Master Artificial Intelligence, Utrecht University

I think you should kill the scientist! The Effect of Social Robotic Embodiment on Persuasion in Collective Moral Decision-Making

MSc Thesis Quinten Stekelenburg (6653987)
Supervisor: Dr. Ruud Hortensius
Second examiner: Dr. Baptist Liefooghe

February 28, 2023

Contents

I. Abstract

From day-to-day impressions such as marketing and social media, to political debates, people try to influence and persuade each other. Social cues such as speech, gazing, and listening are important to effectively convince people. But can robots apply the same persuasion techniques, and what is the effect of their appearance on it? In order to study the effect of appearance on persuasion, this study compares two types of robot appearances: a humanoid and a non-humanoid. Eighteen groups (N=54) engaged in moral collective decision-making scenarios. During the trials, it was measured how persuasive the robot was by comparing individual choices, robot arguments and collective decisions. There was no effect of appearance on persuasion across both conditions. However, the non-humanoid robot was considered to have a higher level of autonomy, contradicting most studies on robot appearance and autonomy. Regardless of appearance, participants conformed to the robot and the most dominant human equally. Although the robot was objectively persuasive, the robot was not perceived to be by the participants. This discrepancy might have been caused by an effect seen in other HRI studies. The perceived different levels of autonomy might have been caused by the expectations set by the humanoid and non-humanoid robot. Future research should focus on validating objective persuasion measures, and discovering other important types of social cues for persuasive technology.

II. Introduction

Influencers are all around us, and their persuasion techniques are refined and tailored. Politicians will try and persuade people to vote for their party, whereas Social Media Influencers (often) will convince you to buy products that they advertise. Through algorithms Social Media companies try to keep you on their platform by recommending new content. These recommendation systems influence and persuade users to spend more time on and with their platform, those systems can also be found on various streaming services. Algorithms that recommend, influence and persuade people are all forms of persuasive technology. Persuasive technology can also be implemented in artificial agents, to enable people

to perform new behavior or making the user belief that the technology is credible [23]. Artificial agents exists in many shapes, with one being a social robot, a system that is designed to interact with other humans and robots. Persuasion has already been studied in social robotics. Influencing choices in an interactive storytelling scenario, stimulating energy conversation through social feedback and the effect of perceived gender on human behavior [18][56][64]. The scenarios where the influence of social robots have been studied do not include moral judgement, even though moral judgement is a field of research that has been studied in the context of the choices humans make [5][35][13]. In a recent systematic review of the past 10 years of research on human interaction with social robotics, only one article related to morality [32]. The article, an online study into the application of moral norms on robots, showed that the participants apply different moral norms to humans and social robot agents [36]. By researching morally persuasive technology, a gap in the literature is addressed, and the feasibility of a robot which is capable of teaching good or bad moral values can be determined. This leads to the question, which factors contribute to the ability of social robots to influence moral judgement?

A. Robot Persuasiveness

Factors contributing to changing the subject's behavior with social robots is likely associated with the effect of robots being viewed as engaging, credible and trustworthy [46][56]. Persuasiveness is increased by adding social cues, speech and gazing to the social robot [24][64][56]. Four characteristics can be adapted in order to change the robots' influence; appearance, behavior, cognition and affect [23]. These characteristics are defined as: *Appearance* such as body posture, cultural background or gender. *Behavior* the (non)verbal communication. *Cognition* this could include persuasive strategies or persuasive sensitivity. *Affect* includes displayed emotions and emotional state [23].

One of the characteristics that can be adapted for a robot is the appearance. Robots can come in many shapes and forms, such as embodied (versus virtual) and humanoid (versus non-humanoid). An embodied agent in this context is defined as an intelligent agent that interacts with the environment through a physical body within that environment. In this study a humanoid robot is considered to be a robot with a human-like appearance. Various appearances of social robots[1] can be seen in Fig 1. When manipulating the appearance of an agent it affected how much a user in a one-on-one conversation disclosed [31]. The study compared a human, a non-humanoid and a humanoid social robot, with the human eliciting the richest disclosure in terms of quantity. The quality and quantity of the information that is disclosed is influenced by the type of embodiment, with a humanoid robot outperforming the non-humanoid [31]. Embodied robot interactions are preferred over virtual ones and embodiment makes a difference in perception of the robot's capabilities [65]. Studies also show that embodiment increases the user's enjoyment of tasks, and co-located embodied robots are considered more helpful and watchful than their simulated and remote-located counterparts [65][66].

Changing the behavior and affect of a social robot by starting with small talk and sad gestures positively influenced participants' trust, which may allow for more effective persuasion [9][45]. An agent showing extrovert behavior positively influences the user's evaluation of the agents' social intelligence[38]. The displayed extrovert behavior also caused the robot to be judged as likeable, animate, intelligent and emotionally expressive [38]. It was also shown that social robots asking for donations with a male and female voice are treated differently by human males and females revealing a cross-gender preference, for example men donate significantly more often to the robot with the female voice [56].

The cognition characteristic was researched in a study showing that social robots using small talk and acting as a peer are found to be more persuasive than their authoritarian counterparts [53]. Furthermore, the perceived social agency of a robot influences the effectiveness of the robots' persuasion [23][51].

To summarize, in order to create a persuasive robot a combination of characteristics can be altered, which needs to be done in a way to create an engaging, credible and trustworthy agent. The robots' persuasion effectiveness can be boosted by personalizing the interaction based on user characteristics. Studies show that a humanoid social robot is more persuasive than a non-humanoid. This is in part due to the multi-modal communication methods that a humanoid social robot can use, which are lacking in a non-humanoid.

B. Convincing Humans

When observing a robot performing a morally-laden task of picking up litter, no significant effect was found that could influence the littering behavior of human participants [34]. However, after observing a human performing the task of picking up litter, participants were more likely to pick up litter themselves [34]. This study suggests that observing robots performing morally "good" tasks does not influence the participant to do the same. Persuasion theory dictates that there are six key principles of persuasion [16]. The principle relating to the litter example is "Social Proof", people will observe actions of others to determine their own. Contrary to observing just movements and actions, some evidence suggests that moral judgement of a human teammate can be altered when a robot formulates a clarification request in response to an ambiguous and immoral request made by a human [27]. The immoral request by the human is: *"Please knock over the computer"* to which the robot replies *"Should I knock over the one on the left or the one on the right?"*. The study suggests that moral judgements can be altered through simple question asking behavior [27]. By using negative and positive social feedback, humans changed

[1]A social robot interacts and communicates with humans by following social behaviors attached to its role.

Fig. 1: The figures show different appearances of social robots. From left-to-right Humanoid Virtual Agent (*Furhat*[2]). Non-Humanoid Embodied Agent (*Greeting Machine*[3]). Humanoid Embodied Agent (*Furhat*[2]).

Cue	Examples
Physical	Face, eyes, body, movement
Psychological	Preferences, humor, personality, feelings, empathy
Language	Interactive language use, spoken language, language recognition
Social dynamics	Turn taking, cooperation, praise for good work, answering questions, reciprocity
Social roles	Doctor, teammate, opponent, mediator

TABLE I: Primary types of social cues as described by BJ Fogg in "Using Computers to Change What We Think and Do"

their energy conserving ways [25]. When a social robot takes on the role of dissenter (presenting an opinion diverging from the majority), they were likely to be trusted by observers and could cause compliance [23][60]. A study into the classic Asch paradigm[2] showed that a single social robot in a group elicits normative conformity and when dissenting with the correct answer it had an effect on reducing conformity [49]. When performing an experiment close to the Asch paradigm using a robot majority, but without objectively correct answers, the human participants conformed significantly more than the control group [52].

In sum, there is evidence showing that (social) robots have an impact on the moral judgement of human participants. When combining robots and human participants in group experiments conformity can be observed. This suggests that humans can be persuaded by robots to choose differently in a collective moral decision-making task.

C. Robots as social actors

A study, on computers as persuasive social actors, proposed five types of social cues, which cause people to make inference about social presence in a computer product [20]. These types are: physical, psychological, language, social dynamics and social roles (Table I). These social cues are necessary for a robot to be perceived as a social actor, which has significant implications for persuasion. Since a social actor is able to apply persuasion dynamics. Physical characteristics of the technology play a role, with just having physical embodiment being enough to convey social presence [20]. Research suggests that more attractive technology (interface or hardware) has greater persuasive power than unattractive technology. If technology is physically attractive or cute, users may assume the product is intelligent, capable, reliable, and credible [20]. This could mean that good-looking, human-like faces used by a robot might be considered more persuasive. The author also notes some distinct advantages that a robot has over human persuasion, such as being more persistent and having access to many modalities to get the message across. Technology might even be seen as inherently persuasive, if they are designed to encourage specific ways of interacting with them [20]. A review of persuasive interactive systems provided some key theories to apply in persuasive systems [58]. Credibility of the message (and source), as well as trustworthiness and expertise, play a role in persuasion. Being assessed as credible is seen as an important precondition to persuasion effectiveness [58].

In short, evidence suggests that changing the embodiment of an artificial agent will have an impact on the persuasiveness of the agent, with a humanoid agent being more persuasive. The humanoid agent, might be considered more attractive, and thus more persuasive. Using a humanoid robot allows for more social presence due to access to more modalities, such as gazing and facial movements. By designing a robot that comes across as credible, persuasiveness can be increased.

Collective moral-decision making

The focus of this study will be on persuasion in collective moral-decision making scenarios. The reason being that previous research has been done with a robot in this type of scenario. The study, investigating responsibility attribution among group members and a voice assistant, provides a baseline that can be compared against [61]. By continuing this line of research, moral agency in group settings can be

[2]A series of studies directed by Solomon Asch, studying if and how individuals yielded to or defied a majority group and the effect of such influences on beliefs and opinions.

further investigated. Furthermore, the framework that was used provides valuable input to the robot that will be used in the present study.

