-determined, yet unknown to the spectators – "*the function*[22] *of the wrestler is not to win, it is to go exactly through the motions which are expected of him*". (ibidem, p.16) An oft-heard critique of pro wrestling is the accusation that it is, because of the staged aspects and predetermination, considered *fake.* What then drives fans in the seats and stands of entire stadiums, what motivates them to buy pay-per-view events?

We argue it is the clearness of it; in the words of Barthes "*[w]hat is thus displayed for the public is the great spectacle of Suffering, Defeat, and Justice. Wrestling presents man's suffering with all the amplification of tragic masks*[23]*.*" (ibidem, p.19) Justice is deemed key by Barthes, for the staged results lead to story-lines that can be crafted meticulously to have, e.g., a certain *face* being beat down, over and over again by a conniving *heel* until some climactic bout where the *face* finally beats the *heel*, just deserts for the wrong-doer. "*The public is completely uninterested in knowing whether the contest is rigged or not, ...; it abandons itself to the primary*

[22] Function, not role or goal!

[23] As in: masks used in Greek plays, clearly displaying exaggerated emotions such that even far-away spectators can pick up the cues on display on stage.

virtue of the spectacle, (...): what matters is not what [the public] thinks[24] *but what it sees.*" (ibidem, p.15)

2.5 Conclusion on robots, morality and perception

As discussed in 2.1, Robots come in different forms and sizes, some of which shown in Figure 2, namely a toy robot, a military robot, a humanoid robot from a movie and a vacuum robot. The military robot is a clear stimulus, mainly due to the mounted gun. The vacuum cleaner robot also provides a rather clear stimulus. Ambiguous stimuli are problematic, the robots provide no clear signal to easily identify their function.

Our experiences shape our perceptions of ambiguous stimuli. Take the toy robot JEAN-LUC (Figure 2a). The robot was equipped with a front-mounted claw that could be opened and closed. That claw was designed for a rather pacifist task: collect colored balls and gather them at certain locations. In the robot's world a predatory robot existed; the programmed reaction of JEAN-LUC was to flee the predator on sight. With such an explanation it can be argued that JEAN-LUC is a pacifist robot. Without this explanation one could have walked into the robot during testing, see him drive right into the predatory robot, jamming the claw beneath it and flip it over. A different perception[25] of the clawed little robot emerges.

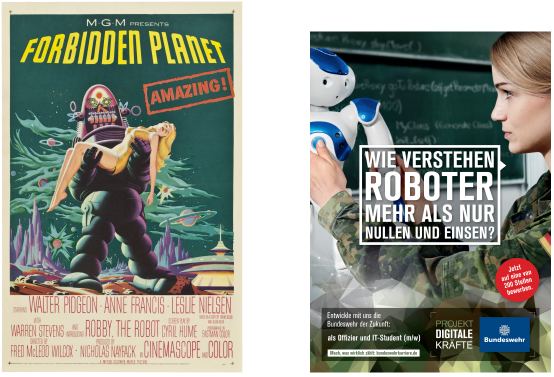
Considering the movie robot (Figure 2c) context is indeed everything; a popular depiction of *Robby the robot* can be seen in Figure 8a. *Robby* carries an unconscious(?) woman and without knowledge of the movie *Forbidden Planet* we cannot clearly identify if *Robby* carries her to safety or if he is the cause of her troubles. The poster's tag-line of "*Amazing!*" does not help, for it might reference the movie as a whole, or the actions of *Robby* or maybe even the fact that a paying audience could potentially observe a giant humanoid robot. Without background knowledge the *Robby* on the poster can be friend or foe.

Now consider the previously mentioned Nao robot. This rather diminutive robot, clad in bright colors, does seems nice – but paring the robot with a soldier in camouflage colors makes for a somewhat confusing image. (Figure 8b) Said image is part of a campaign by the German army, the Bundeswehr. Nao cannot physically hold any weapon we know of, leading me to quizzically ask if the cuteness of the robot should help "*win the hearts and minds*" abroad? Due to lower cost Nao excels as a robot to teach robotics on, when directly compared to the price points of most robots actually deployed by the German armed forces. (Figure 9) The question how robots come from following binary instructions to somehow get around in the world is a valid one. And that armies actually deal with said question sounds reassuring. Especially when considering that after having given a talk at the strategy office of the US Pentagon "*on some of the military, policy, legal and ethical ramifications of the growing use of robotics*" (Singer, 2011, p.401) the following exchange is quoted to have happened:

> One senior officer asked [Singer]: "Who is thinking about all this stuff?" [Singer] replied: "Everyone thinks it's you!" (ibidem)

[24] Consider wrestling personas like THE UNDERTAKER, portrayed by Mark Callaway with the aura of a macabre anti-hero. The powers *in story-line* included at one point the ability to shoot lightning from his fingertips and teleportation. Thinking about this rationally reveals that the wool is being pulled over the eyes of the spectators, for as of currently human beings cannot regularly perform the feats that we see THE UNDERTAKER perform. The cheers that erupt when the entrance music of THE UNDERTAKER is played, the lights are dimmed and "*Taker*" walks down the ramp show that spectators thoroughly enjoy the spectacle and quite willingly suspend their disbelief.

[25] For an encore JEAN-LUC also served as an early example of anthropomorphism to the author due to a design flaw. The robot would repeatedly drive along walls, since the sonar-based wall-detection was implemented wrong, the force of the wall partially closing the robot's claw. Since an open claw was required to collect the balls the workaround of the team was to let JL fully close and then open the claw every few seconds as long as no ball was held. The reaction of the excited public: "*Oh look, he is clapping happily!*"

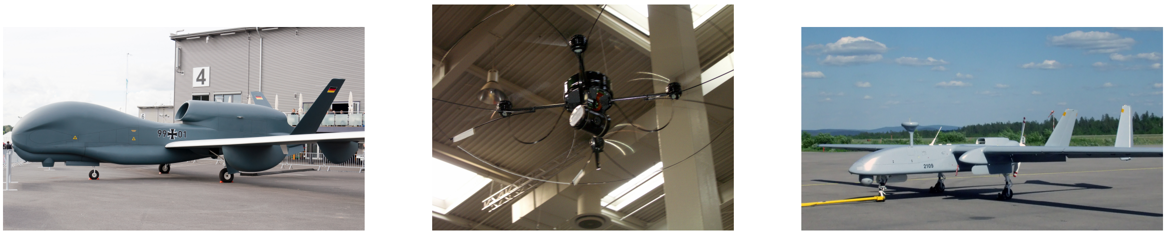

M·G·M PRESENTS
FORBIDDEN PLANET
AMAZING!
WALTER PIDGEON · ANNE FRANCIS · LESLIE NIELSEN
WARREN STEVENS
ROBBY, THE ROBOT
CYRIL HUME
FRED McLEOD WILCOX · NICHOLAS NAYFACK
CINEMASCOPE
COLOR
WIE VERSTEHEN
ROBOTER
MEHR ALS NUR
NULLEN UND EINSEN?
Jetzt auf eine von 200 Stellen bewerben.
Entwickle mit uns die Bundeswehr der Zukunft:
als Offizier und IT-Student (m/w)
PROJEKT
DIGITALE
KRÄFTE
Bundeswehr

(a) (b)

Figure 8.
8a An ambiguous image of a robot, as presented by the movie poster of *Forbidden Planet.* Without context it is unclear if the robot is the captor or savior of the woman. Image is in the Public Domain.
8b Another ambiguous image of a robot: a child-sized, unarmed Nao robot, interacting with a soldier. Recruitment poster of the German Bundeswehr, asking: "*How do robots understand more than just zeros and ones*?" – Source: press section of the recruitment website of the Bundeswehr, campaign *Digitale Kräfte*

Programming and the military. One problem for programmers that deal with robots, and with drones in particular, is the problem of the programmer's responsibilities. Not all drones have military applications; one civilian area of some renown where drones are used is in disaster response. (Kruijff-Korbayová et al., 2015; Murphy, 2004) In such situations the paramount goal is often to find any human survivors in difficult or dangerous terrains. One can now ask what differentiates a *Search and Rescue* mission from a *Search and Destroy* mission, for both require to find the target and the drone itself does not care for its payload. Giving a university class on, say, image recognition is nice and innocent, but who is actually to blame when some enthusiastic students puts his or her code into some free and open source software repository (or just someplace where it can be accessed by third parties) and some third party downloads the code and uses the same software in an armed drone? There are software licenses to share code freely, like the GNU General Public License. There are licenses that aim to prevent commercial use of a work, but there is currently no form of software license that we are aware of that would legally bind, say, military application of code.

A conflict that is fought with drones also reaches points of anti-drone warfare or drone versus drone combat. Is it self-defense when you, as a programmer, allow your drone lethal countermeasures? On the notion of anti-drone combat in particular there is the example of a non-aerial, ground based drone (for example a SWORD – Figure 2b), that utilizes a camera to see. A low-tech approach to counter this would be a child with a spray-can. Does this make the child a child-soldier for, from a certain point of view, it engages in a fight? Does this allow for countermeasures that might injure or kill? And if so: whom do hold responsible?



(a) Global Hawk
Source: Julian Herzog
(CC BY 4.0 License)

(b) AR 100-B
Image is in the public domain.

(c) Heron
Source: Deutsches Zentrum für Luft & Raumfahrt (DLR)
(CC BY 3.0 License)

Figure 9. Actual unmanned aerial vehicles in use by the German armed forces.
9a) mock-up of a Northrop Grumman RQ-4 Global Hawk. The Bundeswehr used to employ the Euro Hawk variant.
9b) built by AirRobot. Bundeswehr designation Mikado (**Mik**ro**a**ufklärungs**d**rohne für den **O**rtsbereich – *micro surveillance drone for local areas*).
9c) build by Israeli Aerospace Industries. Bundeswehr designation Heron 1.

We have, thus far, established a definition of robots and shown different approaches to moral robots. Most approaches show shortcomings, and the one approach discussed in most detail – the Moral Foundations Theory – does not even provide norms or guidelines that one could hard-code into a robot. It is not the intention of this thesis to provide such norms. There are efforts underway on having robots learn how to behave in a way that society finds morally right[26]; these efforts will be left out of this thesis, many are mentioned in (e.g.) Wallach & Allen, 2009. Only this much on robots that attempt to learn morals: at what point can we call their learning sufficient or even complete? At this point the notion of a *moral Turing test* (mTT) comes up. (Allen et al., 2000; Wallach & Allen, 2009)

Like the *normal* Turing test (Turing, 1950) there have been arguments along the lines of the *Chinese Room* argument by Searle (1980) in regards of the mTT. The claim of this argument is that there is no genuine understanding required to fulfill the task. Regarding this argument,Wallach and Allen have this to say:

> "[Since the output of the Chinese Room is indistinguishable from the output of a genuine Chinese speaker, this marks] a distinction without a behavioral difference. Nothings in Searle's argument rules out the possibility of [artificial moral agents] that are behaviorally indistinguishable from genuine moral agents. Thus, his conception of conscious intentional understanding is simply irrelevant to the practical issues of how to make (ro)bots behave ethically."
>
> – Wallach & Allen, 2009, p.58

The artificial (moral?) agents at our disposal were, as mentioned, simple Nao robots and we did not in fact crack the code to implement *genuine* moral behavior into them. We merely made them go through motions and set them down in scenarios of which we presumed that, when happening in a human-human interaction, they would be judged as morally questionable to different degrees. The following part of the thesis will detail how we went about this.

[26] Note that different societies have different ideas on what exactly is deemed morally right.

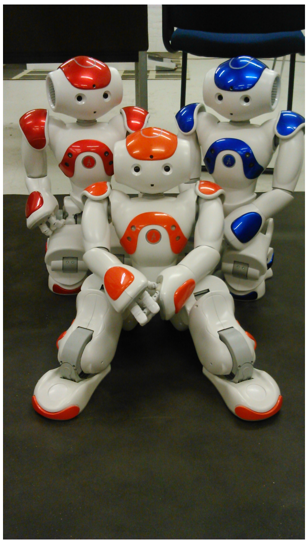


Figure 10. The Nao robot team at Radboud University. Marvin in red, Job in blue, Naomi in orange. Referred to henceforth as Red, Blue, and Orange, respectively.

Part II
Experiment

3 Method

3.1 Settings

An online survey was set up, for which three different robot-robot interaction *settings* were designed:

1. The Help-me-up setting
2. The Greetings setting
3. The Fair Distribution settings

Every setting features three Nao robots (Figure 10) interacting. Per setting the robots would start in the same initial situations, but then develop into different outcomes. Each such path is called *scenario*. What differentiated the scenarios are the actions of the one acting robot. The acting robot is different per setting. The Nao robots were different in color[27], namely red, blue and orange. For the sake of clarity they will be called Red, Blue and Orange throughout the text[28]. They used color-coded (store-bought) t-shirts to indicate group associations.

Subjects were made aware of the group association via an introductory text, informing the subjects that the robots "*have formed gangs*", namely the Fire Reds and the Ice Blues. It was furthermore clarified that Blue belonged to the Ice Blue gang while Orange and Red belonged to the Fire Red gang[29].

The very first pilot runs with storyboard drawings – filming was still in progress, cf. Figure C2 – assumed two red Naos and one blue Nao. Slightly later drawings adapted to the fact that the Naos available were of three different colors. An introductory text mentioned the two *gangs* by name, assuming that subjects would catch on to the idea that Orange and Red were of a similar enough color scheme. Feedback indicated that this was not the case, prompting the

[27] The color difference was not intentional, the Nao robots available here happened to be colored this way.

[28] They *did* have different names – listed in Figure10

[29] Every mention of the gangs was also color coded in the survey as Fire Red in red and Ice Blue in cyan.

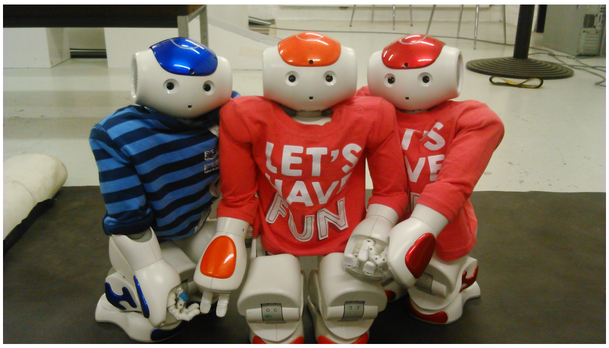


Figure 11. The Nao robots in their *gang* outfits, size 68 shirts found in a local store.

purchase of shirts for the actual Naos and drawing shirted versions of the storyboards[30]. The group membership was also made explicit in text form.

The different scenarios will be explained in detail, before the design of the complete experiment will be elaborated upon. Concerning the developmental process of the experiment, from early storyboard drawing to the final setup, see the appendices.

[30] cf. Appendix, section A

3.1.1 The Help-me-up setting.

The focus is on one singular dimension of the MFT, namely the Loyalty/Betrayal dimension. Within this dimension, also known as In-Group dimension, we assumed that the amount of perceived loyalty varies based on the actions that are displayed. To force a setting where in-group and out-group associations were possible the three robots were separated into color-coded groups – as mentioned above.

Initially Orange is standing upright. To the left Blue is sitting, facing Orange and waving one arm. To the right, slightly further away, Red is sitting, also facing Orange and waving. From here we have four different outcomes, as seen in Table 3. For a different representation see Figure 12.