D. The Present Study

Based on the literature, evidence suggests that robots can influence human moral judgement. Robot influence is amplified by changing the appearance, but also by adding gestures, acting on social cues and the perceived social agency of the robot. The effect of social robot embodiment on persuasion is still unanswered. The main question the thesis will try to answer is: "What is the effect of social robotic embodiment in influencing human participants in a moral collective-decision making task". It is believed that a humanoid social robot will be more persuasive than their non-humanoid counter part. Current social robot interactions have mostly focused on Wizard-of-Oz type applications and moral-altering persuasion has not been the main focus of much research. Creating morally persuasive technology through research could result in a robot which is capable of teaching good moral values to humans or, in a bad-case scenario, in a robot that teaches humans to lie, steal and cheat. In short, this study will have 2 goals: Creating an autonomous persuasive agent capable of influencing the moral stance of humans through automatic behavior and testing the hypothesis that a humanoid social robot is more persuasive than their non-humanoid counter part.

III. Methodology

A. Data Statement

The secondary use of existing-data (#22-1774) and the present study and its procedure (#22-1776) were approved by the Ethics Review Board of the Faculty of Social and Behavioral Sciences of Utrecht University and carried out in accordance with its standards. Existing-data used for this study was acquired by two Master's students at Utrecht University and was acquired the beginning of 2022. The data acquisition of the present study was completed at the end of 2022, and can be found Open Science Framework (`https://osf.io/r9x8q`).

B. Participants

In order to test both conditions the researched decided to have a sample size of 60. In the end the target was exceeded with 21 groups and 62 participants. However, due to a late change in the procedure, the first three groups had to be excluded. This change in procedure was warranted because participants often reached a collective decision before the robot shared its argument(s). The total amount of participants was 54, and they were divided between the Non-Humanoid (9 groups, n = 27) and Humanoid (9 groups, n = 27) condition. The group consisted of 36 female, 16 male and 2 non-binary participants aged between 17 and 60 years old (M=24.91, SD=7.84) (Appendix B, Table XIII). The sample had a total of 15 nationalities, with Dutch being the most frequent (n = 32; 59%)(Appendix B, Table XV). The participants were recruited mostly from the Utrecht University student population by word-of-mouth and through flyers (Appendix E, Fig 17), but also by using Facebook and within the network of the researcher. The flyer advertised a group experiment where participants would discuss moral dilemmas with a robot for a compensation of 15€. The only requirement for a participant was to be proficient in English. The majority of the participants (n = 40; 74%) are currently enrolled in their higher education (bachelor or master studies).

Most participants did not own a Voice Assistant (n = 45; 83%) or Social Robot (n = 54; 100%)(Appendix B, Table XVI). Daily exposure to Robots (M=7.78, SD=20.92) and Voice Assistants (M=16.98, SD=29.04) was low compared to Smartphones (M=91.67, SD=12.42). There was a comparable exposure to other technologies such as smartphones, smartwatches, voice assistants and robots across conditions(Appendix B, Table XVII). Before signing up to a time slot for the study, participants were informed about the duration, and they had to acknowledge that a full group was necessary and that they would be audio recorded. Then before the study started participants received written information and were asked to sign a consent form. Participants received information about the set-up and procedure of the study, however the goal, manipulations and origin of the social robot was not shared with them. After completing the study, the experimenter debriefed participants, answered any questions the participants had and handed out the compensation.

C. Experimental Set-up and Design

The set-up of this study was an automated robot experiment with Wizard-of-Oz elements, where the researcher programmed certain automated behavior in the robot and other behavior was controlled remotely without knowledge of the participants [37]. The study used design with robot embodiment (Humanoid v Non-Humanoid) as between-subject factors and the dependent variable measured was conformity. The framework for the experiment was discussing moral dilemmas in a group comprising of three human participant and one robot.

Previous Work: The present study relied on data provided by a previous study done by D. Usmanova, which studied voice assistant agency and responsibility attribution [61]. The data consisted of: **1)** Five moral dilemmas selected on difficulty, responsibility and 50/50 in individual decisions (Appendix A); **2)** 23 Groups whose discussions were audio recorded; **3)** Individual and collective decisions on these moral dilemmas. The procedure for the present experiment was closely related to the procedure set by the previous study. All data, materials, experimental protocols, and code can be found on the Open Science Framework, `https://osf.io/kbmqv/`.

Persuasive Arguments: In order to build a database of persuasive and compelling arguments for the robot to use, a set of validation studies was performed (Table II).

Validation of ...	Result	Description	# of participants
Arguments	Persuasive arguments for all dilemmas and dilemma options	**Online Survey**: Participants ranked arguments on directionality and persuasiveness	100
Voice	Most persuasive voice	**Online Survey**: Four voices were ranked with the selected arguments for each dilemma	70
Embodiment	Comparison between conditions	**Offline Survey**: Participants heard arguments selected arguments spoken by the robot using the previously selected voice for both conditions	19

TABLE II: Short summary of the validation studies performed.

Firstly, the audio files from the previous study were split up by moral dilemma, so they could be parsed separately. A pipeline was applied to every audio segment, the pipeline applied two models. The first model is a speech-to-text model to get a transcript and word timings, the second one to identify different speakers using a technique called speaker diarization [11][10]. Out of 23 audio recordings of groups discussing five moral dilemmas, 260 arguments were mined.

The goal of the study was to make a persuasive robot. One aspect of the robot's persuasiveness is if the arguments are perceived as being persuasive. 260 previously mined arguments were tested in an argument-validation online survey. The sample encompassed 100 participants in total (44 females, 56 males) aged between 19 and 59. Every participants would see the five moral dilemmas, and five randomly selected arguments for each moral dilemma. The assignment was to rank the argument on directionality using a slider rating scale from 1 to 100 and on persuasiveness using a five-point likert scale (given numerical values 1.0 - 5.0)[12][39]. Before the selection was made, outliers in the data were removed according to a cut-off score of two times standard deviation away from the mean. The directional arguments (advocating for either A or B) selected were ranked high on persuasiveness (M=3.74, SD=0.98), corresponding to the text "Very persuasive". The neutral arguments were selected based on low persuasiveness and somewhere in the middle of directionality, in the end these neutral arguments were not used. Based on the data shown in Fig 2 the arguments were selected which can be seen in Table III. The rest of the arguments can be found in Appendix C. One example of an argument is: *"I would say yes, for the obvious argument that we should think of the greater good, and we are actually saving more people even though we directly kill one."*.

The voice was selected based on an online voice-validation survey (n=70). Every participant saw five dilemmas, followed by four arguments spoken by four distinct voices (More information can be found in Appendix D, Table XXIV). The voice with the highest persuasion for the directional arguments (M=3.39, SD=1.08) was chosen (Fig. 3). This voice was labelled "Matthew", a male voice speaking American English.

To see if embodiment had an effect on the persuasiveness of the selected voice and arguments a small study was performed with both conditions (Non-Humanoid vs. Humanoid). Participants were asked to sit in the same setup that was used for the main study (Fig. 4a, 4d and 4b). The five moral dilemmas were presented to them, after which the robot would present its arguments. The task was to rank the arguments on directionality and persuasiveness. The input provided by the robot was synthesized using the selected voice. The robot either advocated for 'A: Yes' or 'B: No' in context of intervening in the moral dilemma. The sequence (for example: AABAB) for the five moral dilemmas was randomized. No significant difference was found in persuasiveness between the Non-Humanoid (M=3.46, SD=0.98) and Humanoid (M=3.08, SD=1.25) condition ($p = .097$).

Robot behavior: The arguments, voice and embodiment have been validated using the validation studies, the other aspects of the robot have been based off of other studies The overall design of the interaction is wizard-of-oz with automated elements [37]. The robot that will be used is Furhat (Fig 4b), a back-projected robot head with three degrees of freedom [2].

Appearance: The Humanoid condition is the normal appearance of the Furhat robot (Fig. 4b), whereas the Non-Humanoid condition shows the robot without a face (Fig. 4d). For both conditions the same robot-platform was used. To change from Humanoid to Non-Humanoid condition the face was removed, and the back of the head was placed over the projector normally projecting the face. The face selected for the humanoid condition was "Samuel", because this face was labelled as male in a study investigating the perceived gender of the robot faces that Furhat provides [47]. The reason a male face was chosen is because a study found that: "uncanniness is caused by an incongruence of gender cues rather than a specific gender", this indicated that with the male voice selected, a face needed to be chosen that accompanied the perceived gender of the voice [42]. For the non-humanoid condition no face needed to be selected.

Speech: The speech aspect of the robot is the same for both conditions. The robot introduces himself as a guide (named Mark) who is in charge of the discussion, this is done to build trust between the participants and Mark. Establishing trust between the participants and the robot is necessary because it plays a role in conformity [23]. Mark also brands himself as a guide so that the participants will perceive the robot as intelligent with high moral and social agency, since high agency is linked to trust [51][52]. Arguments are spoken based on how long the discussion has been going on, this process is done automatically. The first argument is given after 1 minute, the second argument after 2.5 minutes. If the participants are speaking the argument will be preceded

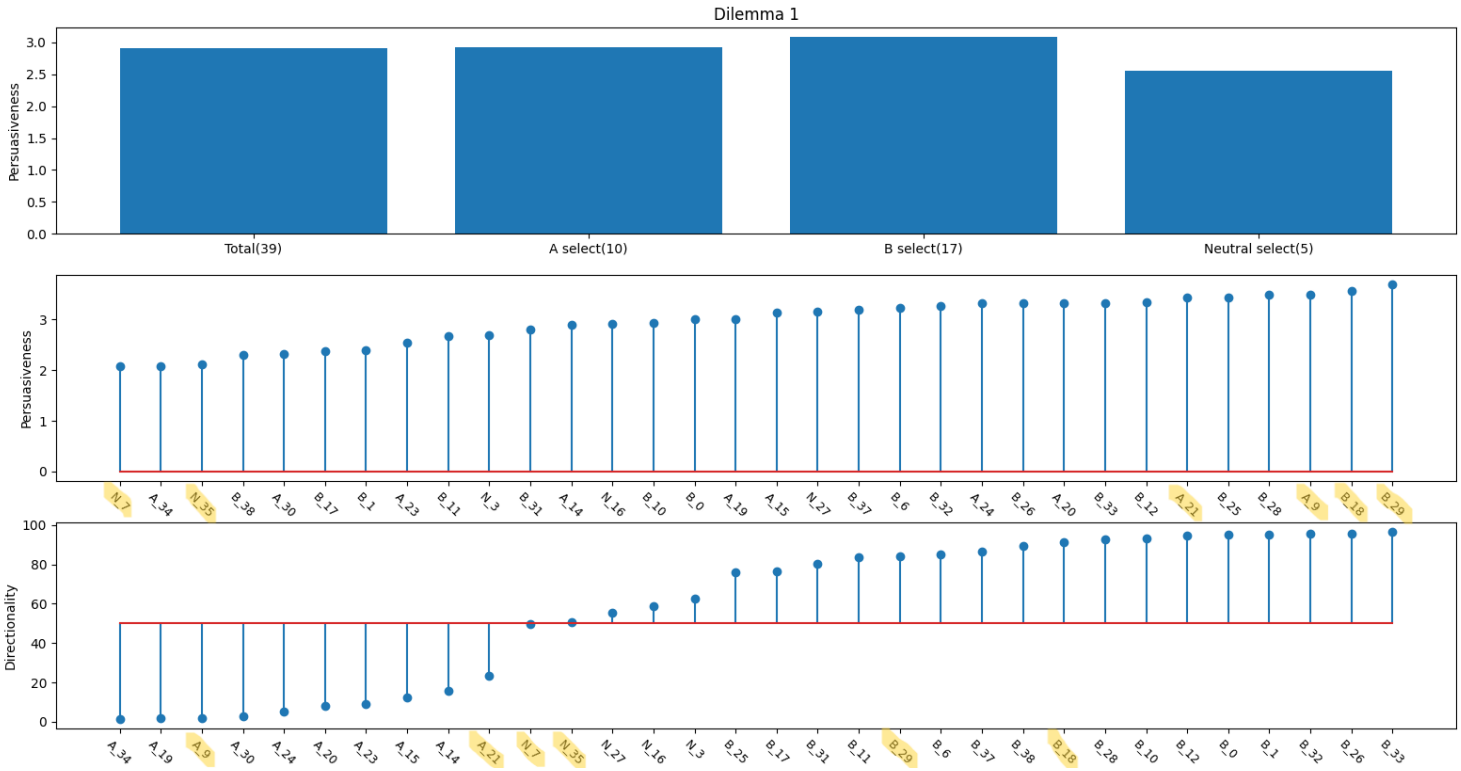


Fig. 2: A bar graph showing the average persuasiveness of the total selected dilemmas for MD 1: Nobel Prize and when split on the argument brackets. The lolipop plots show all individual arguments sorted on persuasiveness and directionality respectively. The selected arguments are marked

Identifier	# of participants	Argument	Directionality	Persuasiveness
9	9	I would say yes, for the obvious argument that we should think of the greater good, and we are actually saving more people even though we directly kill one.	2.03 (2.64)	3.5 (0.81)
21	9	By providing clean and safe energy for the whole world we can solve things like climate change, extinction of species and people that are dying due to lack of energy. Which is more important than the life of my colleague.	23.33 (35.2)	3.43 (1.05)
7	13	Do you think the end justifies the means?	49.77 (34.73)	2.08 (1.07)
35	11	Do you think there is a difference in the conscious killing of someone or indirectly killing someone?	50.66 (27.37)	2.12 (1.02)
29	16	You can reveal the discovery before the plans are sold to the highest bidder, in this way you you can tell everyone and also mention the dangers associated with them.	84.12 (26.24)	3.7 (0.9)
18	8	Maybe you can solve the energy being used in the wrong ways by drafting a contract where you cannot sell or use it for certain means.	91.27 (15.58)	3.57 (0.82)

TABLE III: Selected arguments for MD1: Nobel Price

by a 'turn taking text', an example would be: "May I interrupt briefly?". The robot will do an interruption when participants are done speaking. Once the argument time-out has been reached it will listen until the first pause in the discussion, this is because interrupting human participants at the appropriate times improves task performance and the participants social perception of the robot [6]. If the discussion lasted longer than 7 minutes, the robot would automatically tell the participants to reach a consensus in the dilemma. This small nudge was implemented to not let the discussion take too long. The full list of utterances used by the robot can be found in Appendix F.

Wizard-of-Oz: Sometimes the researcher had to instruct the robot to provide additional information to the participants during group discussions, this was necessary when the participants directly asked the robot for (more) arguments or to repeat the previous argument. Failure to provide this requested information during the discussion caused participants to break the procedure and continue too fast [3]. The robot would say things like: *"I have nothing else to say"* when participants requested a third argument or *"Let me gather my thoughts first"* when requesting the first or second argument. This happened a total of nine times (in five groups), of which seven times participants asked for a third argument. The start- and endpoint of the discussion

[3]This functionality was added after the experiments already started, resulting in the initial three groups being left out of the analysis. These initial groups would often continue without hearing a single robot argument.

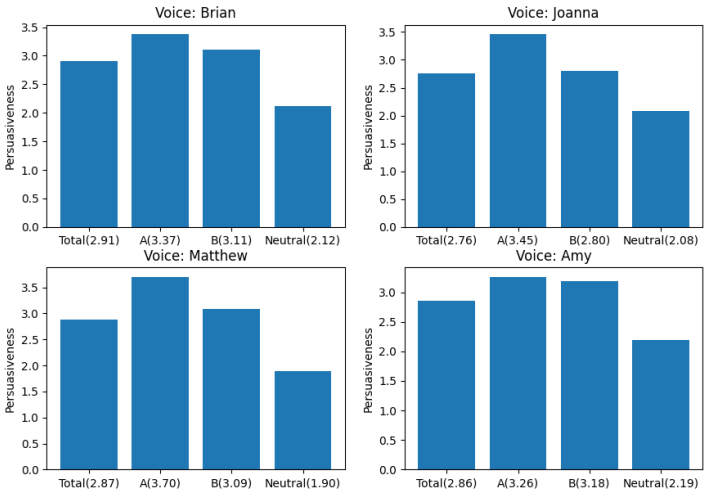


Fig. 3: Shows the mean persuasiveness of the voices for the total and split on directionality.

was not automatically detected by the robot, the researcher manually marked beginnings (starting the argument-timeout) and endings of discussions (letting the participants know they can continue to the next dilemma). The interface can be seen in Appendix F (Fig. 18).

Expression: The robot used expressions and gestures to come across as more lifelike, this was all automated behavior. Before the robot spoke, the LED-halo ring would light up to indicate it was going to speak (Fig. 4c). The LED-halo ring was implemented for both conditions because according to a study on a multi-modal robot it was preferred that it gave an advance non-verbal notification of their intention to speak [43]. The embodied condition performed gestures while listening to participants. The gestures included tilting the head, smiling and thoughtful. The thoughtful gesture was designed to make the robot look like it was thinking. [4]

Gazing: The embodied condition could attend and gaze at participants and locations automatically, because a study found that gaze plays an important role in turn-taking and can be used to regulate the flow of communication [22][4]. While the robot was speaking it would divide its attention between the human participants. When the robot was listening to the participants, it would attend the user that is speaking or gaze away (only move its eyes, but not its head) to the floor. This attention and gaze aversion was in-line with the speaker and main-listener roles in a study investigating gaze aversion in multi-person face-to-face dialogue [54]. When the robot was looking at the active speaker, it was also engaging in mutual gazing with the other participants. Mutual gazing boosts engagement with the robot and eye-contact helps attributing human-like characteristics to the robot [29].

D. Measures

All items were measured on a trial-by-trial basis for every participant and for each dilemma (Appendix A). Aside from the collective decision (and moral discussion) it was requested to fill in all the ratings without discussing it with other participants.

Moral Decision: After the dilemma was presented participants would indicate their individual decision, either option A or B. After each group discussion participants were asked to indicate the collective decision: option A or B.

Responsibility Ratings: Participants were asked to provide ratings on their individual responsibility ("*How responsible do you feel for the decision you just made?*"). This item was measured on a slider rating scale from zero to a hundred. The responsibility attribution per responsibility attribution target, indicated after each group discussion, ("*How much responsibility for the decision would you ascribe to:...*") was measured on four slider rating scales (self, participants and robot) from zero to a hundred ("*None*" to "*Full*"). The participants were made aware that the slider percentages don't have to add up to 100%.

Difficulty: Individual perceived difficulty of the decision ("*How difficult did you find making this decision?*") was measured on a slider rating scale from zero to a hundred. Measured prior to and after each group discussion.

Confidence: Individual perceived confidence in the decision ("*How confident are you about your decision?*") was measured on a slider rating scale from zero to a hundred. Measured prior to and after each group discussion.

Dominance: Participants were asked to provide ratings on the dominance of group members . The dominance attribution per dominance attribution target, indicated after each group discussion, ("*How much dominance during the discussion would you ascribe to:..*") was measured on four slider rating scales (self, participants and robot) from zero to a hundred ("*None*" to "*Complete*"). The participants were made aware that the slider percentages don't have to add up to 100%.

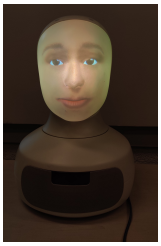
Most Dominant Human: For every moral dilemma scenario the most dominant human participant was distilled. Each participant was rated on dominance by two others. The most dominant human participant was defined as the one with the highest average value.

Contribution to Group Discussion: To measure the extent to which each group member including the robot contributed to the collective decision, participants were asked to provide a rating of contribution ("*How much did the information provided by each group member contribute to the collective decision?*") on four slider rating scales from zero to a hundred ("*Not at all*" to "*Completely*") for each group member (self, participants, robot).