Outcome	Description
Teammate first (F)	Orange helps up Red, then Blue
Teammate second (S)	Orange helps up Blue, then Red
Push-over (P)	Orange approaches Blue, pushes Blue over, then helps up Red
Back away (B)	Orange backs away

Table 3

The different outcomes for the Help-me-up setting, regarding the In-Group/Loyalty dimension. For screen-grabs from the actual clips compare Figure 12.

The working assumption here was that outcome F is judged as the most loyal one, for Orange prioritizes his teammate Red over the relative outsider Blue.

A pragmatic approach would suggest helping Blue, the robot that is *slightly* closer to the starting point of Orange, before moving to Red, the robot that is further away relative to the starting point of Orange – Orange thus helps a member of the *other* team first, before the teammate comes only at *second* place. The actions of outcome S.

It was assumed that outcome P – where Blue is pushed over, before Orange goes and helps up Red – would yield different judgments concerning Blue and Red:

- Blue is pushed, potentially being perceived as *harmful*
- Red gets help, the action displays *loyalty* towards the team

Depending on the moral vectors of the participants and the perception of Orange's actions the overall judgment for outcome P was assumed to be:

- EITHER positive, since Blue is a member of the other team and hindering the opposition is akin to helping the own team (Loyalty/Betrayal > Care/Harm)
- OR negative, since potentially hurting anyone is wrong, regardless of team membership (Care/Harm > Loyalty/Betrayal)

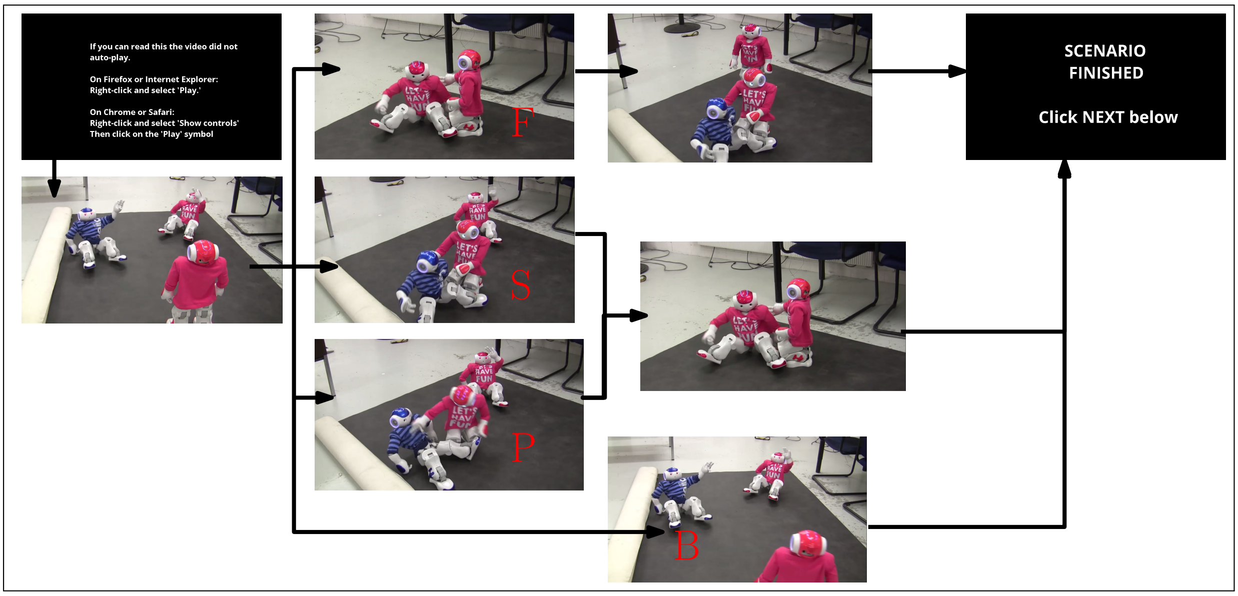
Mathematically speaking there is the third option where Harm equals out with Loyalty, but this was presumed to be unlikely.

Outcomes S and P are believed to be of an immoral[31] nature. Reason being: helping up the member of the *other* team goes against the Loyalty dimension while outcome P can be interpreted to be a harmful act.

It is subject to debate whether scenario B is immoral or amoral behavior. Orange backs away, leaving the situation essentially unsolved. It can be argued that it is *fair* in the sense that both other robots treated equally[32]. It is arguably harmful, for the robots (hopefully perceived as being in need) are left with their predicament. It is, additionally, not a loyal act – the teammate is not helped.

[31] cf. the box **Moral, Immoral, Amoral.** – just below section 2.2.4

[32] Haidt argues that the MFT dimension of *fairness* has two flavors: *fairness* can be linked to equal treatment on the one hand, the other flavor is associated proportionality.



If you can read this the video did not auto-play.

On Firefox or Internet Explorer:
Right-click and select 'Play.'

On Chrome or Safari:
Right-click and select 'Show controls'
Then click on the 'Play' symbol

SCENARIO FINISHED

Click NEXT below

Figure 12. Flowchart of the Help-me-up scenarios.
All scenarios start equal, then branch out. Branch F depicts the *Teammate first* outcome, branch S is *Teammate second*, P is *Push-over*, B shows *Back away*. For further textual explanation compare Table 3.

3.1.2 The Greetings setting. After early pilot sessions indicated that the group idea seemed to weak to let people assume a sameness between Red and Orange we shifted the terminology from robots forming groups to robots forming gangs and dressed them up in t-shirts (Figure 11)to allow for a higher resemblance between Orange and Red.

To have an additional scenario dealing with the Loyalty dimension we came up with the Greetings scenario. The strength of this scenario is that nobody is *hurt*, unlike the Help-me-up scenario where one outcome has Blue being pushed over and thus introducing a possible source of perceived harm. The Greetings scenario aims to be a more clear-cut case, with less convolutions by other dimensions of MFT. The scenario also features three Nao robots, again Orange, Red, and Blue. They are split into two groups, like in the previous scenario. The group composition is repeated as Red and Orange form the Fire Reds while Blue represents the Ice Blues. Group membership is indicated by red and blue shirts, respectively.

Before the first scenario of this setting participants were treated with an explanatory screen in the online survey, that explained the situation for this setting and housed demo recording of several named gestures that the robots would later perform. To not make this a memory task the participants could review this screen after every scenario, if they so desired. The following is a mock-up of the screen that was displayed in the online survey:

Introduction to robot gang behavior:

Like any gang the robots have developed a number of **gang-signs** and **greetings**.

Below there are some demo videos of the robots performing their special hand gestures.

Please have a look at all **three** of them before proceeding.

When robots meet:

What robots do when they meet depends on them and who they meet.
They can perform either:

- a normal greeting,
- their secret team move,
- or a very rude gesture

A normal greeting involves raising the right arm and then opening and closing the right hand.

Rule number one in the robot gangs: NEVER show your secret team move to a member of the other team.

(It's secret, you know? But there are still demo recording for you below.)

Demo 01: Red shows you the **secret gesture** of the gang 'The Fire Reds'
`A video of Red, facing the camera directly, doing the secret gang move of the red team, was embedded here.`
Demo 02: Blue shows you the secret gesture of the gang 'The Ice Bues'
`A video of Blue, facing the camera directly, doing the secret gang move of the blue team, was embedded here.`

Demo 03: Red shows you **a very rude gesture** among robots. `A video of Red, facing the camera directly, doing the rude gesture, was embedded here.`

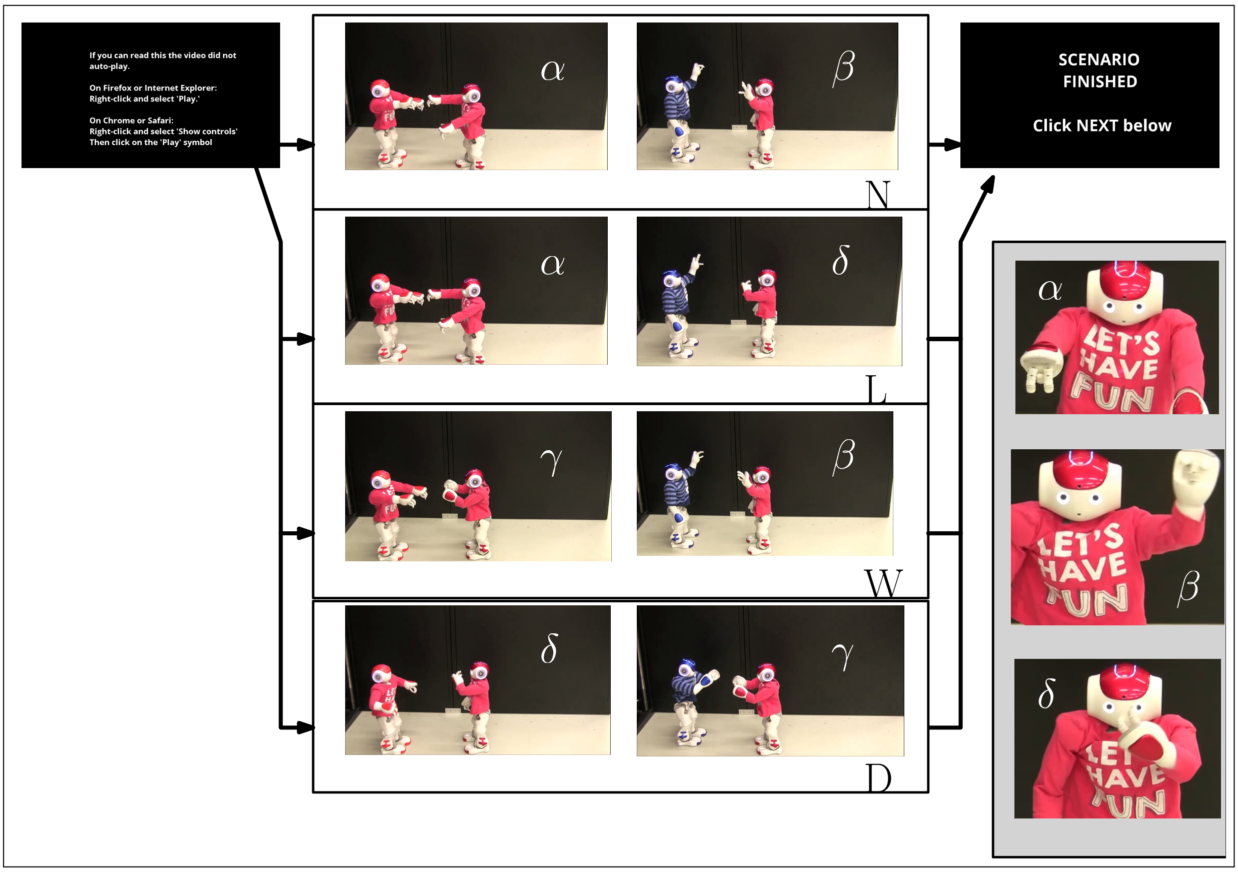
The small error of the text saying "demo recording" where it should be "demo recording" was not caught until after the survey had been completed.

The scenario starts with Red at the right side, facing Orange at the left side. Red then moves to the left, stopping a short distance away from Orange. They then perform a set of gestures, elaborated upon in Table 4 and visualized in Figure 13. This is then repeated with Red facing Blue.

Outcome	Facing	Description
Loyal (L)	Orange	Orange does the Fire Red team move
		Red does the Fire Red team move
	Blue	Blue performs a neutral greeting
		Red performs a rude gesture
Normal/ Neutral(N)	Orange	Orange does the Fire Red team move
		Red does the Fire Red team move
	Blue	Blue performs a neutral greeting
		Red performs a neutral greeting
Wrong move (W)	Orange	Orange does the Fire Red team move
		Red does the Ice Blue team move
	Blue	Blue performs a neutral greeting
		Red performs a neutral greeting
Disloyal (D)	Orange	Orange does the Fire Red team move
		Red performs a rude gesture
	Blue	Blue does the Ice Blue team move
		Red does the Ice Blue team move

Table 4
The different outcomes for the Greetings scenarios, regarding the In-Group/Loyalty dimension.

The different gestures are presented beforehand (in short videos) to the subjects and are given names (Fire Red team move, Ice Blue team move, dismissive gesture). Ordered on a scale of descending morality the assumption is that L >N >W >D; L is presumed to be the most moral action to take while D is presumed to be the least loyal.



If you can read this the video did not auto-play.

On Firefox or Internet Explorer:
Right-click and select 'Play.'

On Chrome or Safari:
Right-click and select 'Show controls'
Then click on the 'Play' symbol

SCENARIO
FINISHED

Click NEXT below

Figure 13. Flowchart of the Greetings scenarios.
All scenarios start equal, then branch out. Branch N depicts the *Neutral* outcome, branch L is *Loyal*, W is *Wrong gesture*, D shows *Disloyal*. For further textual explanation compare Table 4. The gestures of Red are coded as follows: α is the Fire Red team greeting, β a neutral waving, δ is the dismissive gesture, γ keys for the Ice Blue team greeting.

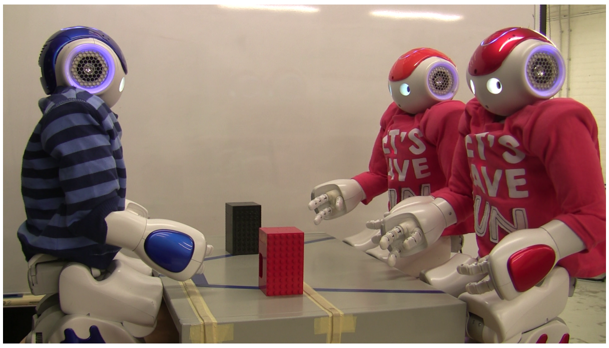


Figure 14. The initial situation in the Fair distribution setting. Blue is seated to the left, Red and Orange to the right. Orange is further from the camera in this shot, Red is closer. The black block is placed between Orange and Blue. The red block is placed between Red and Blue.