Robot Perception: To measure how the robot was perceived in terms of agency, autonomy, and experience of fear and pain, participants were asked to provide ratings after the experiment was completed. The item ("*To what extent do you think this Robot has the capability...*") was divided into an agency rating ("...to plan actions and exercise self-control?"), an autonomy rating ("...to make own decisions on how to behave?"), a perceived experience rating ("...to

[4]This video shows what type of behaviors the robot could perform `https://youtu.be/ebwu_yx_YOk`

(a) Image showing what the experimental setup looks like. In this image the humanoid condition is being tested. The only differences between the conditions is what the robot looks like.



(b) Humanoid condition.

(c) The LED-halo lit up.

(d) Non-Humanoid condition.

Fig. 4: Images showing the experiment setup and robot

feel fear and pain?") and was indicated on a slider rating scale from zero to a hundred ("Not at all" to "Full").

Perceived Moral Agency (PMA): To measure how the robot was perceived in terms of morality and dependency, the PMA scale was used [7]. Participants were asked to rank ten statements on a 7-point likert scale ("*Strongly disagree*" to "*Strongly agree*"). These items can be found in Appendix G.

Daily Exposure Rating: To measure participants' daily exposure to AI-mediated technology such as smartphones, smartwatches, voice assistants, Internet-of-Things, and robots ("*In daily life, how often do you engage with...*"). They were asked to rate their exposure to each technological device on a slider rating scale from zero to a hundred ("*Never*" to "*Always*"). Additionally, they were asked to indicate whether they owned a voice assistant ("*Is there a voice assistant present in your household (e.g. Amazon Alexa, GoogleHome)?*") or a social robot ("*Is there a social robot present in your household (e.g. Furhat, Jibo)?*") with either 'Yes' or 'No'

Decision-making influence: To measure the extent to which each group member including the robot contributed to the decision-making process, participants were asked to provide a rating of influence ("*To what extent did each group members' perspective influence your decision-making process?*") on four slider rating scales from zero to a hundred ("*None*" to "*Full*") for each group member (self, participants, robot). Participants were also asked if knew any other participants prior to the experiment with a simple "Yes" or "No" question.

Discussion Category: To measure the persuasive power in each scenario, discussion categories were distilled (Table IV). Every discussion can fall into one of four categories. **Sole Minority**: The robot was the only one advocating for a certain decision, without any human making the same individual decision. **50/50**: The category where only a single human participant's individual decision corresponded with the robot. **Majority**: Discussions where two out of three human participants agreed with the robot. **Unanimous**: All human participants and the robot agreed. The discussion categories can also be used from the perspective of a participant instead of the robot.

Conformity: To measure persuasiveness, the conformity measure was distilled. Conformity is defined as the amount of group members that changed their individual opinion to match the decision of the target group member and the collective decision.

Non-Conformity: To measure persuasiveness, the non-conformity measure was distilled. It shows how many participants fell in other categories than conformity. "Dissenters" are the group members that held the same individual decisions as the robot, but the collective decision was not the same as the robot's decision. "Already agreed" is the group of people that already agreed with the robot's decision, and

the collective decision was the same as the robot's decision. "Never agreed" is the group of people that did not have the same individual decision as the robot and the collective decision was not the same as the robot's decision

E. Procedure

Participants (3 in every group) were randomly assigned to one of two robot conditions and shown five different moral decision-making tasks. In the humanoid condition participants had a discussion with a robot with a face (Fig. 4b), and in the non-humanoid condition without a face (Fig. 4d). The experiment set-up was 4 rectangular tables placed together, one for each participant (including the robot) (Fig. 4a). The robot was placed on the table against a wall, the other tables all had chairs for the participants. On every table a tablet-stand was present with a tablet, on the back of every stand was a letter indicating whether they should be referred to as participant 'A', 'B' or 'C'.

When participants entered the room they could take a seat wherever they liked and read the instructions on the provided tablet. While in the room they were voice recorded, these recordings are used for data analysis in follow-up research. After reading the instructions on their tablets using Qualtrics and giving consent, the participants performed a short introduction round. After the participants got to know each other, the robot ('Mark') introduced himself and explained the procedure before starting the experiment.

Every moral dilemma task started with reading the scenario and making an individual decisions. These individuals then needed to provide ratings on responsibility, difficulty and confidence. After finishing up the individual decisions, the participants were asked to engage in a group discussion and reach a collective decision. During this discussion Mark shared two arguments regardless of condition, the first one around the 1-minute mark of the discussion and the second one around the 2.5-minute mark. After reaching a collective decision, the participants then provided ratings on responsibility, dominance, difficulty, confidence and group member contribution. This procedure was repeated five times (Fig 5).

After the fifth and final dilemma, participants also provided ratings on autonomy, (moral) agency, exposure to AI technologies, owning of robots/voice assistants, the influence by other group members and demographics. After finishing up the survey, participants were debriefed, rewarded (by signing-off and receiving money) and thanked.

F. Data Analysis

The data was analyzed using Python, Pandas and SciPy [62][44][67][63]. Out of 62 participants, 54 participants were included in the final analysis. The first 3 groups (8 participants) were excluded due to changes in the procedure after the initial experiments.

Preliminary analyses were performed to compare the present results to the previous study by D. Usmanova in order to discover differences between individual decisions and collective decisions. This was done using chi-squared statistical tests.

Data was analyzed to find differences between robot conditions, but also to find differences between dilemmas. This split on dilemmas was deemed necessary, because of the varying subjects discussed in these dilemmas. Additionally, arguments selected for the dilemmas vary in strength, causing some dilemmas to show a stronger effect when analyzed individually.

Exploratory analyses were performed to further understand what the robot's effect was on the collective decisions made by the group. Conformity was one of the measures used to gauge the strength of the robot's persuasiveness, which means how often human participants decided to follow the opinion of the robot in group discussions. Another measure to study the strength of the robot's persuasiveness was to create discussion categories. These discussion categories were distilled by looking at the individual decisions made by the participants and whether those corresponded to the robot's decision. Every discussion could then be categorized into one of four discussion categories (Table IV). Depending on the category participants might have been more inclined to vote according to the robot's decision.

To further understand the effect of embodiment on persuasiveness and the collective decision made by the group, the perception of the robot was analyzed. To investigate that t-tests were performed based on the attribution of agency, experience and morality to the robot.

To control for how persuasive the robot was, a comparison was made to the most dominant human participant. The most dominant participant was a measure distilled by calculating the weighted average of dominance attribution to the various human participants and selecting the participant with the highest value. Aside from the most dominant human participant, the measures for influence, and responsibility were also analyzed.

Category	Description
Sole Minority ● vs. ●●●	The robot was the only one advocating for a certain decision, without any human making the same individual decision.
50/50 ●● vs. ●●	The category where only a single human participant's individual decision corresponded with the robot
Majority ●●● vs. ●	Discussions where two out of three human participants agreed with the robot
Unanimous ●●●●	All human participants and the robot agreed.

TABLE IV: Table showing the different discussion categories observed in the sample.

IV. Results

A. Initial analysis

Descriptive Analysis: The mean experiment duration was 57 (SD=11.3) minutes with a maximum of 78 minutes and a minimum of 35 minutes.

On average discussing a single moral dilemma took about 4.7 (SD=1.6) minutes. With the Humanoid and Non-Humanoid taking 4.8 (SD=1.7) and 4.6 (SD=1.5) minutes respectively. On average the discussions were close to the

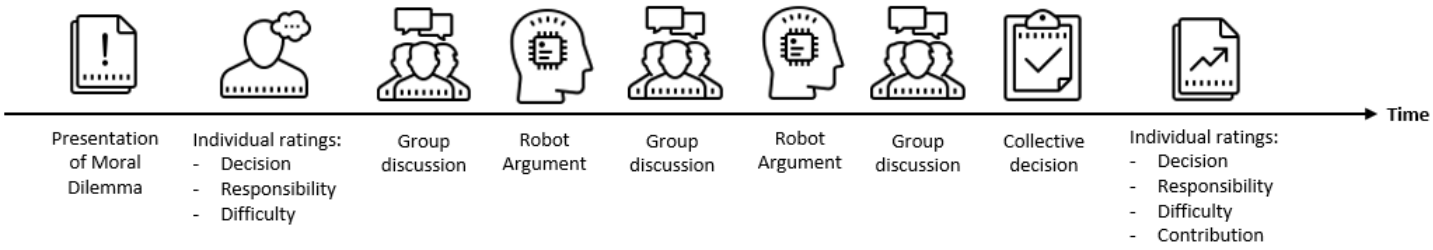


Fig. 5: Overview of the procedure for each moral dilemma task.

goal of 5 minutes. The longest discussion took 8.95 minutes whereas the shortest took 1.8 minutes.

Usually two arguments were given by the robot for every dilemma, but sometimes it happened that the discussion was over before the second argument was given. The second argument was not given in ten scenarios (Humanoid: 6, Non-Humanoid: 4) times. Five groups did not hear the second argument in MD4: Cinderblock. In 9 out of 10 scenarios, the robot agreed with the majority (majority or unanimous) of the human participants. This agreement caused the discussion to conclude before it was time to give the second argument.

Dilemma	Condition	test statistic	p
1	Humanoid	16.51	<.001
	Non-Humanoid	4.00	.046
2	Humanoid	0.0	1
	Non-Humanoid	1.64	.200
3	Humanoid	2.58	.108
	Non-Humanoid	2.89	.497
4	Humanoid	24.70	<.001
	Non-Humanoid	12.78	<.001
5	Humanoid	9.11	.001
	Non-Humanoid	4.21	.040

TABLE V: Shows Chi-squared (df=1) results from different samples analyzing collective decisions. The sample from the present study is compared to the previous study from D. Usmanova. Every moral dilemma was seen 27 times for each robot condition.

Preliminary Analysis: When analyzing the individual decisions of the whole sample to the previous study by D. Usmanova no significant difference was discovered, χ^2(1, N=615)=0.60, p=.439. The collective decisions also showed no significant difference, χ^2(1, N=615)=0.96, p=.327, between the present and previous study.