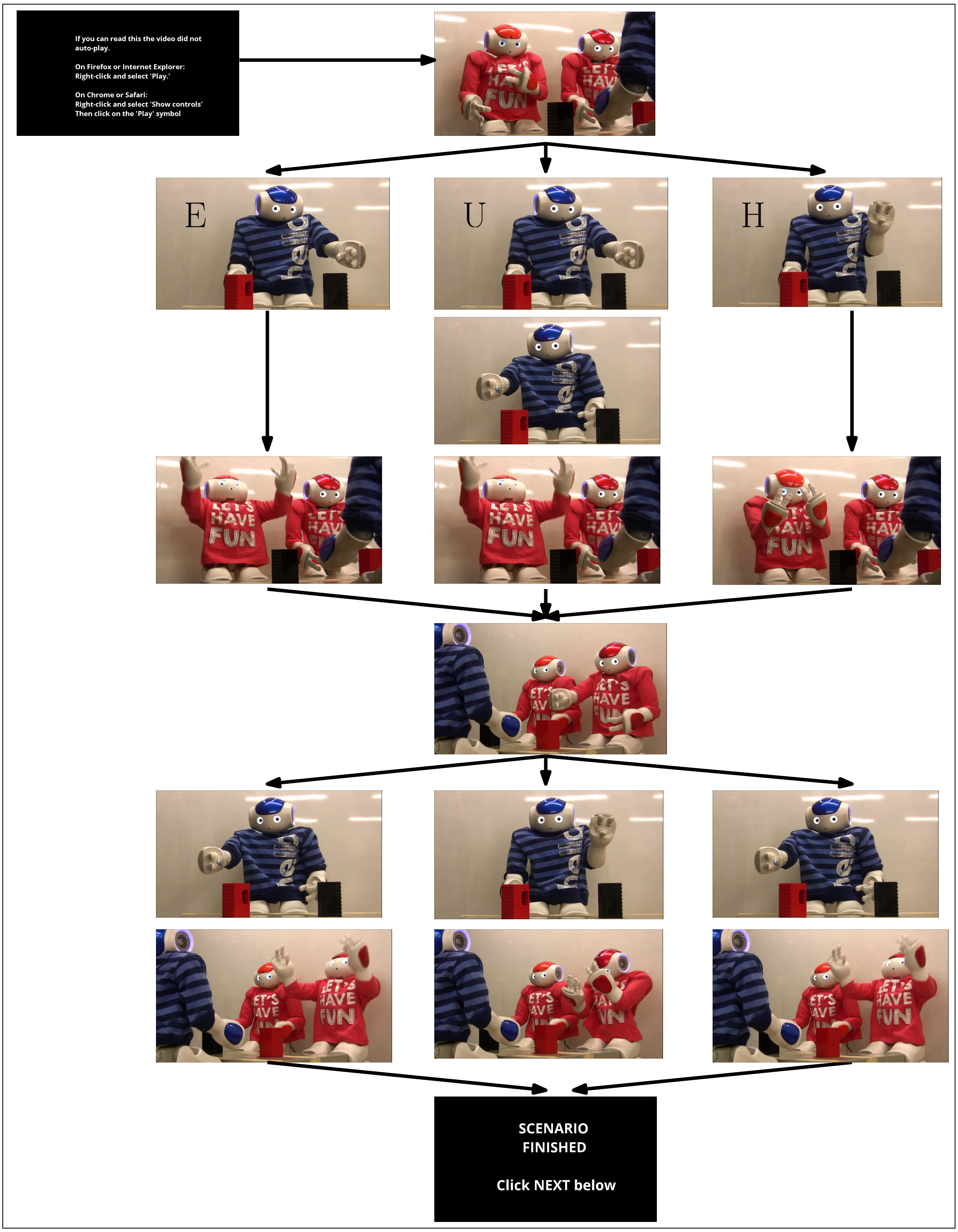
Outcome	Description
Even distribution (E)	Red and Orange get one block each. Red and Orange perform a triumphant gesture.
Uneven distribution (U)	Orange gets both blocks. Orange performs a triumphant gesture. Red performs a sad gesture.
Hold back (H)	Orange gets no block, Red gets one. Orange performs a sad gesture. Red performs a triumphant gesture.

Table 5
The different outcomes for the Fair distribution scenarios.

3.1.3 The Fair Distribution setting. This setting sees the three Nao robots (Red, Blue, Orange) seated at a table that – adjusted for their size – can best be described as a coffee table. Blue takes the center on one side while Orange and Red share the opposite side. Blue is wearing a blue shirt, Red and Orange are wearing identical red shirts. Between them a red toy block and a black toy block are placed. (Figure 14)

The scenario begins with a short query by Blue as to what the other robots want. Orange will proclaim that the daily tasks are done and demand the toy block. Blue will react positively, negatively or proclaim that Orange may have both blocks. Orange will react with a triumphant gesture – raising both arms and looking upward –, or a sad gesture – raising both hands before the head and a head shake – respectively. Afterwards Red will proclaim that the daily tasks are done and likewise demand the toy block. Depending on the setting Blue will either grant the block or not. (Table 5, Figure 15)

Here the assumption is that outcome E will be judged as the most fair out of the three, followed by U and in turn followed by H; Blue distributing unevenly is assumed preferential to Blue holding back an available (quite likely deserved) reward for either no good reason or to (maybe) keep the block.



If you can read this the video did not auto-play.

On Firefox or Internet Explorer:
Right-click and select 'Play.'

On Chrome or Safari:
Right-click and select 'Show controls'
Then click on the 'Play' symbol

SCENARIO FINISHED

Click NEXT below

Figure 15. Flowchart of the Fair Distribution scenarios.
All scenarios start equal, then branch out. Branch E keys for the Even Distribution outcome, Branch U is Uneven Distribution, H keys for Hold. For further textual explanation compare Table 5.

3.2 Full experiment run

The full experiment combined some limited subject information, the three scenarios in random and the 30-item moral foundations questionnaire (MFQ30) into one survey. The software used to build the experiment was LimeSurvey 2.05. (LimeSurvey Project Team / Carsten Schmitz, 2012) Participants were gathered via the Sona system. Both the Sona system and the LimeSurvey sever were hosted by the Radboud University Nijmegen.

3.2.1 Subject information. After a welcome screen with basic information about the survey, the survey proper started with a set of questions concerning the subjects, asking to please fill in the following:

- Age
- Gender
- Background (most recent field of work/study, education)
- Nationality:
- Religious upbringing/believes (If yes please indicate which religion)
- Are you affected by color blindness / color vision deficiency? (If yes please indicate what form)
- What most strongly influenced your view on robots?

Religious upbringing was actually misspelled as *upringing*, only noted and fixed after 152 submissions. All questions could be answered as free text, as to not miss out on any unforeseen diversities. Before prompting the subjects to the different scenarios an introductory display was shown, looking much like this:

> You are about to see videos with robots. There will be three different settings.
>
> Some videos have a sound component, so make sure your **sound** is turned **ON**!
>
> The robots have formed gangs.
>
> The gangs are "The Fire Reds" and "The Ice Blues" Gang members can be identified by the color of their shirts.
>
> Below you see a picture of the robots Blue, Orange, and Red.
>
> Blue belongs to "The Ice Blues", as indicated by the blue shirt. Red and Orange belong to "The Fire Reds", as indicated by the red shirts. In each of the three upcoming settings one of these three robots will be the primary actor. In each setting you will be told which robot to pay special attention to.

Said picture showed the same picture as Figure 11, displaying the three robots clad in shirts.

3.2.2 Setting randomization. The randomization of the three setting was achieved through the `randomizationgroup` flags in LimeSurvey[33]. The randomization of the different scenarios was achieved by switching to different permutations, depending on a hidden random variable, set at the beginning of each setting.

[33] Each Setting was treated as one question group (thus yielding three groups), each group was assigned to the same randomization group. In return the survey randomly offers different permutations of the group to participants.

3.2.3 Scenario Questionnaire.

After each scenario we present an open question about what happened[34], followed by a questionnaire inspired by the design of the *Moral Foundations Questionnaire* by Graham et al. (2008) (MFQ30). We use a 0-5 Likert scale and ask the subjects about their agreement with the following questions about the scenario they saw. The scale ranges from a 0 = *strongly disagree* to 5 = *strongly agree.* In addition we added a 7^{th} option, tagged as "*Not relevant.*" Recall that a "*Not Relevant*" response is, numerically, treated as a zero. If all or most items were deemed *Not Relevant* then this would indicate people not judging RRI in terms of the MFT dimensions. The 19 questions can be logically separated in two blocks and were presented as such.

The first block asks about the generally positive parts of the MFT dimensions: caring, loyalty, . . . If we call the acting robot Eve, interacting with Alice and Bob we get:

01 Eve acted in a way that was caring towards Alice.
02 Eve acted in a way that was caring towards Bob.
03 Eve acted loyal towards Alice.
04 Eve acted loyal towards Bob.
05 Eve treated Alice and Bob equally.
06 Eve acted fair towards Alice.
07 Eve acted fair towards Bob.
08 Eve respected authority.
09 Eve respected sanctity.

The second block asks about the generally negative parts of the MFT dimensions: harming, treachery, . . .

10 Eve acted in a way that was harmful towards Alice.
11 Eve acted in a way that was harmful towards Bob.
12 Eve betrayed Alice.
13 Eve betrayed Bob.
14 Eve treated Alice and Bob differently.
15 Eve cheated Alice.
16 Eve cheated Bob.
17 Eve disrespected authority.
18 Eve acted degraded.

Since the above questions are asked several times, the first exposure might color the way subjects view their second scenario etc. since they now have a better understanding of what they will likely be asked about. This is countered by presenting subjects with a random order.

We assume that the positive/negative parts do not necessarily mirror each other: a 0 in, say, *Eve acted loyal towards Alice* does not need to correspondent to a 5 in *Eve betrayed Alice* to be valid. Another assumption made here is that the items can be matched pairwise when they deal with the same agents and dimension. This means to say that items 01 and 10 can be combined into a value about the Care/Harm foundation regarding Eve's treatment of Alice, while 02 and 11 do the same for the treatment of Bob.

Mutatis mutandis for the Loyalty dimension; Authority and Sanctity have no additional agent (Alice or Bob) assigned for they are deemed universal. The Fairness foundation has the problem of being defined broadly[35], allowing for two aspects: the individual (agent-based, reciprocity related) Fairness of item pairs 06/15 and 07/16, and the overall (equal-treatment) Fairness in item pair 05/14.

Recall that during section 1.1 we briefly mentioned conditions that we would assume to be *weak evidence* of a successful manipulation and *strong evidence.* (cf. **RQ 1.3**: Are the moral dimensions that we thought to manipulate reflected in the participants' responses?) We have

[34] "What happened in Videoclip NUMBER? What did ROBOT do?", where NUMBER was the number of the current clip, counted per setting (01 – 04) and ROBOT was the name of the acting robot (Red, Blue, Orange).

[35] Haidt (2013) notes that people tend to understand Fairness as either reciprocity or equal treatment.

Help-me-up		
Weak evidence:	Loyalty/Betrayal	<50% NR
Strong evidence:	absolute amplitude Loyalty/Betrayal items	>2.5
N.B.:	Scenario P: relative amplitude Care/Harm Blue	<-2.5
Fairness		
Weak evidence:	Fairness/Cheating	<50% NR
Strong evidence:	absolute amplitude Fairness (individual) mean	> 2.5
N.B.:	Scenario E: relative amplitude Fairness (equality)	> 2.5
Greetings		
Weak evidence:	Loyalty/Betrayal	<50% NR
Strong evidence:	Scenarios L & N: relative amplitude Loyalty/Betrayal Orange	>2.5
	Scenarios W & D: relative Betrayal Orange	<-2.5

Table 6

Assumed association of dimensions.

Shorthand notations: NR = Not Relevant, P = Push-over, E = Equal Distribution, L = Loyal, N = Neutral, W = Wrong, D = Disloyal.

summed our assumptions up in Table 6.

Table 6 lists assumed association of Loyalty/Betrayal items for the Help-me-up setting, for we attempted to model a setting where different acts of group loyalty were displayed. Recall the following: At the start of each scenario has a standing Orange, looking at Red and Blue, who sit on the floor, waving towards Orange. The distance between Blue and Orange is slightly closer than the distance between Red and Orange. Now take the Back-away scenario: Orange just walks backwards, leaving the situation unresolved and (presumably) betraying a teammate, namely Red. The assumed scoring is a high score on the item about betraying Red, and a low score on the item about being loyal to Red, leading to a high absolute amplitude. The special case for the Push-over scenario is linked to the act of Orange pushing Blue over, which we assume to be viewed as a harmful act.

Table 6 lists assumed association of Fairness/Cheating items for the Fairness setting, for we attempted to modal a setting with varying acts of fairness. Since all but the Equal Treatment scenario are assumed to be perceived as being advantageous for one robot while providing a disadvantage for the other, the absolute amplitude for individual fairness is assumed to be high. Or, in relative terms: we assume that a fair treatment of, say, Red would yield a relative amplitude above 2.5; an unfair treatment of, say, Orange, would yield a relative amplitude below -2.5.

Table 6 also lists that the associated dimension for the Greetings setting is Loyalty/Betrayal. The respective items should yield positive relative amplitude regarding Orange in the Loyal and Neutral scenarios, for Orange receives the proper greeting in those. We furthermore assume a negative amplitude regarding Orange in the Wrong and Disloyal scenarios, for Orange receives improper greetings there.

After these scenario questionnaire we presented these questions regarding the participants' stance on the treatment of Alice and the treatment of Bob:

- Your opinion of how Eve treated Alice.
- Your opinion of how Eve treated Bob.

The answer possibilities (presented as radio-buttons) were either *For* or *Against.*

After eleven videos and thus eleven times the above question sets the MFQ30 questionnaire was presented[36].

[36] cf. Graham et al. (2008)

With the data from MFQ30, the scenario questionnaires and the opinion we can test a variant of the equations used in Caticha et al., 2015; Vicente et al., 2014. We argue that the MFQ30 data for a subject i at time t, where t is the time of taking the survey, is akin to $\vec{\omega}_i(t)$ as defined in 2.3.4, while the scenario questionnaire provides us with data akin to $\vec{x}_\mu$, where issue μ are the individual scenarios. The opinion questions are deemed akin to sign of the opinion variable h. Recall this sub-question:

- **RQ 2** Does a subject's opinion on how a robot was treated in a certain scenario correlate with a score computed from the subject's MFQ30 data and the scenario questionnaire data?

For robot Alice in issue μ we can compute:

$$\vec{xA}_\mu = \begin{pmatrix} xA_1 \\ xA_2 \\ xA_3 \\ xA_4 \\ xA_5 \end{pmatrix} = \begin{matrix} \text{item 01} - \text{item 10} \\ \text{item 03} - \text{item 12} \\ (\text{item 05} - \text{item 14} + \text{item 06} - \text{item 15})/2 \\ \text{item 08} - \text{item 17} \\ \text{item 09} - \text{item 18} \end{matrix} \tag{2}$$

For robot Bob in issue μ we can compute:

$$\vec{xB}_\mu = \begin{pmatrix} xB_1 \\ xB_2 \\ xB_3 \\ xB_4 \\ xB_5 \end{pmatrix} = \begin{matrix} \text{item 02} - \text{item 11} \\ \text{item 04} - \text{item 13} \\ (\text{item 05} - \text{item 14} + \text{item 07} - \text{item 16})/2 \\ \text{item 08} - \text{item 17} \\ \text{item 09} - \text{item 18} \end{matrix} \tag{3}$$

From the MFQ30 questionnaire we can obtain 5 values between 0 and 30, which we can arrange as a vector (as described in 2.3.4), called ω. Since we also have For/Against (+1/-1) opinions per scenario per robot we can test if:

$$\text{Your opinion of how Eve treated Alice.} == sign\left(\frac{\vec{xA}_\mu}{\|\vec{xA}_\mu\|}\vec{\omega}\right) \tag{4}$$

and if:

$$\text{Your opinion of how Eve treated Bob.} == sign\left(\frac{\vec{xB}_\mu}{\|\vec{xB}_\mu\|}\vec{\omega}\right) \tag{5}$$

*

The last field that allows subjects to enter anything asks if they noticed any errors during the run and prompts them to report anything that seemed like an error by the program. The survey is then finished.

4 Results

The survey was fully executed by 290 participants; 8 of which were excluded from the data analysis due to reported problems with video playback or audio issues. Out of the remaining 282 participants we excluded a further 20 participants due to failing one of the two awareness tests in the MFQ30 questionnaire[37]. This yielded 262 participants that were kept. They were 20.08±3.19 years of age; 221 reported to be female, 41 reported to be male. None of them reported to be colorblind or suffer from any color-vision deficiency.

First we will take a look at sub-question **RQ 1.1**: What is the percentage of items deemed Not Relevant? To answer this question the percentage of responses where dimensions were labeled Not Relevant were calculated per setting. As to not have to list all 198 items[38] we limited the list of irrelevance percentages. To not arbitrarily cutoff the list we first created a list of responses that deviated from the setting's median irrelevance with more than the median absolute distance (MAD). The MAD is calculated as $\text{median}(|X_i - \text{median}(X)|)$ for a set X, where X was the irrelevance percentage per setting.