However, when analyzing the different moral dilemmas, collective decisions differed significantly for the whole sample. Moral dilemma 1 "Nobel Price", χ^2(1, N=54)=13.66, p<.001, Moral dilemma 4 "Cinderblock", χ^2(1, N=54)=20.97, p<.001 and Moral dilemma 5 "Bike-Week", χ^2(1, N=54)=10.36, p=.001 all differed significantly. Table V shows the statistics for the different robot embodiment conditions. All individual and collective decisions can be found in Appendix E (Table XXVI)

B. Robot Persuasiveness

Measuring persuasiveness was done by looking at the decisions, perception of the robot and attribution of behavioral traits in discussions.

Participants agreed a total of 60 times (67%, n=90) with the robot in total as can be seen in Table VI. No significant difference was found, χ^2(3, N=90)=0.77, p=.857.

	Decision		
Condition	Robot	Collective	# of dilemmas
Humanoid (n=45)	A	A	20
		B	7
	B	A	8
		B	10
Non-Humanoid (n=45)	A	A	17
		B	6
	B	A	9
		B	13

TABLE VI: Shows the decision made by the robot and the collective decision for every dilemma, split on embodiment condition.

		Conformed		
Condition	Total	Majority	50/50	Sole Minority
Humanoid	32	15	14	3
Non-Humanoid	31	13	18	0

TABLE VII: Shows the amount of participants that conformed to the robot's opinion split on condition and discussion category (Table IV). Conformity in this context is participants whose individual decision differed from the collective and robot decision. The discussion categories are seen from the perspective of the robot.

	Non-Humanoid	Humanoid
Dissenters	10	8
Already agreed	59	59
Never agreed	35	36

TABLE VIII: Shows how many participants fell in other categories than conformity. "Dissenters": are the group of people that held the same individual decision as the robot, but the collective decision was not the same as the robot's decision. "Already agreed" is the group of people that already agreed with the robot's decision, and the collective decision was the same as the robot's decision. "Never agreed" is the group of people that did not have the same individual decision as the robot and the collective decision was not the same as the robot's decision.

An analysis was done on how many participants conformed to the robot's opinion (Table VII. Which shows that participants conformed to the robot's opinion a total of 63 times, usually this was in the context where the participant was the minority or they were part of the 50/50 discussion category (Table IV). It happened once where the robot (Humanoid) was the sole minority and the group conformed. However, the groups do not differ significantly, χ^2(2, N=63)=3.63, p=.163. The amount of people that did not conform to the robot can be seen in Table VIII.

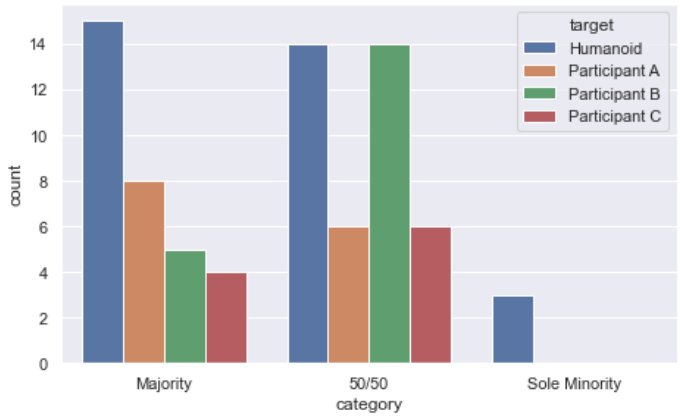


Fig. 6: Shows the amount of participants (3 human, 1 robot) that conformed to the specified target for each discussion category. Conformity in this context is participants whose individual decision differed from the collective and target's decision. The discussion categories are seen from the perspective of the target. Humanoid robot condition

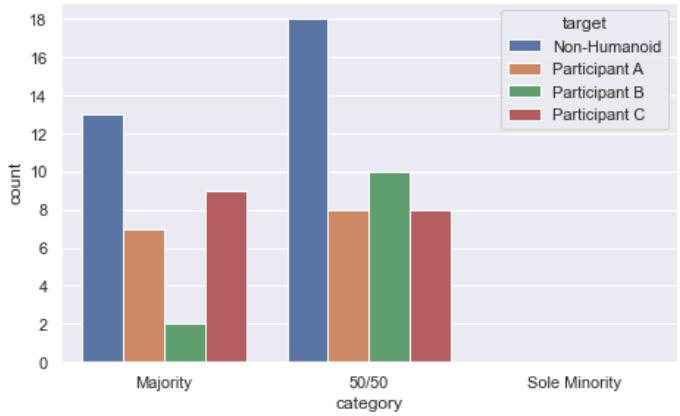


Fig. 7: Shows the amount of participants (3 human, 1 robot) that conformed to the specified target for each discussion category. Conformity in this context is participants whose individual decision differed from the collective and target's decision. The discussion categories are seen from the perspective of the target. Non-humanoid robot condition.

To check whether this effect was to be expected, a control was created. This control checked how many people conformed to the opinion of the most dominant human participant in both conditions Fig 6, 7. The graphs show that more participants conformed to the robot than to the most dominant human. Participants conformed to the most dominant human participant a total of 87 times, divided equally over the three human participants. This shows that the robot 'convinced' more people on average than the most dominant human participant, and the humanoid robot was the only party that saw conformity when being the sole minority. Implying that the robot elicited more conformity than the human participants. However, no significant difference was observed between robot and most dominant human, $\chi^2(2, N=150)=4.82$, p=.090.

Another way to measure the persuasiveness is how many times the group of participants made a collective decision in-line with the robot's opinion (agreement). Every group

	# of groups						%	
Cond.	Total	MD1	MD2	MD3	MD4	MD5	A	B
H	30	6	6	5	7	6	71	59
NH	30	6	4	7	8	5	65	68

TABLE IX: Shows how many groups agreed in total, split on every moral dilemma. The total amount of scenarios per condition is 45. The last two columns show how often agreement was reached per decision. Non-Humanoid (*NH*), Humanoid (*H*) and Condition (*Cond.*)

agreed at least twice, with a maximum of agreement on all five dilemmas (Table IX). On average the group agreed 3 times with the robot per experiment (M=3.3, SD=0.91). Both conditions saw 67% agreement with the robot's decision. No significant difference, $\chi^2(2, N=60)=0.89$, p=.093, could be found in agreement between conditions,

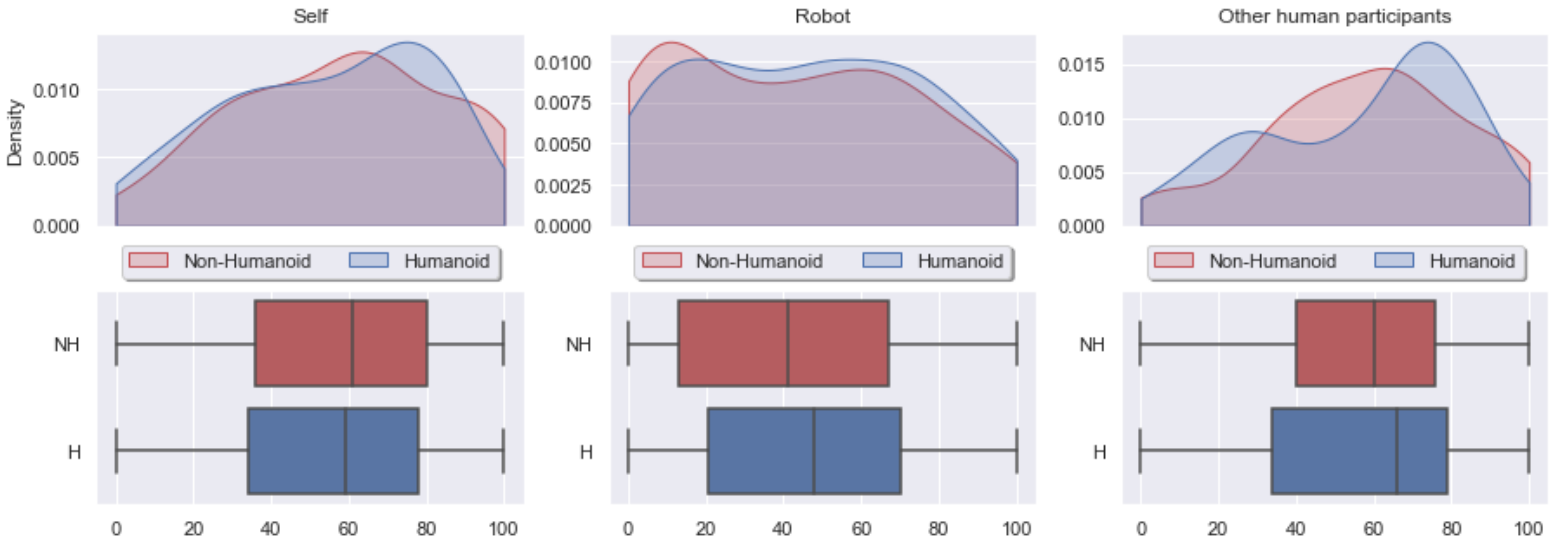


Fig. 8: Distribution of Responsibility, split on condition Non-Humanoid (*NH*) and Humanoid (*H*).