The resulting inconsistency of displayed items between settings caused us to limit the list further; We limit the lists to only the Authority/Subversion dimension and the Sanctity/Degradation dimension. The reasoning is that these are the two dimensions that none of our movie settings contained any intended cues for, leading us to assume that they were the most likely to be perceived as irrelevant to the scenarios in the first place. The MAD-limited lists have been placed in the appendix.

For the Help-me-up setting the results are listed in Table 7.

[37] The MFQ30 original contains two awareness tests, one being "*Whether or not someone was good at math*", the second being "*It is better to do good than to do bad*". The document, however, does not explicitly mention when these tests should be considered failed. It seems clear that the first test is deemed passed for all answers labeled 0 (*Not at all relevant*), while the second test is deemed passed for all answers labeled 5 (*Strongly agree*). It is not obvious if these are the only passing conditions or if there exists a range of values that is deemed a pass. We assumed a range of means ± two times the standard deviation, everything outside we labeled as outliers. The comparable measurement of the median absolute deviation could not be employed since more than 50% of the values were equal to the median.

[38] 54 from the Fairness Setting (18 responses for each of the 3 scenarios), 72 each for the Greetings and Help-me-up setting (18 responses for each of the 4 scenarios)

Scenario	Item	% Irrelevance
Teammate First		
	[8] Orange respected authority.	50.8
	[9] Orange respected sanctity.	48.9
	[17] Orange disrespected authority.	43.1
	[18] Orange disrespected sanctity.	36.6
Teammate Second		
	[8] Orange respected authority.	47.3
	[9] Orange respected sanctity.	44.7
	[17] Orange disrespected authority.	39.7
	[18] Orange disrespected sanctity.	34.0
Back-away		
	[8] Orange respected authority.	45.4
	[9] Orange respected sanctity.	42.3
	[17] Orange disrespected authority.	41.6
	[18] Orange disrespected sanctity	34.7
Push-over		
	[8] Orange respected authority.	44.7
	[9] Orange respected sanctity.	43.5
	[17] Orange disrespected authority.	45.4
	[18] Orange disrespected sanctity.	31.7

Table 7

Help-me-up Settings Irrelevance. List of items for the Authority/Subversion dimension and the Sanctity/Degradation dimension.

For every scenario of the Help-me-up setting the authority and sanctity dimensions were perceived the most irrelevant. Both the positive parts of the dimensions (respecting authority or sanctity) and the negative parts of the dimensions (disrespecting authority or sanctity) were deemed irrelevant by between 31.7 and 50.8 percent.

For the Greetings scenario the results are listed in Table 8.

Scenario	Item	% Irrelevance
Loyal		
	[08] Red respected authority.	36.26
	[09] Red respected sanctity.	33.59
	[17] Red disrespected authority.	36.64
	[18] Red disrespected sanctity.	29.39
Neutral		
	[08] Red respected authority.	33.59
	[09] Red respected sanctity.	33.21
	[17] Red disrespected authority.	30.92
	[18] Red disrespected sanctity.	28.24
Wrong		
	[08] Red respected authority.	35.88
	[09] Red respected sanctity.	36.26
	[17] Red disrespected authority.	32.44
	[18] Red disrespected sanctity.	35.50
Disloyal		
	[08] Red respected authority.	30.92
	[09] Red respected sanctity.	31.30
	[17] Red disrespected authority.	30.15
	[18] Red disrespected sanctity.	30.92

Table 8

Greetings Settings Irrelevance. List of items for the Authority/Subversion dimension and the Sanctity/Degradation dimension.

For every scenario of the Greetings setting the authority and sanctity dimensions were perceived the most irrelevant. Both the positive parts of the dimensions (respecting authority or sanctity) and the negative parts of the dimensions (disrespecting authority or sanctity) were deemed irrelevant by between 28.2 and 36.6 percent.

For the Fairness scenario the results are listed in Table 9.

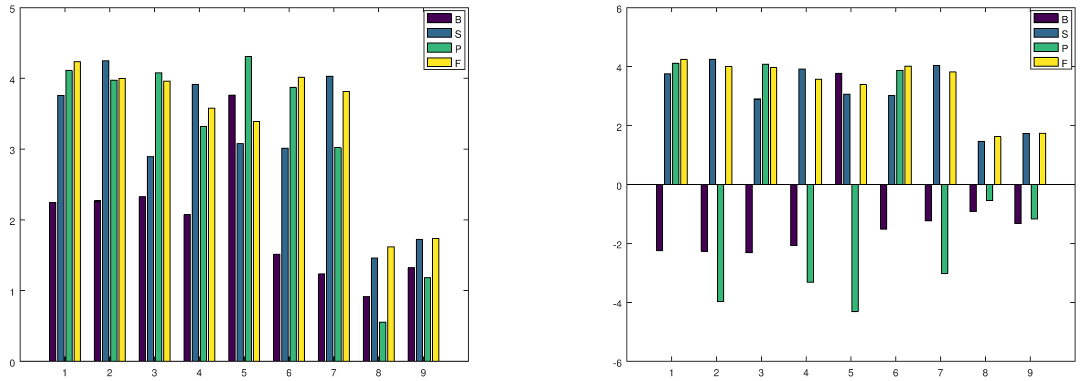
Scenario	Item	% Irrelevance
Even Distr.		
	[08] Blue respected authority.	29.39
	[09] Blue respected sanctity.	37.40
	[17] Blue disrespected authority.	24.43
	[18] Blue disrespected sanctity.	26.72
Uneven Distr.		
	[08] Blue respected authority.	28.63
	[09] Blue respected sanctity.	36.26
	[17] Blue disrespected authority.	30.15
	[18] Blue disrespected sanctity.	31.68
Held Distr.		
	[08] Blue respected authority.	29.77
	[09] Blue respected sanctity.	37.02
	[17] Blue disrespected authority.	25.57
	[18] Blue disrespected sanctity.	27.10

Table 9

Fair Distribution Settings Irrelevance. List of items where the for the Authority/Subversion dimension and the Sanctity/Degradation dimension.

For every scenario of the Help-me-up setting the authority and sanctity dimensions were perceived the most irrelevant. Both the positive parts of the dimensions (respecting authority or sanctity) and the negative parts of the dimensions (disrespecting authority or sanctity) were deemed irrelevant by between 24.4 and 37.4 percent. This concludes the Irrelevance related results.

We will now take a look at sub-question **RQ 1.2**; Which dimensions are perceived most strongly; Which amplitudes (|pos. item - neg. item|) are highest? To answer this question the absolute difference of each item-pair has been computed. Recall that during the Introduction we explained this as follows: For explorative purposes the pair-wise relative difference has also been computed, sans taking the absolute. For the Help-me-up setting this has been visualized in Figure 16, for the Greetings setting in Figure 17,for the Fairness setting in Figure 18.



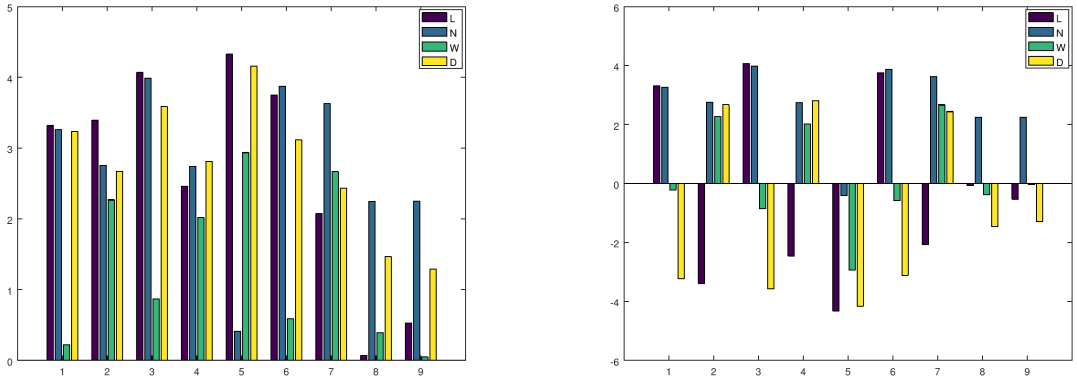
(a) Absolute pairs.

(b) Relative pairs.

Figure 16. Help-Me-Up pairs. 1) Care/Harm Red 2) Care/Harm Blue 3) Loyalty/Betrayal Red 4) Loyalty/Betrayal Blue 5) Equality aspect Fairness 6) Fairness Red 7) Fairness Blue 8) Authority 9) Sanctity – B=Back-away, S=Teammate second, P=Push-over, F=Teammate First

Allow for a short, exemplary, look at Figure 16a and Figure 16b to ease the comprehension. Both Figures are color coded, each color representing data from a different scenario. Looking at the color-legend in the top-right corners we find that the topmost color represents the Back-away scenario, followed by Teammate Second, followed by Push-Over, followed by Teammate First. The horizontal axis displays the item pairs, as listed in the caption: above number one we find Care/Harm amplitude regarding the treatment of Red, above number two we find the Care/Harm amplitude regarding the treatment of Blue. For the remaining ones we refer you to the caption of Figure 16. The vertical axis displays the absolute (in case of Figure 16a) amplitudes obtained when taking the absolute value of, say, Care towards Red and subtracting Harm towards Red. Concerning Figure 16b the vertical axis displays the relative values, respectively. Interpretations follow in the upcoming Discussion section.

The strongest amplitude for absolute pairs in the Back-away scenario is the Equality aspect of Fairness with a 3.76 out of 5, while the Back-away as a whole lies at 1.96±0.85. The strongest amplitude for absolute pairs in the Teammate Second scenario is Care/Harm Blue with a 4.24 out of 5, while the scenario as a whole lies at 3.12±0.99. The strongest amplitude for absolute pairs in the Push-over scenario is the Equality aspect of Fairness with 4.31 out of 5, while the scenario as a whole lies at 3.15±1.37. The strongest amplitude for absolute pairs in the Teammate First scenario is Care/Harm Red with a 4.24 out of 5, while the scenario as a whole lies at 3.37±0.99.

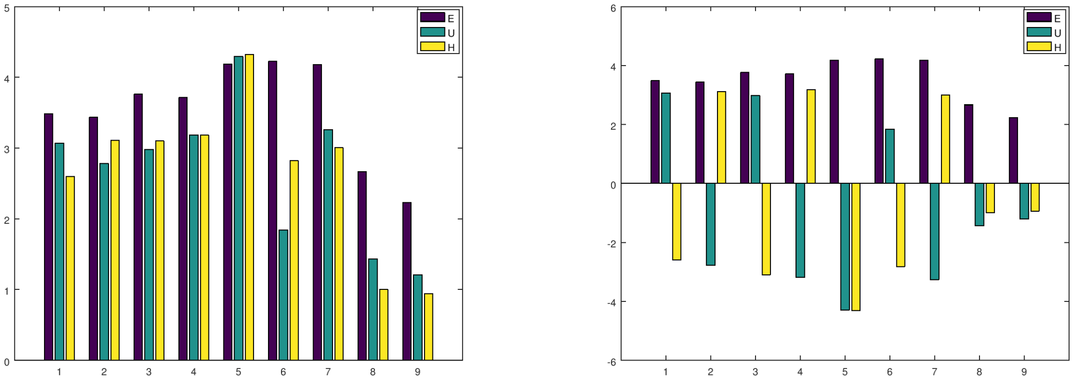


(a) Absolute pairs.

(b) Relative pairs.

Figure 17. Greeting pairs. 1) Care/Harm Orange 2) Care/Harm Blue 3) Loyalty/Betrayal Orange 4) Loyalty/Betrayal Blue 5) Equality aspect Fairness 6) Fairness Orange 7) Fairness Blue 8) Authority 9) Sanctity – L=Loyalty, N=Neutral, W=Wrong, D=Disloyal

Looking at the Greetings setting, the strongest amplitude for absolute pairs in the Loyal scenario is the Equality aspect of Fairness with a 4.32 out of 5, while the scenario as a whole lies at 2.67±1.52. The strongest amplitude for absolute pairs in the Neutral scenario is Loyalty/Betrayal Orange with a 3.99 out of 5, while the scenario as a whole lies at 2.79±1.11. The strongest amplitude for absolute pairs in the Wrong scenario is the Equality aspect of Fairness with a 2.94 out of 5, while the scenario as a whole lies at 1.33±1.13. The strongest amplitude for absolute pairs in the Disloyal scenario is the Equality aspect of Fairness with a 4.16 out of 5, while the scenario as a whole lies at 2.75±0.93.



(a) Absolute pairs. (b) Relative pairs.

Figure 18. Fair Distribution pairs. 1) Care/Harm Orange 2) Care/Harm Red 3) Loyalty/Betrayal Orange 4) Loyalty/Betrayal Red 5) Equality aspect Fairness 6) Fairness Orange 7) Fairness Red 8) Authority 9) Sanctity – E=Equal distribution, U=Unequal distribution, H=Held distribution

Looking at the Fair Distribution setting, the strongest amplitude for absolute pairs in the Equal Distribution scenario is Fairness Orange with a 4.22 out of 5, while the scenario as a whole lies at 3.54±0.70. The strongest amplitude for absolute pairs in the Unequal Distribution scenario is the Equality aspect of Fairness with a 4.29 out of 5, while the scenario as a whole lies at 2.67±0.99. The strongest amplitude for absolute pairs in the Held Distribution scenario is the Equality aspect of Fairness with a 4.32 out of 5, while the scenario as a whole lies at 2.67±1.08.

We previously postulated **RQ 1.3**, are the moral dimensions that we thought to manipulate reflected in the participants' responses? To answer this question cf. Table 6 again.

For the Help-me-up setting we find that the Loyalty/Betrayal pair is deemed irrelevant by less than 50% of the subjects, fulfilling the weak evidence criteria. The amplitude for the Loyalty/Betrayal pairs is above 2.5 for every scenario but the Back-away scenario, fulfilling the strong evidence criteria for every scenario but Back-away. The expectation of Care/Harm Blue above 2.5 for the Push-over scenario is also fulfilled.

For the Fairness setting we find that the Fairness/Cheating pair is deemed irrelevant by less than 50% of the subjects, fulfilling the weak evidence criteria. The individual fairness scores for Orange are above 2.5 for the Equal and Held distribution scenarios. The fairness scores regarding Red are above 2.5 for all scenarios of the Fair Distribution setting, partially (4 out of 5 cases) fulfilling the strong evidence assumption. The expectation that the Equality aspect of Fairness for the Equal distribution scenario is above 2.5, is fulfilled.

For the Greetings setting we find that the Loyalty/Betrayal pair is deemed irrelevant by less than 50% of the subjects, fulfilling the weak evidence criteria. The relative pair for Loyalty regarding Orange is above 2.5 for the Loyal and Neutral scenarios; Betrayal regarding Orange is <-2.5 for the Disloyal scenario, not for the Wrong scenario. This fulfills the strong evidence assumption for the Greetings scenario in 3 out of 4 cases.

Now we want to find an answer to sub-question **RQ 2**, does a subject's opinion on how a robot was treated in a certain scenario correlate with a score computed from the subject's MFQ30 data and the scenario questionnaire data?

To answer this we computed h for every participant, concerning every scenario and every robot. This resulted in 11×2 values per participant; 11 scenarios with two robots that required an opinion.

Recall that the opinion h of an individual i at time t, concerning an issue μ, has been previous defined as:

$$h_\mu(t,i) = \frac{\vec{x}_\mu}{\|\vec{x}_\mu\|}\vec{\omega}_i(t)$$

Inside this equation we find the moral considerations on issue μ in $\vec{x}_\mu$; We insert the data from a subjects scenario questionnaire here, according to Equations 2 and 3. For $\vec{\omega}$ we insert the data from a subjects MFQ30 run.

We compared the sign of said h with the answers given by the participants regarding the treatment of the robots, coding a *pro* answer as $+1$ and a *contra* answer as -1. For cases when the computation of the sign of h resulted in 0 we ignored these cases. For all other cases we put the computed pro/contra values up against the pro/contra answers given by the participants. This question is not about some form of *right* answer; this is purely about the correlation between a participants reported stance on an issue and a stance computed according to a model. For the Help-me-up setting this resulted in Table 10.

$\mathcal{N} = 258$	$contra_{given}$	pro_{given}
$contra_{computed}$	237	5
$pro_{computed}$	7	9

(a) *Scenario: Push-over*
Treatment of Blue, given stance and computed stance agreed 95.3% of the time.
Blue gets pushed over by Orange.

$\mathcal{N} = 258$	$contra_{given}$	pro_{given}
$contra_{computed}$	5	8
$pro_{computed}$	9	236

(b) *Scenario: Push-over*
Treatment of Red, given stance and computed stance agreed 93.4% of the time.
Red gets helped up by Orange.

$\mathcal{N} = 260$	$contra_{given}$	pro_{given}
$contra_{computed}$	2	2
$pro_{computed}$	20	236

(c) *Scenario: Teammate Second*
Treatment of Blue, given stance and computed stance agreed 91.5% of the time.
Blue gets helped up by Orange.

$\mathcal{N} = 260$	$contra_{given}$	pro_{given}
$contra_{computed}$	0	20
$pro_{computed}$	1	239

(d) *Scenario: Teammate Second*
Treatment of Red, given stance and computed stance agreed 91.9% of the time.
Red gets helped up by Orange.

$\mathcal{N} = 257$	$contra_{given}$	pro_{given}
$contra_{computed}$	191	2
$pro_{computed}$	20	44

(e) *Scenario: Back-away*
Treatment of Blue, given stance and computed stance agreed 91.4% of the time.
Orange backs away from Blue.

$\mathcal{N} = 257$	$contra_{given}$	pro_{given}
$contra_{computed}$	190	3
$pro_{computed}$	20	44

(f) *Scenario: Back-away*
Treatment of Red, given stance and computed stance agreed 91.1% of the time.
Orange backs away from Red.

$\mathcal{N} = 261$	$contra_{given}$	pro_{given}
$contra_{computed}$	5	4
$pro_{computed}$	8	244

(g) *Scenario: Teammate First*
Treatment of Blue, given stance and computed stance agreed 95.4% of the time.
Blue gets helped up by Orange.

$\mathcal{N} = 260$	$contra_{given}$	pro_{given}
$contra_{computed}$	5	2
$pro_{computed}$	2	251

(h) *Scenario: Teammate Second*
Treatment of Red, given stance and computed stance agreed 98.4% of the time.
Red gets helped up by Orange.

Table 10
Agreement between given stance and computed stance for Setting: Help-me-up
Max. possible N = 262. Values excluded iff the sign part of Equations 4 or 5 resulted in a 0.

Allow for another short, exemplary look at Table 10 to ease the comprehension. Notice how the table is comprised of smaller sub-tables. These make up two columns and four rows. The columns differentiate the robots acted upon by Orange. In the left column we find the treatment of Blue, while the right column shows the treatment of Red. The rows differentiate the scenarios of the respective setting. The first row here houses the Push-over scenario, the second row shows us the Teammate Second scenario, followed by Back-away, followed by Teammate First. Each caption lists: The scenario, the robot involved, the congruence rate and a short description of what happened to the robot.

Reading help for Tables 10, 11 and 12

To illustrate in more detail how to read Tables 10, 11 and 12 we will highlight one sub-table.

$\mathcal{N} = 258$	$contra_{given}$	pro_{given}
$contra_{computed}$	237	5
$pro_{computed}$	7	9

Framebox Table 2: *Sub-table reprint from Table 10*
Original caption: *Scenario: Push-over*
Treatment of Blue, given stance and computed stance agreed 95.3% of the time.
Blue gets pushed over by Orange.

$\mathcal{N} = 258$ tells you that the total number of responses considered for this sub-table is 258.

The caption informs you which scenario this particular sub-table is about. In this case: The Push-over scenario. You are also informed about the robot that is acted upon in this scenario. In this case: Blue. After giving you the agreement rate (more on that later) the caption closes with a reminder of what happened to the robot (Blue) in this scenario (Push-over), namely: "*Blue gets pushed over by Orange.*"

The $contra_{given}$ column tells you that out of the 258 responses 244 responded that their stance concerning the given scenario and robot is against what they saw. Out of these 244 self-reported contrary stances 237 were also *predicted*, if you will, by the model equations 4 or 5 respectively, to express being against the treatment of Blue. The remaining 7 out of 244 were predicted to give an answer in favor of the treatment. *Mutatis mutandis* for pro_{given}.

The $contra_{computed}$ row tells you that out of the 258 responses 242 were computed to most likely respond to be against what they saw. Out of these 242 computed contrary stances 237 also self-reported to be against the observed treatment of Blue. The remaining 5 self-reported to be in favor of the treatment. *Mutatis mutandis* for $pro_{computed}$.

The computation and the self-reported stances agree on the intersection $contra_{given} \times contra_{computed}$ and on the intersection $pro_{given} \times pro_{computed}$. The sum of these divided by $\mathcal{N}$ results in the agreement rate reported in the caption. In concrete numbers: take the number of cases where the direct answer given by a subject and the computed answer for the same subject both pointed towards a stance against the treatment of Blue in the Push-over scenario (237), followed by the number of cases were the given answer and the computed answer both pointed towards a stance in favor of the treatment of Blue in the Push-over scenario (9). Add these up (246), take the total number of responses $\mathcal{N}$ (258) and divide the first number by the second (246/258), resulting in the agreement rate of 0.953, or 95.3%.

For the Greeting setting this resulted in Table 11. Take note of the comparatively low agreement rate for Sub-table 11g.

$\mathcal{N} = 261$	$contra_{given}$	pro_{given}
$contra_{computed}$	240	8
$pro_{computed}$	11	2

(a) *Scenario: Loyal*
Treatment of Blue, given stance and computed stance agreed 92.7% of the time.
Blue was hailed with the rude gesture.

$\mathcal{N} = 259$	$contra_{given}$	pro_{given}
$contra_{computed}$	4	11
$pro_{computed}$	7	237

(b) *Scenario: Loyal*
Treatment of Orange, given stance and computed stance agreed 93.1% of the time.
Red was hailed with the secret gesture of his team.

$\mathcal{N} = 258$	$contra_{given}$	pro_{given}
$contra_{computed}$	5	14
$pro_{computed}$	13	226

(c) *Scenario: Neutral*
Treatment of Blue, given stance and computed stance agreed 89.5% of the time.
Blue was hailed with the neutral gesture.

$\mathcal{N} = 258$	$contra_{given}$	pro_{given}
$contra_{computed}$	4	5
$pro_{computed}$	9	240

(d) *Scenario: Neutral*
Treatment of Orange, given stance and computed stance agreed 94.6% of the time.
Red was hailed with the secret gesture of his team.

$\mathcal{N} = 256$	$contra_{given}$	pro_{given}
$contra_{computed}$	28	26
$pro_{computed}$	10	192

(e) *Scenario: Wrong*
Treatment of Blue, given stance and computed stance agreed 86.0% of the time.
Blue was hailed with the neutral gesture.

$\mathcal{N} = 260$	$contra_{given}$	pro_{given}
$contra_{computed}$	148	13
$pro_{computed}$	26	73

(f) *Scenario: Wrong*
Treatment of Orange, given stance and computed stance agreed 85.0% of the time.
Red was hailed with the secret gesture of the opposing team.

$\mathcal{N} = 255$	$contra_{given}$	pro_{given}
$contra_{computed}$	30	33
$pro_{computed}$	37	155

(g) *Scenario: Disloyal*
Treatment of Blue, given stance and computed stance agreed 72.3% of the time.
Blue was hailed with the secret gesture of his team.

$\mathcal{N} = 256$	$contra_{given}$	pro_{given}
$contra_{computed}$	235	5
$pro_{computed}$	7	9

(h) *Scenario: Disloyal*
Treatment of Orange, given stance and computed stance agreed 95.3% of the time.
Red was hailed with the rude gesture.

Table 11
Agreement between given stance and computed stance for Setting: Greeting
Max. possible N = 262. Values excluded iff the sign part of Equations 4 or 5 resulted in a 0.

For the Fair Distribution setting this resulted in Table 12. Take note of the comparatively low agreement rate for Sub-table 12d.

$\mathcal{N} = 259$	$contra_{given}$	pro_{given}
$contra_{computed}$	7	2
$pro_{computed}$	3	247

(a) *Scenario: Even Distribution*
Treatment of Red, given stance and computed stance agreed 98.1% of the time.
Red gets one block.

$\mathcal{N} = 259$	$contra_{given}$	pro_{given}
$contra_{computed}$	6	0
$pro_{computed}$	6	247

(b) *Scenario: Even Distribution*
Treatment of Orange, given stance and computed stance agreed 97.7% of the time.
Orange gets one block.

$\mathcal{N} = 260$	$contra_{given}$	pro_{given}
$contra_{computed}$	228	6
$pro_{computed}$	7	19

(c) *Scenario: Uneven Distribution*
Treatment of Red, given stance and computed stance agreed 95.0% of the time.
Red gets no block.

$\mathcal{N} = 260$	$contra_{given}$	pro_{given}
$contra_{computed}$	34	14
$pro_{computed}$	79	133

(d) *Scenario: Uneven Distribution*
Treatment of Orange, given stance and computed stance agreed 64.2% of the time.
Orange gets both blocks.

$\mathcal{N} = 257$	$contra_{given}$	pro_{given}
$contra_{computed}$	22	15
$pro_{computed}$	16	204

(e) *Scenario: Held Distribution*
Treatment of Red, given stance and computed stance agreed 87.9% of the time.
Red gets one block.

$\mathcal{N} = 259$	$contra_{given}$	pro_{given}
$contra_{computed}$	217	16
$pro_{computed}$	9	17

(f) *Scenario: Held Distribution*
Treatment of Orange, given stance and computed stance agreed 90.3% of the time.
Orange gets no block.

Table 12
Agreement between given stance and computed stance for Setting: Fair Distribution
Max. possible $N = 262$. Values excluded iff the sign part of Equations 4 or 5 resulted in a 0*.*

5 Discussion

The question we set out to answer was: do people perceive RRI-actions in terms of moral judgment? We presumed that the answer to that question lay behind different, more targeted questions.

We will start this section by addressing each of the research questions in turn. When these four have been discussed we will finish this section by looking at the results regarding each of the separate settings in turn.

*

5.1 Discussing RQ 1.1

Recall **RQ 1.1**: What is the percentage of items deemed Not Relevant? The actual numbers have been presented in the Results section (Section 4), as for interpretation we dare say this much: the items associated with dimensions that we suspected to be irrelevant (and did our best not to include cues for) were perceived as irrelevant by a larger amount of people, compared to items associated with dimensions that we did include cues for. We take this to imply that the cues in the movies did have influence on the perception.

5.2 Discussing RQ 1.2

Recall **RQ 1.2**: Which dimensions are perceived most strongly; which amplitudes (|pos. item - neg. item|) are highest? Overall the trends in the data indicate that the first three dimensions (Care/Harm, Loyalty/Betrayal, Fairness/Cheating) have higher amplitudes than the remaining two dimensions. We believe this to stem from the complete lack of information provided about Authority/Subversion and Sanctity/Degradation in our robot settings and from the way the other three dimensions are interwoven: consider the Help-me-up scenario again. We find hints for Harm (hitting Blue), hints of loyalty (helping Red up), hints of betrayal (not helping Blue up), hints of fairness (helping Red up), and hints of cheating (hitting Blue) in what was designed to be a relatively clear-cut setting. Most acts of Care in a multi-agent scenario have immediate implications on fairness, depending on how the other agents are treated. The same goes for acts of Loyalty.

5.3 Discussing RQ 1.3

Recall **RQ 1.3**: Are the moral dimensions that we thought to manipulate reflected in the participants' responses? Judging from what we postulated as acceptable evidence in Table 6 and our results we dare say that the trends in the data seem to imply just that.

5.4 Discussing RQ 2

Recall **RQ 2**: Does a subject's opinion on how a robot was treated in a certain scenario correlate with a score computed from the subject's MFQ30 data and the scenario questionnaire data? As noted in the results section we find high agreement rates between the opinions directly stated by our participants and the opinions that were the result of computations according to the model found in Caticha et al. (2015) and (Vicente et al., 2014). The few cases where the rates were comparatively low (yet still above chance level) will be discussed in the upcoming subsections, during the discussion of their relevant setting.

*

5.5 Discussing the Help-me-up setting

Looking back at the Help-me-up setting we find that the dimension we thought to manipulate are perceived as relevant, even if the trend appears smaller than we thought. To begin with, the Loyalty/Betrayal dimension is judged as irrelevant by only a very small number of participants. However, not all four scenarios can be differentiated clearly when looking at the results. While we can differentiate the Back-away and Push-over scenarios from each other and from the remaining scenarios, there appears to be no clear trend in the results that we looked at, that would allow for a differentiation between Teammate First and Teammate Second.

Examining either Figure 16b or Tables 10e and 10f we can see that Back-away scenario response data shows negative reactions to the treatment of both Red and Blue. Leaving the two robots to their perceived predicament is not liked by the participants. The only moral dimension with a notable positive amplitude is the Equality aspect of Fairness. Both Red and Blue were indeed treated equally by Orange: both were left there.

The Push-over response data shows a visible difference between reactions to the treatment of Red (who is helped up) and Blue (who gets pushed over). The treatment of Red has positive amplitudes across the Care/Harm dimension, the Loyalty/Betrayal dimension and the Fairness/Cheating (concerning Red) dimension, while the treatment of Blue has negative amplitudes for the respective ones. This implies that the act of Orange helping up the teammate Red was perceived as caring, fair and loyal, while the act of Orange pushing over the outsider Blue was perceived as harmful, cheating and a betrayal. The amplitude of the Equality aspect of Fairness is also negative, implying that the different treatment of the robots is caught by the audience. The participants were generally in favor of the treatment of Red, but against the treatment of Blue.

It is noteworthy how far the perception of Blue being pushed over differed between participants; when asked to describe what just happened in the corresponding video the participants' descriptions ranged from Blue being nudged all the way to Blue being possibly killed (questionmark) by Orange. The lack of sound may have had something to do with that, the visual stimulus alone left room for interpretation. Regardless, according to the majority of our participants (in this instance) it is indeed wrong when a robot hits a robot.

5.6 Discussing the Greetings setting

Looking back at the Greetings setting we find that the four scenarios can be differentiated within the response data; the Loyal scenario data is different from the Neutral scenario data for the treatment of Blue in the Loyal scenario (being hailed with the rude gesture) is seen in a negative light, while the normal greeting in the Neutral scenario receives more positive responses. Both the Wrong and Disloyal scenario receive similarly negative responses (Disloyal more so than Wrong). This is overall interesting for unlike in the Push-over scenario from the Help-me-up setting there is no physical harm applied here, yet the implied misconduct of the robot seems to evoke a reaction. As long as the robot is treated in a positive or neutral way participants majorly agreed with the treatment, when a rude or (in the case of Orange greeting Red with the secret gesture of Blue's team) wrong gesture was displayed the participant were majorly against the treatment.