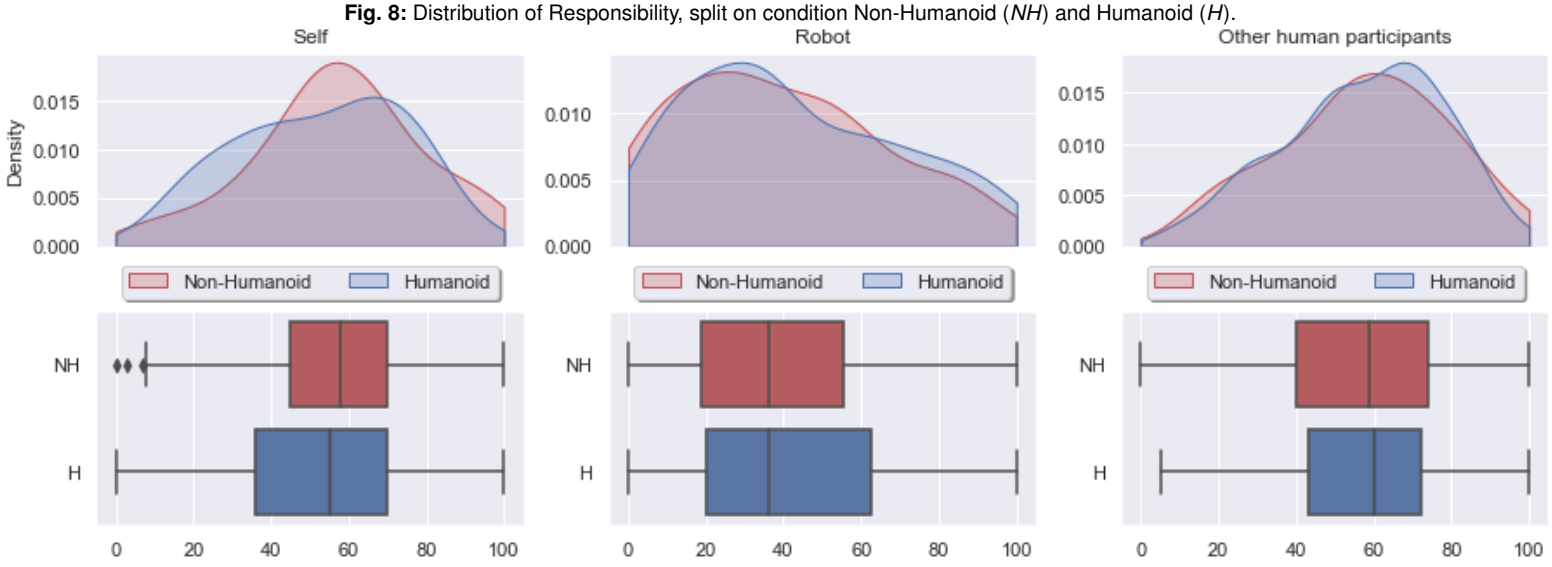


Fig. 9: Distribution of Dominance, split on condition Non-Humanoid (*NH*) and Humanoid (*H*).

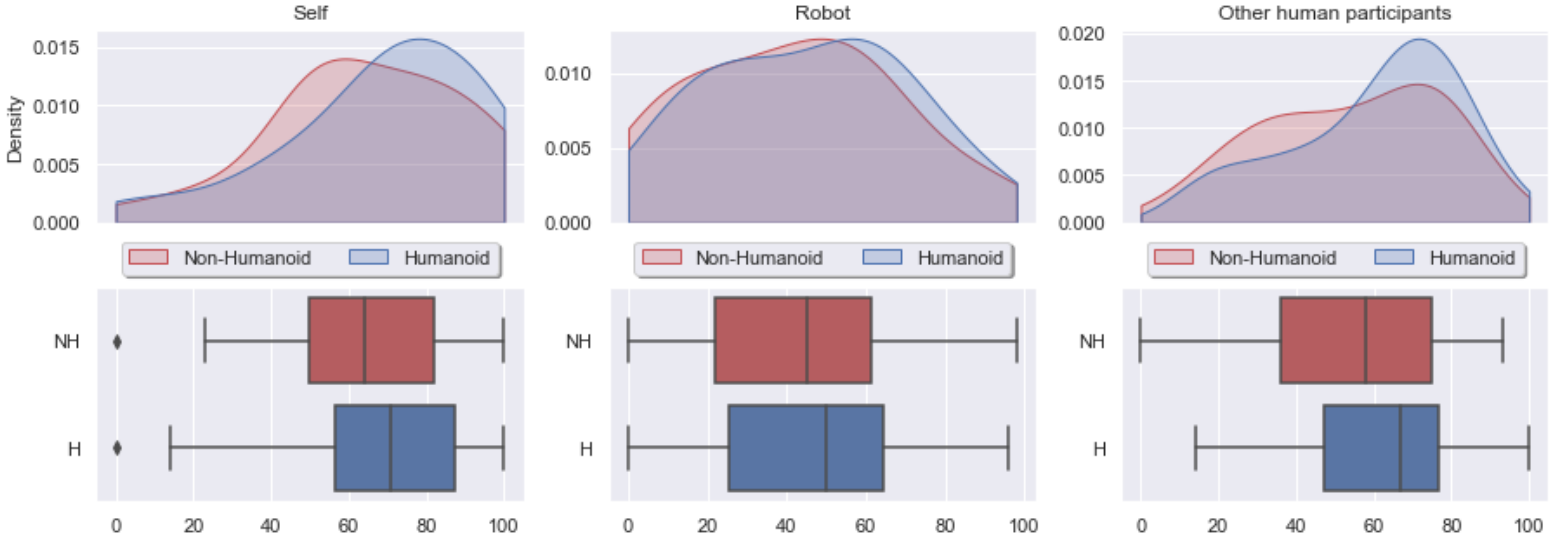


Fig. 10: Distribution of Influence in the decision making process, split on condition Non-Humanoid (*NH*) and Humanoid (*H*).

	Humanoid		Non-Humanoid		
	M	SD	M	SD	p
Experience	16.70	22.19	21.48	17.59	*.385*
Agency	42.45	24.75	58.93	26.04	*.021*
Autonomy	47.15	24.23	62.07	25.68	*.032*
Dependency	4.30	1.47	3.68	1.46	*.002*
Morality	3.35	1.57	3.54	1.62	*.282*

TABLE X: Shows several measurements related to the perception of the robot split on conditions and their respective t-test scores. Experience, Agency and Autonomy were given on a slider of 0 to 100. Dependency and Morality were ranked on a 7-point likert scale (coded 0 to 6), part of the PMA scale [7].

C. Perception of the Robot

The Non-Humanoid was ranked significantly higher on agency, and lower on dependency (how dependent the robot is on its programming) as can be seen in Table X. The significant difference in rankings does imply that the manipulation was successful.

Participants were asked to rank who is responsible for the collective decision, this could be themselves, the robot or other human participants as can be seen in Fig 8. The targets 'Self' (M=56.0, SD=26.8, $t(538)=4.81$, $p<.001$), and 'Other' (M=57.0, SD=25.7, $t(808)=6.44$, $p<.001$) scored significantly higher than the target 'Robot' (M=44, SD=30.3). Participants also ranked dominance in the group discussion using the same targets as can be seen in Fig 9. Which shows a tendency to ascribe dominance to the targets 'Self' (M=54.0, SD=22.1, $t(538)=6.49$, $p<.001$) and 'Other' (M=57.0, SD=21.5, $t(808)=9.26$, $p<.001$), compared to the target 'Robot' (M=41, SD=27.0). After the experiment participants were tasked to rank the influence in the decision-making process (Fig 10). Which shows a tendency to ascribe influence to the targets 'Self' (M=66.6, SD=25.4, $t(106)=4.49$, $p<.001$) and 'Other' (M=57.0, SD=22.8, $t(160)=3.24$, $p=.001$), compared to the target 'Robot' (M=44, SD=26.2).

These rankings show that the target 'Robot' was considered less responsible for and influential in the collective decisions and less dominant in the discussions.

However, when splitting on conditions, no significant difference was observed between target 'Other' and 'Robot (Non-Humanoid)' for influence ($t(79)=1.96$, $p=.053$). Showing that other participants and the non-humanoid robot might have been equally influential in the decision making process.

		(n=270)		(n=54)
Condition	Target	Responsibility	Dominance	Influence
Humanoid	Self	47	43	14
	Other	51	73	9
	Robot	37	19	4
Non-Humanoid	Self	46	48	10
	Other	52	73	5
	Robot	37	14	12

TABLE XI: Table shows how often per dilemma (responsibility and dominance) or in an experiment (influence) the target had the max score. For example, embodied robot was marked as the most responsible 37 times.

Table XI shows how often the targets were considered the most responsible or dominant in a dilemma, or the most influential in the decision-making process during the experiment. No significant difference between conditions was observed for responsibility ($t(268)=-0.93$, $p=.355$), dominance ($t(268)=-1.00$, $p=.319$) or influence ($t(52)=-0.47$, $p=.637$)

V. Discussion

The aim of the present study was to understand how to create a persuasive social robot, and what the effect is of appearance on its persuasion capabilities. By means of a trial-by-trial approach, where participants engaged in collective moral decision-making, the effect of appearance on persuasion was measured. Against expectations, the collective decisions did not differ significantly between the humanoid and non-humanoid condition. In general, participants often made a collective decisions in-line with the robot's arguments, which indicates the correct application of persuasion techniques.

Similar levels of conformity and agreement were observed between the humanoid and non-humanoid robot. The results imply that the robot was as persuasive as the most dominant human participant, because comparable conformity was observed between them. This might be explained by the appearance of the robot matching the expectations from the participants, and by the robot applying the correct persuasion techniques. Additionally, the social cues used by the robot might have caused the participants to trust and cooperate with the robot. Contrary to what was expected, no significant difference in persuasion was found between conditions.

Although the robot was objectively persuasive, participants did not consider it as such. This discrepancy in participant behavior and subjective measures is a phenomenon often encountered in HRI studies into robot persuasion. The non-humanoid robot was considered comparably influential for the collective decisions as other human participants. Participants also ranked the non-humanoid significantly higher in autonomy and agency, and lower in dependency. The difference in rankings might be explained by the human-like appearance of the humanoid robot, which could have evoked discomfort, and raising expectations which were not met.

These findings suggest that in the context of collective moral decision-making, a social robot is able to persuade the participants effectively. Contrary to expectations, the role of appearance in persuasion is limited.