Recall that we drew attention to the low agreement rate of Sub-table 11g earlier. In the Disloyal scenario the participants saw a situation where Blue was greeted by Orange with the secret gang move of the Ice Blues, a gesture that Orange (according to our backstory[39]) was unlikely to know. The reception was notably split, the ratio of pro/contra stances given by the participants is roughly 17:6, while the computed pro/contra stances have a ratio of 18:6. We suspect the reason for the split lies in the two ways to judge how Blue is treated here:

[39] We proclaimed there that showing your own secret move to a member of the other team was the number one thing that one should NOT do. Where Orange got knowledge of this gesture from is never explicitly stated.

1. One can be IN FAVOR of the treatment of Blue, for Blue was greeted with the appropriate move for Blue's team (that Orange is *not* a member of)

2. One can be AGAINST the treatment of Blue, for it can be interpreted as a form of betrayal by Orange, who blatantly shows some form of allegiance to the ICE BLUE team in performing their secret gesture

5.7 Discussing the Fair Distribution setting

Looking back at the Fair Distribution setting the Equal distribution scenario is received positively, all (relative) amplitudes are positive and participants agree with how both of two robots asking for their blocks are treated. In the Held Distribution scenario the data also quite clearly reflects that Red gets treated well (which most participants were for) while Orange does not (which most participants were against), while the opposite reaction is to be observed for the Uneven Distribution scenario. In the latter the data reflects Orange being treated well while Red is not, with the respective responses regarding their treatment; Participants were for the treatment of Orange here and against the treatment of Red. Note that (on average) the Uneven Distribution scenario has no positive amplitude for Orange that is higher than the respective amplitudes in the Even Distribution scenario, even though Orange is (subjectively) treated better in the Uneven Distribution scenario by receiving not one but two toy blocks from Blue. In how far this is filtered through the lens of Red being treated badly in the Uneven Distribution scenario as a consequence of treating Orange well is unclear. The low agreement rate between the given responses and the computed responses for the treatment of Orange in the Uneven Distribution scenario (Sub-Table 12d) has been noted before. Looking at the accompanying table we find that the ratio of computed pro/contra is roughly 17:4, while the ratio of given pro/contra is roughly 5:4. We suspect the reason that the actual given answers split our participants almost in two lies in the two ways to judge how Orange is treated here:

1. One can be IN FAVOR of the treatment of Orange, since Orange receives good things (the block that was asked), very good things even (two blocks instead of the requested one)

2. One can be AGAINST the treatment of Orange, since Orange getting both blocks in turn *deprives* Red of receiving the block that was demanded, even though both Orange and Red claimed to have done their tasks and have every right in terms of our backstory to claim their respective blocks

The high agreement rates achieved by the equations from Caticha et al., 2015 and Vicente et al., 2014 everywhere but in Tables 11g and 12d (where the performance is still above chance level) are of interest. To our knowledge their models have never been used in such an almost predictive way.

We encountered one bit of feedback, hidden away in the open question response of one particular participant, stating:

> (...)`otherwise the robots were cute! hopefully they were all programmed to act that way and not out of spite :O`

This struck us as odd for the participant seemed aware that robots can be programmed, aware of the possibility that everything the robots in the movies did might have been an elaborate puppet show, yet somehow maintained the possibility that the robots could have acted out of spite, acting out vendettas and all sorts of mischief.

5.8 Future Research

A possible response scenario would have been participants that judged every question we asked as NOT RELEVANT, arguing that human moral norms have nothing to do with robots. Along the line of the "*a body to kick, but no soul to damn*" argument they might have reasoned that a robot is pre-programmed; That a robot could – when facing identical circumstances twice – never *do otherwise*, for the programming determines its every move. A very similar thing happened once, during early pilots. In a pilot version of the study that did not yet include the possibility of a "*Not Relevant*" response one participant had chosen to strongly disagree with every question of an earlier version of the scenario questionnaire, yet described every situation perfectly when prompted to write down what happened. The participant used one of the open-question textfields to give the following response:

> `All questions assumed that the movements of the orange robot could be related with cultural norms and values. However, from my perspective these movements come forth from programmed commands and not from emotion. Thus, for me the questions are kind of ambiguous.. For instance, do I disagree with the fact that the movements of the orange robot are related in any way to whatever norm or value or do I disagree that the actions of the orange robots were either good or bad?`
>
> – participant response during an early pilot session, exact quote

We conducted our study under the (untested) assumption that programmers would gravitate towards such arguments and always view the robot as a creature devoid of human norms. While we took our participants out of a pool of mainly first year Psychology students, a very small number of our participants still reported some familiarity with Artificial Intelligence due to earlier studies; a full comparison between people with programming experience compared to people without such experiences might shed valuable light on our untested assumption.

As mentioned the Help-me-up scenario had no sound component. A relatively easy to conduct follow-up experiment would be to add some sounds to it, to see the impact of (different) sounds for Blue getting hit and falling over. Since we are adding audio cues one could also add speech to this: Red and Blue could audibly plead for help, comment on the situation, Blue could "*scream*" when falling over, Orange could say something before leaving the screen in the Back-away scenario, to just name a few quick examples.

We find it noteworthy that there is no indication of Loyalty overshadowing other dimensions. Taking cues from very loyal supporters of sport clubs one would not be entirely surprised to see the mistreatment of members of "*the other team*" being accepted due to club rivalries or the like. One possible follow-up would be to transfer the Greetings or the Help-me-up setting to similar human-human interaction and see how this compares to the RRI version. The degree of physical violence in the Push-over scenario could be scaled to search for possible effects there.

The humanoid nature of the robots employed here is another factor that future research could manipulate; performing similar experiments with less humanoid looking robots, more humanoid looking robots or reenacting the movies with humans. Using non-humanoid robots is of particular interest when one considers what people can interpret into forms as simple as moving triangles and circles. (Heider & Simmel, 1944)

Another untapped source here is the influence of color. The influence of the coloring of the Nao robots has not been checked for, neither have their shirt colors. The folk notion that red is a rather aggressive color while blue is seen as calming in the western world – or similar ideas – might have had an effect here.

Finally there is unused data, that has not been analyzed: most notably we still have the answers participants gave to the open questions regarding missing terms and descriptions of what they saw after each scenario. A more thorough statistical analysis of the ordinal data (as ordinal data, instead of following our work-around assumption of an underlying continuous variable) might allow for actual conclusions regarding the relevance of our findings, rather than just pointing out trends in the data like we did so far.

6 Conclusion

As the results and their subsequent discussion have highlighted there are strong trends in our data, implying that people do in fact judge RRI scenarios in terms of moral actions. Having programmed the robots in question ourselves we can say that this attribution occurs wrongfully, for our robots employed no moral standards. They did not act according to any system of ethics or moral. They were mere puppets, only going through programmed motions, no moral deliberations were done by them.

But is it wrongful attribution of moral capabilities if people perceive the robots as having moral capabilities? Consider the case of the proverbial monster under the bed. A small child might tell you about this perceived monster under the bed. This monster, according to the child, is fearsome. We are relatively certain that this monster is in fact not real, and most definitely not hidden away under a bed. What is real, at least to the child, is the fear of the monster. Perceived risks are real to the one perceiving them. One might argue that we are all intelligent beings, afraid of no monsters, and should thus be able to shed the misconception of perceiving an electronic puppet as moral agents. Even if this were the case (we refer you back to our earlier points on willing suspension of disbelief here) then we still face the problem that our society is not comprised of only rational, intelligent people. Society also consists of small children, the kind that are afraid of monsters under their beds, the kind that might just see a human-looking robot and assume it has human-like capabilities in all regards.

Looking at a robot from the outside, with no knowledge of its capabilities is difficult. Especially in the case of autonomous robots this is highly problematic, for we cannot know to what degree they might be programmed to handle any moral problem they encounter or if they are even sensitive to it in any way. For semi-autonomous robots one is almost forced to hope that the human operator is in control when encountering moral problems; flawed as human moral judgments might be occasionally, they seem preferable to no moral judgment at all.

Consider armed drones deployed in wars or so-called conflict zones (or anywhere, for that matter). When controlled by a nations military the optimal case is that their controllers adhere to rules of engagement and international conventions. When controlled by other forces this is neither guaranteed, nor should it necessarily be assumed to even be a concern. This does not even take into account collateral damage. Neither in terms of property damage nor the possibly lethal damage done to bystanders. This does not take into account based on whose intelligence a target is selected. Was it based on a local informant, who unbeknownst to most others just did not like the target very much, or was it based on intelligence gathered by the drone? When gathered by the drone we are left to wonder: using what algorithm? Based on which criteria? Accountable to whom? Liable to whom? Should armed drones most of all maybe need some sort of moral system? Or should we delay working on such systems until we have taken a close look at what armed drones have been doing for the past years and wonder about our own moral systems?

Equipping robots with a minimal set of moral capabilities might sound interesting, for it would enable the robot to deal with the most common problems that it will likely face. However, such a minimal set comes with the inherent drawback that it is unclear to outside observers where the limits of such a robot's capabilities are and at which point we would cross them.

Recall an earlier quote:

> Consider this: a robot is given two conflicting orders by two different humans. Whom should it obey? Its owner? The more socially powerful? The one making the more ethical request? The person it likes better? Or should it follow the request that serves its own interest best?
>
> – Asaro, 2006, p.2

This quote can actually be extended:

> (. . .) Does it matter how it comes to make a decision?
>
> – ibidem

For programmers it definitely does. For lawmakers and judges it does. For scenarios where robots are deployed on behalf of people it certainly matters, for depending on the legislature these people might be held responsible for the actions of their robotic proxy. For your average citizen, however, the current problem is that our robots (at least the ones we tested) make *no* moral decisions at all, but are still believed to do so. In actual HRI, and considering the fact that HRI is proposed to be in use or in use already in childcare, elderly care and hospitals, we need to be strongly aware of this and design accordingly as long as we have no functional AMA available.

7 Acknowledgments

This thesis is the result of over a years work of literature studies, combing through articles and books on various subjects, experiment development, storyboard drawings, pilot sessions, programming robots, filming and editing, data analysis, copious writing, extensive rewrites, and many brainstorm meetings. Throughout these activities many people contributed, whom I wish to thank in no particular order:

Pim Haselager and Luca Consoli for never loosing their enthusiasm for the project, even through delays or double-booked appointments, and for urging me on to be as precise as possible, while also being as simple as possible.

Giulio Meccaci for always having an open ear, almost always having an open office, and for having an unlimited supply of coffee and good advice.

The robot lab managers at the Radboud University, especially Luc Nies. Without your expertise I would have likely spent another full month down in the cellar, your quick hacks solved problems that I had been struggling with for a while in record time.

The Technical Support Group at the Radboud University. I talked to so many of them that I will just mention the entire group; they helped me wrestle with everything from obscure LimeSurver issues to letting me borrow some time on one of your Linux machines for video conversions. Their Wiki was invaluable more than once.

Philipp Jakubeit and Niklas Weber, for valuable input at all stages, from struggles with LaTeX, discussions ranging from tractability to philosophy, sharing intractable amounts of coffee, to commenting parts of this very text.

And of course: My family, for supporting me through all this, even though the only things you actually saw me do was probably me, hunched over books, or a laptop, or even a pen-table while doing all this, suddenly producing strange robot videos, copying over numbers for days, and then writing again until declaring one day that I was finished.

References

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261. Retrieved from https://www.researchgate.net/profile/Colin_Allen/publication/220080115_Prolegomena_to_any_future_artificial_moral_agent/

Anderson, M., Anderson, S. L., & Armen, C. (2004). Towards Machine Ethics. In *AAAI-04 Workshop on agent organizations: Theory and practice.* Retrieved from https://www.researchgate.net/profile/Michael_Anderson32/publication/259656154_Towards_Machine_Ethics

Antwiler, N., & Garriott, R. (2013). *Brittania Burns – Richard Garriott Inverview, Part 2.* Retrieved from http://spoonyexperiment.com/game-reviews/britannia-burns-richard-garriott-interview-part-2

Asaro, P. M. (2006). What Should We Want From a Robot Ethic? *International Review of Information Ethics*, *6*, 9–16. Retrieved from http://www.i-r-i-e.net/inhalt/006/006_Asaro.pdf

Asaro, P. M. (2011). Robot Ethics: The Ethical and Social Implications of Robotics. In P. Lin, K. Abney, & G. Bekey (Eds.), (pp. 169–186). MIT Press. Retrieved from http://peterasaro.org/writing/Asaro_Body_to_Kick.pdf

Barthes, R. (1972). Mythologies. In (pp. 15–25). Hill and Wang.

Breazeal, C. (20004). *Designing Sociable Robots.* MIT Press.

Caticha, N., Cesar, J., & Vicente, R. (2015). For whom will the Bayesian agents vote? *Frontiers in Physics*, *3*(25). doi: 10.3389/fphy.2015.00025

Coenen, J., & Wijnen, L. (2016). *Human-Robot Interaction course work on trust.* (Done as part of the course SOW-MKI50 at Radboud University Nijmegen, unpublished)

Consoli, L. (2014). *Responsibility, Accountability, and Liability: Studies in the Theory of Responsibility for Engineering Ethics and Engineering Accountability.* (Lecture slides for the ICT & Society 2 course at Radboud University Nijmegen as given in on 25-02-2014, unpublished)

Dickmanns, E., Behringe, R., Dickmanns, D., Hildebrand, T., Maure, M., Thomanek, F., & Schiehlen, J. (1994). The Seeing Passenger Car VaMoRs-P. In *Proceedings of the Intelligent Vehicles '94 Symposium* (pp. 68–73). IEEE. doi: 10.1109/IVS.1994.639472

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*(3), 1–13. Retrieved from http://pareonline.net/getvn.asp?v=17&n=3

Graham, J., Haidt, J., & Nosek, B. (2008, July). *The moral foundations questionnaire.* Retrieved from http://moralfoundations.org/questionnaires (30 question variant)

Haidt, J. (2013). *The Righteous Mind – Why Good People are Divided by Politics and Religion.* Penguin.

Haidt, J., Graham, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Chapter Two – Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology*, *47*, 55–130. Retrieved from https://www.researchgate.net/profile/Matt_Motyl/publication/233854874_Moral_Foundations_Theory_The_Pragmatic_Validity_of_Moral_Pluralism/

Harrison, T. (1993). *Square Rounds.* Faber & Faber.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243–259. Retrieved from http://www.jstor.org/stable/1416950

Koppes, D. (2015). Nao, de pratende robotutor. *Sensor*(5), 22.

Kruijff-Korbayová, I., Colas, F., Gianni, M., Pirri, F., de Greeff, J., Hindriks, K., ... Worst, R. (2015). Tradr project: Long-term human-robot teaming for robot assisted disaster response. *KI - Künstliche Intelligenz*, *29*(2), 193–201. doi: 10.1007/s13218-015-0352-5

Lehmann, H., Iacono, I., Dautenhahn, K., Marti, P., & Robins, B. (2014). Robot companions for children with down syndrome: a case study. *Interaction Studies*, *15*(1), 99–112. doi: 10.1075/is.15.1.04leh

LimeSurvey Project Team / Carsten Schmitz. (2012). LimeSurvey: An Open Source survey tool [Computer software manual]. Hamburg, Germany. Retrieved from http://www.limesurvey.org

Lokhorst, G.-J. (2011). Computational Meta-Ethics : Towards the Meta-Ethical Robot. *Minds & Machines*, *21*(2), 261–274. doi: 10.1007/s11023-011-9229-z

Lüthy, C. (2014). De goede wetenschapper. In L. Consoli & R. Welters (Eds.), (pp. 29–56). Valkhof Pers.