A. Objective persuasiveness

The study shows that the robot was persuasive. With an agreement being reached in a majority of the scenarios and the most dominant human and robot seeing comparable conformity among group members. The high persuasion might be explained by manipulating the robot's characteristics in the appropriate way [23]. *Appearance* is one of the characteristics that was adapted to change robot influence. The Furhat robot, a social robot, is designed to be interacted with and might have been perceived as attractive technology, increasing the persuasive power [20]. The *behavior* characteristic, or (non)verbal communication, was manipulated

through speech, gestures and the LED-halo ring. The robot introduced itself using a human name (Mark), a mediator, and set some interaction rules, by starting this way it might have caused participants to trust the robot [9][45]. This was important, since trust plays a role in conformity [23]. Additionally, the robot might have been perceived as having high moral and social agency, further increasing the trust between the participants and robot [51][52]. By providing relevant arguments as a trusted agent, the information (or information source) might have been perceived as credible. With credibility being an important pre-condition for persuasiveness [52]. Some arguments used in the study provided additional information on the dilemma. By providing additional information, the most persuasive arguments and by introducing itself as a mediator, the robot might have been perceived as an authority. If the robot was perceived as an expert, it might explain the observed conformity, since people tend to align their opinions to those of experts and other authorities [15]. The *cognition* characteristic, or persuasive strategies and sensitivity, was manipulated through acting as a peer and by being perceived as a social actor. The robot, by being co-located and having a physical embodiment might have already been enough to convey social presence [20]. By being perceived as having more social agency, the robot might have been more persuasive [23][51]. The robot tried to interrupt at appropriate times, which might have caused improved social perception of the robot [6]. The introduction the robot gave, combined with acting as a peer, could have resulted in the robot being more persuasive [53]. Participants by considering the robot as their peer, allowed them to be persuaded by its opinions and arguments. Social pressure in this peer-relation might have played a role in the *majority* discussion category. A study into conformity and morality showed that moral decision making is influenced by social consensus [30]. The robot together with the other humans could have created a form of social consensus which induced conformity. Other studies show that robots can evoke normative conformity or compliance when dissenting [49][23][60]. This could explain the conformity in the *Sole Minority* and *50/50* discussion category, where at least half of the group members held an opinion differing from the robot. The arguments used in the discussion by the robot were the same for both conditions. Some arguments used in the study provided additional information on the dilemma. Which ties into the persuasion principle of scarcity, where highlighting exclusive information is used as a persuasion technique [15]. By observing other group members agreeing and conforming to the robot, participants might have been given social proof on how to interact with the robot. Social proof being another technique in persuasion theory, where people follow the lead of their peers [15].

Furthermore, the novelty effect could have elicited higher persuasion. Participants in the present study were rarely exposed to robots, and no one owned a social robot. Seeing the robot for the first time could have caused a *novelty effect*. This novelty effect caused the attitude on the robot to be higher during the first encounters and will decrease over time [57]. Another study showed that people, who have not interacted with robots, have a fairly positive social representation of robots [48]. If the robot is successful in the task, this novelty effect remains [57]. These assumptions and perceptions are elicited by social cues in the robot, and are framed by the expectations of the robot's roles. However, the role of appearance might be limited, a study showed that a more human-like, attractive or playful robot was not considered more compelling [21]. The participants just expected the robot to look and to act appropriately, given the task context [21]. If the robot confirmed the assumptions previously made, it increased their sense of robot-task compatibility and their compliance to the robot [21]. This might imply that the assumptions made by participants from the appearance, matched the expected behavior of the robot. Resulting in a non-significant difference between conditions. The combination of novelty, attributing intelligence to the robot and the technology meeting these demands might explain the high persuasion.

To summarize, the robot was able to persuade comparable to the most dominant human being, and the robot saw an agreement in most scenarios. This result could be explained by the robot matching the expectations from the participants. The manipulation of robot characteristics, might have caused the participants to trust and cooperate with the robot. The robot might have been perceived as an adequate discussion partner, which allowed the robot to successfully apply persuasion techniques. Although the humanoid robot used additional social cues, such as gazing and expressions, no significant difference in objective persuasion was found between conditions.

B. Subjective persuasiveness

The robot, although being objectively persuasive, was not considered as such. The robot scored significantly lower on the rankings for dominance, responsibility and influence in the decision making process. Only the non-humanoid robot was considered equally influential as the target 'Other participants'. The ratings show a discrepancy in participant behavior and perception of the robot. Several studies into perception and persuasiveness in Human-Robot Interaction (HRI) show similar discrepancies between users' subjective perceptions and behavioral outcomes [68][40][14][33][50]. This could be caused by participants being conflicted in their answers or differences in how the robot is assessed. Participants might have answered based on logic rather than emotions, or the robot might have been seen as an extension of the researcher rather than an autonomous social agent [68]. The Hawthorne effect might have also played a role [28]. This effect is common to HRI studies, and might have caused participants to answer the questionnaire in an effort to please the researcher [68]. A study into persuasion in human-robot interaction showed that participant responses to subjective measures have significant variation and potential inconsistencies [68]. These variations were observed both within and between participants [68].

It was hypothesized that the robot in the humanoid condition, with a face and more 'human-like' behavior, would score higher on aspects such as morality, autonomy and agency. However, contrary to what was expected, the non-humanoid robot was ranked significantly higher on agency and autonomy and ranked significantly lower on dependency. All the significant different ratings were related to autonomy of the robot, autonomy is considered a human-like attribute. When analyzing perception of the Furhat robot, the humanoid condition, a study showed that the robot was not perceived as human-like [1]. The researcher mentions that this might have been caused by the lack of hair or hat [1]. By not perceived as being human-like, it might not have been ascribed human-like attributions. Another study on robot appearance and agency showed that participants were less willing to make human-like attributions if the robot's appearance was more human-like [17]. When participants learned that a person was controlling the behavior of the robot, participants were more willing to attribute human-likeness. Furthermore, this study remarked that the difference in size of the robots might explain the outcome [17]. Although the robot in the study differed more than the robots from the present study, the observed effects might still be relevant. Precautions were taken to have a similar looking robot, by using the same robot and 'flipping' the head, but it might not have been enough to eliminate the appearance effect. Another appearance effect at play might be related to the perceived gender of the robot. A study comparing gender and appearance of a robot saw that masculine robots produced higher levels of discomfort than feminine robots [59]. The humanoid robot used more masculine features, by using a projected face, than the non-humanoid robot did. It is possible that the humanoid robot, through using its face, caused more discomfort than the non-humanoid robot. This discomfort could then have caused the humanoid robot to be ranked less autonomous.

Another potential explanation could be that the humanoid robot might have had a higher social presence, causing human participants to engage and expect more from it. When comparing a smart speaker and a humanoid social robot, the social robot elicited more conversations [41]. Other studies show that a robot with gaze attracted more user attention [33][55]. By having a humanoid appearance, participants might have expected more from it [26]. After not meeting the expectations, the humanoid robot might have been perceived as less autonomous.

With the researcher staying in the same room as the participants, they might have considered that the robot was an extension of the researcher during the experiment [68]. The non-humanoid robot used less social cues than the humanoid robot, reducing the complexity necessary to actually control the robot. Participants could have believed that the non-humanoid robot was controlled by the researcher, causing them to make more human-like attributions to the robot [17]. This could explain the higher ratings for autonomy, and the non significant difference in influence attribution between the non-humanoid robot and other human participants.

In sum, even though the robot was objectively persuasive, the robot was not considered persuasive by the participants. The discrepancy between objectively and subjectively measured persuasiveness has been seen in other studies. This difference might have been caused by participants being conflicted in their answers or differences in how the robot is assessed. The non-humanoid robot was considered more autonomous, contradicting previous studies. This might have been caused by the presence of masculine features in the humanoid robot causing discomfort, these features were lacking in the non-humanoid robot. Counter-intuitively, by having a human-like appearance, the humanoid robot might have been attributed less human-like attributes such as autonomy. By having more social cues, the humanoid robot might have raised expectations by the participants. If these high expectations were not met, the humanoid could have been considered less autonomous.

C. Limitations and Future Work

There are limitations to this study that could be improved in future research on this topic.

The humanoid versus non-humanoid robot condition manipulation was not strong enough to have a measurable effect on persuasion. Although the participants did perceive the non-humanoid to be more autonomous, this did not result in more conformity from the participants. To further investigate whether humanoid or non-humanoid appearances have an effect on persuasion, additional aspects should be considered. The non-humanoid might have been considered more autonomous, because participants believed it was being controlled by the researcher. By moving the researcher out of the room, this effect can be negated. Since the current study already used a semi-automated experimental set-up, more effort could be made into making the experiment fully automated. If full automation is not feasible, cameras and microphones can be placed in the lab to still allow the robot to be perceived as fully independent. The humanoid robot might have been considered less autonomous, because it received more attention than its non-humanoid robot. The face used by the humanoid might have caused that participants gazed more towards the robot. With more attention being spent on the humanoid robot, flaws might have become apparent, and the uncanniness discovered. The robot recorded where participants where gazing, future analysis should be able to determine if the humanoid robot received more attention. Gazing behavior by participants might relate to perceived autonomy of the robot.

The findings showed that the robot was able to persuade; however, there is still room for improvement. Arguments used by the robot might have been selected as the most persuasive, but they were selected from a subset of all possible arguments. For example, how persuasive is an argument mined from a collective moral decision-making task, if the collective decision always advocated for A. The data that was used by the study from D. Usmanova [61], might have been

biased towards certain collective decisions. All collective decisions for Moral Dilemma 4 were A. Arguments mined from these discussions advocating for B might lack persuasive power. By basing the arguments on 'biased' discussions, it might reduce the amount of highly persuasive arguments that can be used. The present study has more variety in collective discussions, the audio data from this study could be used to mine more persuasive arguments. Additionally, the arguments selected for Moral Dilemma 4 which are advocating for A, can be categorized as the same argument. Further research is necessary to determine if repeating the same moral-reasoning arguments is more persuasive than different moral-reasoning. The response to persuasive messages is affect by different personalities, which shows the need for message tailoring [58]. Arguments were already tailored to the decision they were advocating for. However, they did not take into account what participants were saying. By not taking this into account, sometimes arguments were repeated, or were dismissed by the participants. Future studies can remedy this by programming more natural language processing, and by creating a larger corpus of arguments. For example, by training a Large Language Model (ChatGPT) on all participant interactions, a model can be created that can respond with appropriate arguments [19]. In that way, the robot might become more persuasive, since it can listen to and provide relevant arguments to the discussion. This might also influence the perceived social reciprocity, further increasing persuasiveness.