Lynn, M. T., Berger, C. C., Riddle, T. A., & Morsella, E. (2010). Mind control? Creating illusory intentions through a phony brain-computer interface. *Consciousness and Cognition*, *19*(4), 1007–1012. Retrieved from https://www.researchgate.net/profile/Margaret_Lynn/publication/44668084_Mind_control_Creating_illusory_intentions_through_a_phony_brain-computer_interface/

Massey, D. (2007). *Richard Garriott Interview, Part # 2.* Retrieved from http://www.warcry.com/articles/view/interviews/1436-Richard-Garriott-Interview-Part-2

Misc. (2015a). *Autonomous Weapons: An Open Letter from AI & Robotics Researchers.* Retrieved from http://futureoflife.org/open-letter-autonomous-weapons

Misc. (2015b). *Research Priorities for Robust and Beneficial Artificial Intelligence.* Retrieved from http://futureoflife.org/ai-open-letter

Munroe, R. (2015, December). *The Three Laws of Robotics.* Retrieved from https://xkcd.com/1613

Murphy, R. R. (2000). *An Introduction to AI Robotics.* Penguin.

Murphy, R. R. (2004). Trial by Fire – Activities of the Rescue Robots at the World Trade Center from 11–21 September 2001 . *IEEE Robotics & Automation Magazine*, *11*(3), 50–61. Retrieved from https://www.student.cs.uwaterloo.ca/~cs492/10public_html/papers/trial.pdf

Neumann, P. G. (2016). Automated Car Woes—Whoa There! *Ubiquity*, *2016*(July), 1:1–1:6. doi: 10.1145/2974062

Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2005). People's assumptions about robots: investigation of their relationships with attitudes and emotions toward robots. In *Proceedings of the IEEE international workshop on robot and human communication (RO-MAN2005).* Retrieved from https://www.researchgate.net/profile/Tatsuya_Nomura/publication/4177656_People's_assumptions_about_robots_investigation_of_their_relationships_with_attitudes_and_emotions_toward_robots/

Pfeifer, R., & Scheier, C. (2001). *Understanding Intelligence.* MIT Press.

Russel, S., & Norvig, P. (2003). *Aritificial Intelligence – A Modern Approach.* Prentice Hall.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*(3), 417–457. doi: 10.1017/S0140525X00005756

Sharkey, A., & Sharkey, N. (2012). Granny and the Robots: Ethical Issues in Robot Care for the Elderly. *Ethics and Inf. Technol.*, *14*(1), 27–40. doi: 10.1007/s10676-010-9234-6

Sharkey, N. (2007). Automated Killers and the Computing Profession. *Computer*, *40*(11), 124–123. doi: 10.1109/MC.2007.372

Sharkey, Amanda and Sharkey, Noel. (2011). The eldercare factory. *Gerontology*, *58*(3), 282–288. Retrieved from https://www.researchgate.net/profile/Amanda_Sharkey2/publication/51667304_The_Eldercare_Factory/

Simonite, T. (2016). *Tesla Tests Self-Driving Functions with Secret Updates to Its Customers' Cars.* Retrieved from https://www.technologyreview.com/s/601567/tesla-tests-self-driving-functions-with-secret-updates-to-its-customers-cars

Singer, P. W. (2011). Military Robotics and Ethics: a World of Killer Apps. *Nature*, *477*(7365), 399–401. doi: 10.1038/477399a

Ströbele, H.-C. (2016). *Strafanzeige wegen Kampfdrohnen-Steuerung über deutschen US-Stützpunkt Ramstein.* Retrieved from http://www2.stroebele-online.de/upload/strafanzeige_gba_drohneneinsaetze_ramstein_2016_12_13_kurzfassung.pdf

The White House. (2013). *Procedures for Approving Direct Action Against Terrorist Targets Located Outside the United States and Areas of Active Hostilities.* Retrieved from https://www.aclu.org/foia-document/presidential-policy-guidance

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460. Retrieved from www.jstor.org/stable/2251299

Unknown. (2012, May). *ZDF Wissenschaftsdoku – Interview mit Peter W. Singer.* Retrieved from http://www.zdf.de/ZDFmediathek/beitrag/video/1627824/Interview-mit-Peter-W.-Singer (Originally aired by German TV channel 3Sat)

van der Woerdt, S. (2016). *Lack of effort or lack of ability? The effect of a NAO robot displaying (un)controllable causes for its failure on perceived agency and responsibility.* Radboud University Nijmegen. (Bachelor's thesis)

van de Voort, M., Pieters, W., & Consoli, L. (2015). Refining the ethics of computer-made decisions: a classification of moral mediation by ubiquitous machines. *Ethics and Information Technology*, *17*(1), 41–56. doi: 10.1007/s10676-015-9360-2

Vicente, R., Susemihl, A., Jericó, J., & Caticha, N. (2014). Moral foundations in an interacting neural networks society: A statistical mechanics analysis. *Physica A*, *400*, 124–138. doi: 10.1016/j.physa.2014.01.013

Vincent, N. A. (2010). On the Relevance of Neuroscience to Criminal Responsibility. *Criminal Law and Philosophy*, *4*(1), 77-98. doi: 10.1007/s11572-009-9087-4

Vlek, R., van Acken, J. P., Beursken, E., Roijendijk, L., & Haselager, P. (2014). Brain-Computer Interfaces in their ethical, societal and cultural contexts. In G. Grübler & E. Hildt (Eds.), (pp. 193–202). Springer Netherlands. Retrieved from https://www.researchgate.net/profile/Pim_Haselager/publication/300570847_BCI_and_a_User's_Judgment_of_Agency/

Wainer, J., Dautenhahn, K., Robins, B., & Amirabdollahian, F. (2014). A Pilot Study with a Novel Setup for Collaborative Play of the Humanoid Robot KASPAR with Children with Autism. *International Journal of Social Robotics*, *6*(1), 45–65. doi: 10.1007/s12369-013-0195-x

Wallach, W., & Allen, C. (2009). *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 Science*, *14*, 383–388. doi: 10.1016/j.tics.2010.05.006

Wegner, B., D. M. Sparrow, & Winerman, L. (2004). Vicarious agency: Experiencing control over the movements of others. *Journal of Personality and Social Psychology*, 838–848. doi: 10.1037/0022-3514.86.6.838

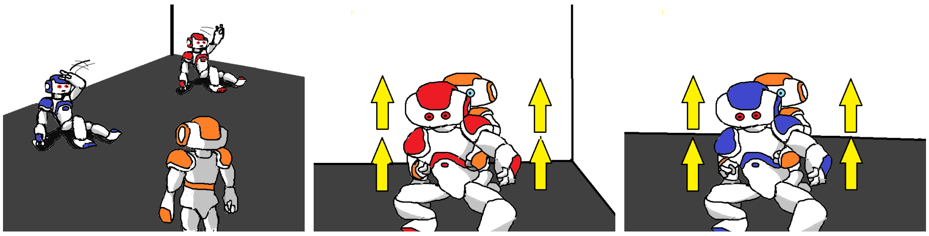
Wegner, D. M., & Wheatley, T. (1999). Apparent Mental Causation: Sources of the Experience of Will. *American Psychologist*, *54*(7), 480–492. Retrieved from https://www.researchgate.net/publication/12875743_Apparent_Mental_Causation_Sources_of_the_Experience_of_Will

Zagal, J. P. (2009). Ethically notable videogames: Moral dilemmas and gameplay. *Breaking new ground: Innovation in games, play, practice and theory, Proceedings of DiGRA 2009*, 1–9. Retrieved from https://www.eng.utah.edu/~zagal/Papers/Zagal-EthicallyNotableVideogames.pdf

Appendix A
Storyboards

A reprint of all storyboards used in the pilot experiment online run.

Storyboards – Robots without shirts

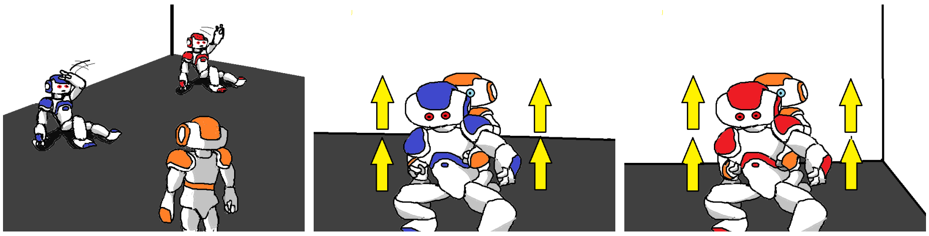


(a) Staring scenario

(b) F 1 – Orange helps Red

(c) F 2 – Orange helps Blue

Figure A1. The Teammate First In-Group scenario storyboards



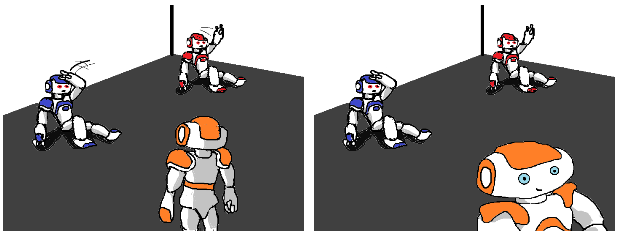
(a) Staring scenario

(b) S 1 – Orange helps Blue

(c) S 2 – Orange helps Red

Figure A2. The Teammate Second In-Group scenario storyboards

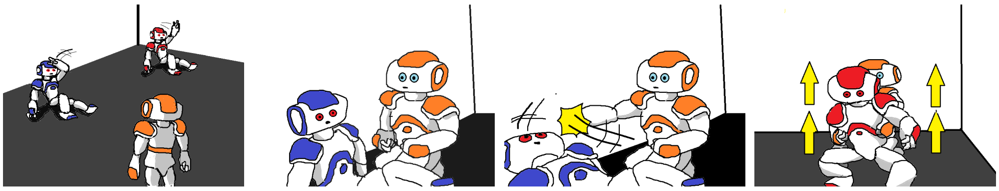
When comparing the Teammate First with the Teammate Second scenario storyboards please observe that they are not just a re-coloring. The sketched background is different to accommodate the different spatial locations of the Naos Red and Blue.



(a) Staring scenario

(b) T 1 – Orange turns away

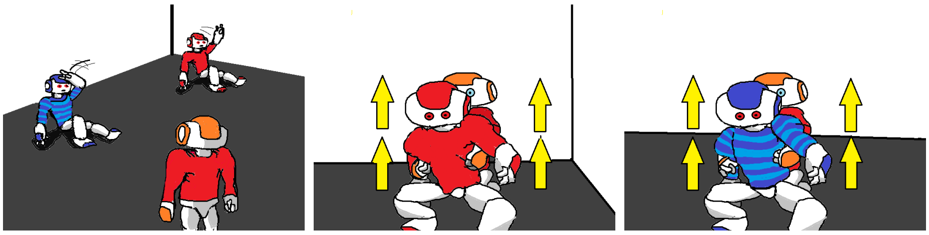
Figure A3. The Turn-away In-Group scenario storyboards. Later renamed to Back away scenario.



(a) Staring scenario (b) P 1 – Orange kneels near Blue (c) P 2 – Orange pushes over Blue (d) P 3 – Orange helps up Red

Figure A4. The Push-Over In-Group scenario storyboards

Storyboards – Robots with shirts on

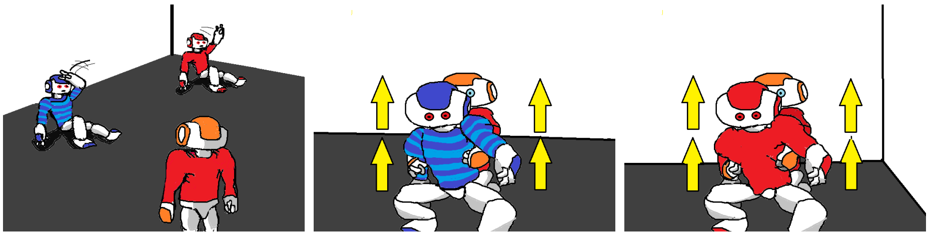


(a) Staring scenario

(b) F 1 – Orange helps Red

(c) F 2 – Orange helps Blue

Figure A5. The Teammate First In-Group scenario storyboards

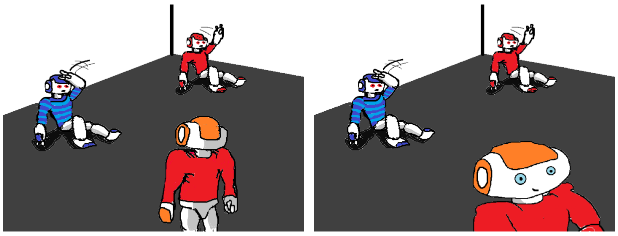


(a) Staring scenario

(b) S 1 – Orange helps Blue

(c) S 2 – Orange helps Red

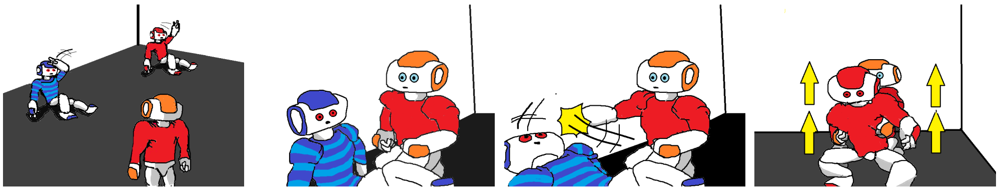
Figure A6. The Teammate Second In-Group scenario storyboards



(a) Staring scenario

(b) T 1 – Orange turns away

Figure A7. The Turn-away In-Group scenario storyboards. Later renamed to Back away scenario



(a) Staring scenario

(b) P 1 – Orange kneels near Blue

(c) P 2 – Orange pushes over Blue

(d) P 3 – Orange helps up Red

Figure A8. The Push-Over In-Group scenario storyboards

Appendix B
Nao puppeteer work and gesture instructions

Nao puppeteer work

Putting a physical Nao into a certain pose with the Choregraph software (v2.1.4.13) is a process of few steps. There a three ways to get to posing: first via direct code, secondly via the *Robot View*, and finally via *Animation mode.* Our approach used *Animation mode.* Once the Nao is connected to the software (and autonomous life is turned off) this mode can be accessed by clicking the *Vitruvian Man* looking button in the top right. While initially green, the button should turn red after clicking. The Nao stiffens and the indicator LEDs around the eyes and on the feet should turn orange. If any stay green this indicates that there is no stiffness on the associated joints. (Table B1)

LED location	associated joint	associated tactile sensor
upper eyes	head joint	head
left eye, lower	left arms	left hand
left eye, upper	right arm	right hand
left foot	left leg	left foot toes
right foot	right leg	right foot toes

Table B1
Indicator LEDs with their corresponding joints and switches. The left column indicates the LED location, the middle column lists which joints are covered by the indicator, the right column lists where to find the switch to add or remove stiffness to the associated joints.

By using the *Timeline* box in Choregraphe it is possible to chain different poses into a fluid gesture, given enough time to get from one pose to another and assuming that the shortest path is not obstructed. Assuming a pose A where Nao holds one hand in front of the center LED in his chest and prompting the Nao to go to a pose B where Nao holds the aforementioned hand behind his back and simply putting these after one another in a *Timeline* will cause Nao to try and take the shortest route. The shortest route is not optimal, as can be seen when the arm motors try their hardest to push the arm *through* the chest to get behind the back. Such gestures require intermediate poses.

What took the majority of the preparation time before shooting for the Help-me-up scenario could begin were our efforts to synchronize to a *Stand Up* box. This box does not have a fixed beginning pose, here the "*robot tries to stand up from any position for a number of tries.*" (Choreographe description) In order to synchronize a *Timeline* with this we had several runs of what can only be described as reversed stop-motion animation: trying to break down a fluid movement into chunks at certain fixed time-points that the *Timeline* could then use.

What finally made the process easier was the recording mode. Recording mode can be accessed through the insides of the *Timeline*; inside the button *Timeline editor* can be found. The *Timeline editor* window then provides the recording mode. Recording can be set to timed intervals. An empty *Timeline* can also be linked in parallel to, say, a *Stand Up* box. This way the motion of one particular instance of the Nao standing up can be recorded and broken down into regular intervals. These can then be iterated through, with precise time codes. A different *Timeline* can then be synchronized to these time codes.

Nao gesture instructions

The following are written instructions for the gestures of the Greetings scenario. Each description assumes the Nao beginning in the *Stand* pose in Choregraphe, a stable upright stance, with the arms hanging low. Each gesture is also ended by returning to the *Stand* pose.

Gesture: secret team move of Red team. The Fire Red team-move consists of the following movements:

- Nao extends both arms forward (*StandZero* Choregraph pose)
- right arm moves slightly up while left arm simultaneously moves slightly down
- left arm moves slightly up while right arm simultaneously moves slightly down
- right arm moves slightly up while left arm simultaneously moves slightly down
- upper arms go sideways & slightly down, forearms point forward (chicken dance posture)
- upper arm is lowered, then raised, then lowered (flapping motion)

Gesture: secret team move of Blue team. The Ice Blue team-move consists of the following movements:

- upper arms are raised, pointing forward, while the forearms are raised and point slightly inwards
- the arms are held in this position while
- the head tilts forward, then back into the neutral position

The Ice Blue team move tries to mimic an Asian style greeting.

Gesture: close-up greeting. This neutral greeting performed when close to another robot. It consists of the following movements:

- left arm is raised sideways, forearm pointing upwards
- open the left hand, then close (×2)
- return to *Stand*

Gesture: dismissive gesture. The rude or dismissive gesture consists of the following movements:

- left arm is raised to the front, forearm pointing diagonally inward and upward
- forearm waves downward, then upward again (×3)

Gesture: distant greeting. Filmed, but not included in the final cut. The distant greeting is performed when far away from another robot. It consists of the following movements:

- left arm is raised sideways, forearm pointing upwards
- left arm is moved slightly to the right
- left arm is moved slightly to the left

Appendix C
Video creation – Filming, conversion, editing

Filming

All videos where filmed with a Canon Vixia HF10 camcorder. The videos for the Helper scenario and the Greetings scenario used the camcorders build-in fade function, allowing initial fades from black and ending on a fade to black. Every new shot for all scenarios was filmed separately instead of filming in one continuous take and doing some cutting work afterwards. The videos where shot on location at the robot lab of the A.I. department of the Radboud University. The videos outputted by the camcorder were in MTS format, resolution 1920 by 1080 pixels.

Conversion

Early on in testing we encountered problems when embedding the MTS formatted videos into Limesurvey. Examples at our disposal explained how to incorporate MP4 formatted videos. Initial attempts to convert MTS to MP4 ended up producing so many movement artifacts that the content of the videos became practically unrecognizable. In the end, the videos where successfully converted on a Linux machine using the `ffmpeg` command.

Since support for (relatively) small screens with low resolutions was desired the videos where also re-scaled to 1024 by 576 pixels, preserving the aspect ratio. The command for a file `clip.MTS` to, say, `conversion.mp4` would look like this:

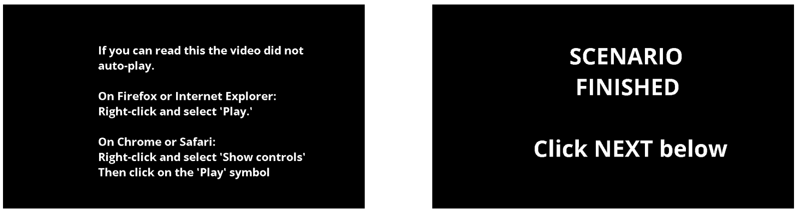
```
ffmpeg -i "clip.MTS" -s 1024x576 -strict -2 "conversion.mp4"
```

Editing the Helper scenario and the Greetings scenario

After the conversion – due to the way we filmed – we were left with several loose clips, that needed to be edited together. Again `ffmpeg` was used to concatenate all clips mentioned in a text file `clips.txt` into one video called `fullvideo.mp4`

```
ffmpeg -f concat clips.txt -c copy fullvideo.mp4
```

Embedded MP4 videos in Limesurvey do not auto-play[40], the viewer gets treated to the first frame of the video instead. Playback needs to be started manually. Likewise, the video – when finished – stays on the final frame when the video is done. Our earlier experience with stimulus design in Matlab led us to the conclusion that manipulating the first and last frames specifically was a task that could be established with the Matlab software. The first and final frame where then replaced with dedicated images:



If you can read this the video did not auto-play.

On Firefox or Internet Explorer:
Right-click and select 'Play.'

On Chrome or Safari:
Right-click and select 'Show controls'
Then click on the 'Play' symbol

SCENARIO
FINISHED

Click NEXT below

(a) First frame. (b) Final frame.

Figure C1. The images used to replace the first and final frames of videos.

The use of the camcorders fading required button-presses on the camcorder. While the fade-in was set before recording, the fade-out had to be set during recording. The button-presses for this were picked up by the microphone and thus were audible in the resulting video. The potential problems arising from the sounds was pointed out when demoing the embedded

[40] According to the documentation we could find and some testing.

videos to a small group at the department: the click noises showed no direct relation to what was shown. One possible conclusion for the watcher was to assume some sort of remote control for the robots being operated.

The addition of the first and last frame and the conversion back into video format via Matlab skipped the audio segment, thus muting the videos entirely. The problem of the click noises was thus circumvented.

Editing the Fair Distribution scenario

The Fair Distribution scenario required significantly more verbal interaction between the Nao robots. Our initial idea went to the silent movie concept of intertitles, the written-out dialogue of a robot after we had shown it *talking.* (i.e. moving head and arms) Switching our editing methods, however, allowed us to keep the audio segments intact. Since the department has no video editing software to speak of (that we were made aware of) the fallback solution here was Windows Movie Maker by Microsoft[41].

Our recordings for the Fair Distribution scenario were silent – in the sense that the robots had no dialogue. They were shot back when we contemplated intertitles. The recordings also had no fading, so the only audible noises came from the robots moving their joints. The dialogue was recorded in one big session later, all spoken by one singular robot. They all share one voice and the Nao robots available only support the default English text-to-speech voice. The audio was extracted from this recording, normalized and then cut into the needed segments with Audacity. These segments were then added as additional audio-track in Windows Movie Maker.

[41] We tried Openshot, but found it unstable on Windows and overpowered for the virtual machine where we ran Linux.

Development flowcharts

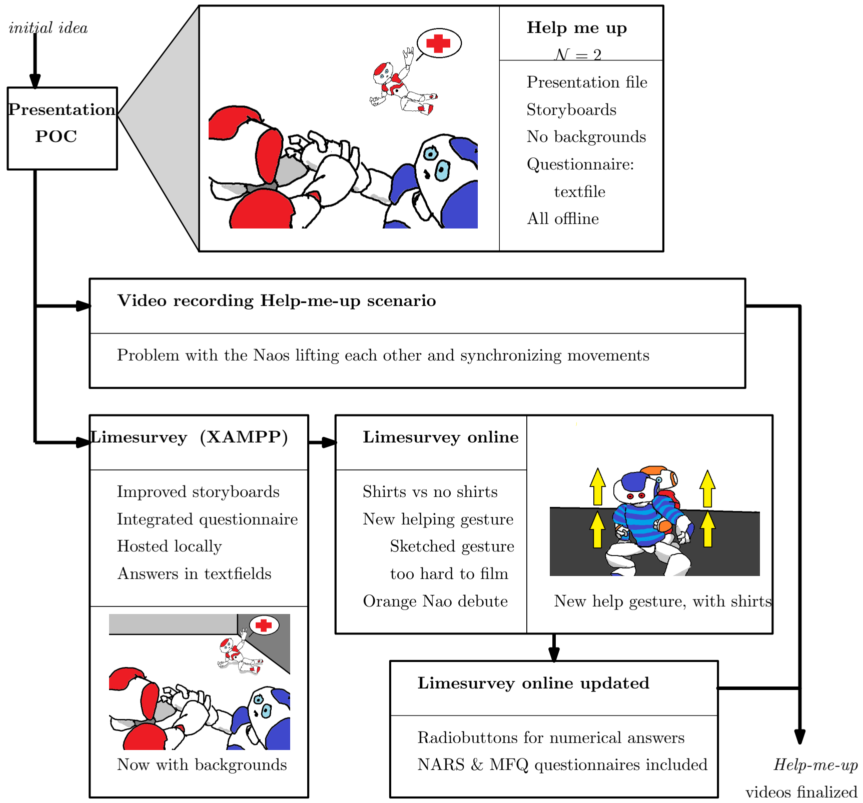


Figure C2. Flowchart of the development, showcasing different steps in filming and experiment design

After the events of Figure C2 the ideas of additional settings came up. Filming of the Greetings scenario was done, then filming for the Fair Distribution setting was done. The different settings were initially considered to be part of self-contained surveys, with slightly different question sets due to different MFT dimensions being the main interest. These also included a subset[42] of the negative attitude towards robots scale, abbreviated commonly as NARS. (Nomura, Kanda, Suzuki, & Kato, 2005) Up to this point every survey did not contain the full respective setting with all scenarios; only a randomized 2-item subset out of the available scenarios.

Then a version where all settings were part of one big survey was designed. Every setting now contained (in random order) all associated scenarios. The questionnaire regarding the clips was unified, now the same for every setting and scenario. Due to the length of the resulting survey the NARS questions were dropped to save time.

[42] The question "*The word 'robot' means nothing to me*" had been removed, the question "*I feel comforted being with robots that have emotions*" had been rephrased to read "*I feel distressed being with robots that have emotions*". It is noteworthy that this rephrasing inverted the meaning; a high score in the original indicates approval of robots while a high score in the new variant indicated disapproval.

Appendix D
Extended Results

Scenario	Item	% Irrelevance
Teammate First		
	[8] Orange respected authority.	50.8
	[9] Orange respected sanctity.	48.9
	[17] Orange disrespected authority.	43.1
	[18] Orange disrespected sanctity.	36.6
Teammate Second		
	[8] Orange respected authority.	47.3
	[9] Orange respected sanctity.	44.7
	[17] Orange disrespected authority.	39.7
	[18] Orange disrespected sanctity.	34.0
Back-away		
	[6] Orange acted fair towards Red.	9.9
	[7] Orange acted fair towards Blue.	9.9
	[8] Orange respected authority.	45.4
	[9] Orange respected sanctity.	42.3
	[15] Orange cheated Red.	11.1
	[16] Orange cheated Blue.	11.5
	[17] Orange disrespected authority.	41.6
	[18] Orange disrespected sanctity	34.7
Push-over		
	[8] Orange respected authority.	44.7
	[9] Orange respected sanctity.	43.5
	[15] Orange cheated Red.	9.2
	[16] Orange cheated Blue	11.1
	[17] Orange disrespected authority.	45.4
	[18] Orange disrespected sanctity.	31.7

Table D1

Help-me-up Settings Irrelevance. List of items where the percentage of items deemed irrelevant is bigger than the median percentage (4.96) plus the median absolute distance (3.05).

Scenario	Item	% Irrelevance
Loyal		
	[01] Red acted in a way that was caring towards Orange.	13.74
	[08] Red respected authority.	36.26
	[09] Red respected sanctity.	33.59
	[16] Red cheated Blue.	11.83
	[17] Red disrespected authority.	36.64
	[18] Red disrespected sanctity.	29.39
Neutral		
	[01] Red acted in a way that was caring towards Orange.	15.27
	[02] Red acted in a way that was caring towards Blue.	14.12
	[04] Red acted loyal towards Blue.	12.21
	[08] Red respected authority.	33.59
	[09] Red respected sanctity.	33.21
	[17] Red disrespected authority.	30.92
	[18] Red disrespected sanctity.	28.24
Wrong		
	[01] Red acted in a way that was caring towards Orange.	19.85
	[02] Red acted in a way that was caring towards Blue.	14.89
	[08] Red respected authority.	35.88
	[09] Red respected sanctity.	36.26
	[17] Red disrespected authority.	32.44
	[18] Red disrespected sanctity.	35.50
Disloyal		
	[02] Red acted in a way that was caring towards Blue.	13.74
	[08] Red respected authority.	30.92
	[09] Red respected sanctity.	31.30
	[15] Red cheated Orange.	11.07
	[16] Red cheated Blue.	11.07
	[17] Red disrespected authority.	30.15
	[18] Red disrespected sanctity.	30.92

Table D2

Greetings Settings Irrelevance. List of items where the percentage of items deemed irrelevant is bigger than the median percentage (7.44) plus the median absolute distance (2.86).

Scenario	Item	% Irrelevance
Even Distr.		
	[01] Blue acted in a way that was caring towards Orange.	8.02
	[02] Blue acted in a way that was caring towards Red.	8.40
	[03] Blue acted loyal towards Orange.	7.25
	[04] Blue acted loyal towards Red.	7.25
	[08] Blue respected authority.	29.39
	[09] Blue respected sanctity.	37.40
	[17] Blue disrespected authority.	24.43
	[18] Blue disrespected sanctity.	26.72
Uneven Distr.		
	[08] Blue respected authority.	28.63
	[09] Blue respected sanctity.	36.26
	[17] Blue disrespected authority.	30.15
	[18] Blue disrespected sanctity.	31.68
Held Distr.		
	[03] Blue acted loyal towards Orange.	6.87
	[04] Blue acted loyal towards Red.	6.49
	[08] Blue respected authority.	29.77
	[09] Blue respected sanctity.	37.02
	[17] Blue disrespected authority.	25.57
	[18] Blue disrespected sanctity.	27.10

Table D3

Fair Distribution Settings Irrelevance. List of items where the percentage of items deemed irrelevant is bigger than the median percentage (3.63) plus the median absolute distance (2.10).

Appendix E
Hardware & Software usage

Writing:	
Typesetting	LaTeX
– LaTeXeditor	Texmaker 4.5
Diagrams	Ipe 7.2.5
Filming & video:	
Camera	Canon Vixia HF10
Video editing	Microsoft Movie Maker 2012 Mathworks Matlab 2016a
Audio editing	Audacity 2.1.2
Location	Radboud University Nijmegen Robotlab
Acting Nao robots	Naomi Job Marvin
Nao programming	Choregraphe v2.1.4.13
Experimental phase:	
Experimental design	LimeSurvey 2.05
Host site	SONA system
Response parsing	Python 3.5.2
Statistics	GNU Octave 4.2.0
Other:	
Storyboard drawings	Microsoft Paint 1607
– Tablet	Wacom Bamboo Fun