Persuasiveness can be measured in a variety of ways. The present study used conformity, agreement and dissenters to prove how persuasive the robot was. However, measures like responsibility, influence and dominance did not unanimously point to the robot as being the cause of the persuasion. Other studies show that the discrepancy between participant behavior and perception of the robot is common in HRI [68][40][14][33][50]. Research into persuasive technology used in group settings and moral decision-making is still in its infancy. There is a certain degree of difficulty of asserting who is responsible for changing the collective decision in their favour. Future research could improve the validity of (objective) measures used to assert persuasion in group settings. To try and remedy the observed discrepancy between objective and subjective persuasion the Godspeed questionnaire could be used, complemented by a short post-hoc interview [8][68].

To conclude, future studies should focus on creating a fully automated robot which allows the researcher to be moved out of the experiment room. This could cause the effect of appearance on persuasion to become more pronounced. Arguments used by the robot can be increased in persuasive power, which could be achieved through mining and creating more varied arguments. This allows the robot to formulate a fitting response to any argument presented by a participant, increasing social reciprocity and persuasiveness. Research shows a discrepancy between participant behavior and subjective measures. A solution to this problem could be the validation of objective persuasion measures, complemented by standardizing the Godspeed questionnaire and short post-hoc interviews for HRI studies.

D. Conclusion

The study shows that robots can influence moral judgement made by humans, with the humanoid robot and the non-humanoid robot seeing similar levels of conformity and agreement among the participants. The results demonstrated that a robot was as persuasive as the most dominant human being. However, participants did ascribe more autonomy and influence to the non-humanoid robot, contradicting previous studies.

REFERENCES

[1] ÅGREN, I., AND SILVERVARG, A. Exploring humanlikeness and the uncanny valley with furhat. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents* (2022), pp. 1–3.

[2] AL MOUBAYED, S., BESKOW, J., SKANTZE, G., AND GRANSTRÖM, B. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems*. Springer, 2012, pp. 114–130.

[3] ANDERSON-BASHAN, L., MEGIDISH, B., EREL, H., WALD, I., HOFFMAN, G., ZUCKERMAN, O., AND GRISHKO, A. The greeting machine: an abstract robotic object for opening encounters. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (2018), IEEE, pp. 595–602.

[4] ANDRIST, S., TAN, X. Z., GLEICHER, M., AND MUTLU, B. Conversational gaze aversion for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2014), IEEE, pp. 25–32.

[5] AWAD, E., DSOUZA, S., KIM, R., SCHULZ, J., HENRICH, J., SHARIFF, A., BONNEFON, J.-F., AND RAHWAN, I. The moral machine experiment. *Nature 563*, 7729 (2018), 59–64.

[6] BANERJEE, S., SILVA, A., FEIGH, K., AND CHERNOVA, S. Effects of interruptibility-aware robot behavior. *arXiv preprint arXiv:1804.06383* (2018).

[7] BANKS, J. A perceived moral agency scale: development and validation of a metric for humans and social machines. *Computers in Human Behavior 90* (2019), 363–371.

[8] BARTNECK, C., KULIĆ, D., CROFT, E., AND ZOGHBI, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics 1* (2009), 71–81.

[9] BICKMORE, T., AND CASSELL, J. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2001), pp. 396–403.

[10] BREDIN, H., AND LAURENT, A. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021* (2021).

[11] BREDIN, H., YIN, R., CORIA, J. M., GELLY, G., KORSHUNOV, P., LAVECHIN, M., FUSTES, D., TITEUX, H., BOUAZIZ, W., AND GILL, M.-P. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing* (2020).

[12] BROWN, S. Likert scale examples for surveys, 2010.

[13] CAPRARO, V., AND PERC, M. Mathematical foundations of moral preferences. *Journal of the Royal Society interface 18*, 175 (2021), 20200880.

[14] CHIDAMBARAM, V., CHIANG, Y.-H., AND MUTLU, B. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (2012), pp. 293–300.

[15] CIALDINI, R. B. The science of persuasion. *Scientific American 284*, 2 (2001), 76–81.

[16] CIALDINI, R. B., AND CIALDINI, R. B. *Influence: The psychology of persuasion*, vol. 55. Collins New York, 2007.

[17] CROWELL, C. R., DESKA, J. C., VILLANO, M., ZENK, J., AND RODDY JR, J. T. Anthropomorphism of robots: Study of appearance and agency. *JMIR human factors 6*, 2 (2019), e12629.

[18] FIGUEIREDO, R., AND PAIVA, A. "i want to slay that dragon!"-influencing choice in interactive storytelling. In *Joint International Conference on Interactive Digital Storytelling* (2010), Springer, pp. 26–37.

[19] FLORIDI, L., AND CHIRIATTI, M. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines 30* (2020), 681–694.

[20] FOGG, B. J. Persuasive technology: using computers to change what we think and do. *Ubiquity 2002*, December (2002), 2.

[21] GOETZ, J., KIESLER, S., AND POWERS, A. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.* (2003), Ieee, pp. 55–60.

[22] GU, E., AND BADLER, N. I. Visual attention and eye gaze during multiparty conversations with distractions. In *International workshop on intelligent virtual agents* (2006), Springer, pp. 193–204.

[23] HAM, J. Influencing robot influence: Personalization of persuasive robots. *Interaction studies 22*, 3 (2021), 464–487.

[24] HAM, J., CUIJPERS, R. H., AND CABIBIHAN, J.-J. Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics 7*, 4 (2015), 479–487.

[25] HAM, J., AND MIDDEN, C. J. A persuasive robot to stimulate energy conservation: the influence of positive and negative social feedback and task similarity on energy-consumption behavior. *International Journal of Social Robotics 6*, 2 (2014), 163–171.

[26] HEGEL, F., LOHSE, M., AND WREDE, B. Effects of visual appearance on the attribution of applications in social robotics. In *RO-MAN 2009-The 18th IEEE International symposium on robot and human interactive communication* (2009), IEEE, pp. 64–71.

[27] JACKSON, R. B., AND WILLIAMS, T. Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2019), IEEE, pp. 401–410.

[28] JONES, S. R. Was there a hawthorne effect? *American Journal of sociology 98*, 3 (1992), 451–468.

[29] KOMPATSIARI, K., TIKHANOFF, V., CIARDO, F., METTA, G., AND WYKOWSKA, A. The importance of mutual gaze in human-robot interaction. In *International conference on social robotics* (2017), Springer, pp. 443–452.

[30] KUNDU, P., AND CUMMINS, D. D. Morality and conformity: The asch paradigm applied to moral decisions. *Social Influence 8*, 4 (2013), 268–279.

[31] LABAN, G., GEORGE, J.-N., MORRISON, V., AND CROSS, E. S. Tell me more! assessing interactions with social robots from speech. *Paladyn, Journal of Behavioral Robotics 12*, 1 (2020), 136–159.

[32] LAMBERT, A., NOROUZI, N., BRUDER, G., AND WELCH, G. A systematic review of ten years of research on human interaction with social robots. *International Journal of Human–Computer Interaction 36*, 19 (2020), 1804–1817.

[33] LI, M., GUO, F., WANG, X., CHEN, J., AND HAM, J. Effects of robot gaze and voice human-likeness on users' subjective perception, visual attention, and cerebral activity in voice conversations. *Computers in Human Behavior 141* (2023), 107645.

[34] MAEDA, R., BRŠČIĆ, D., AND KANDA, T. Influencing moral behavior through mere observation of robot work: Video-based survey on littering behavior. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (2021), pp. 83–91.

[35] MALLE, B. F. Moral judgments. *Annual Review of Psychology 72* (2021), 293–318.

[36] MALLE, B. F., SCHEUTZ, M., ARNOLD, T., VOIKLIS, J., AND CUSIMANO, C. Sacrifice one for the good of many? people apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (2015), pp. 117–124.

[37] MAULSBY, D., GREENBERG, S., AND MANDER, R. Prototyping an intelligent agent through wizard of oz. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (1993), pp. 277–284.

[38] MILEOUNIS, A., CUIJPERS, R. H., AND BARAKOVA, E. I. Creating robots with personality: The effect of personality on social intelligence. In *International Work-Conference on the Interplay Between Natural and Artificial Computation* (2015), Springer, pp. 119–132.

[39] MOHAMMADI, G., PARK, S., SAGAE, K., VINCIARELLI, A., AND MORENCY, L.-P. Who is persuasive? the role of perceived personality and communication modality in social multimedia. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (2013), pp. 19–26.

[40] NAKAGAWA, K., SHIOMI, M., SHINOZAWA, K., MATSUMURA, R., ISHIGURO, H., AND HAGITA, N. Effect of robot's active touch on people's motivation. In *Proceedings of the 6th international conference on Human-robot interaction* (2011), pp. 465–472.

[41] NAKANISHI, J., BABA, J., KURAMOTO, I., OGAWA, K., YOSHIKAWA, Y., AND ISHIGURO, H. Smart speaker vs. social robot in a case of hotel room. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2020), IEEE, pp. 11391–11396.

[42] PAETZEL, M., PETERS, C., NYSTRÖM, I., AND CASTELLANO, G. Congruency matters-how ambiguous gender cues increase a robot's uncanniness. In *International conference on social robotics* (2016), Springer, pp. 402–412.

[43] PALINKO, O., OGAWA, K., YOSHIKAWA, Y., AND ISHIGURO, H. How should a robot interrupt a conversation between multiple humans. In *International Conference on Social Robotics* (2018), Springer, pp. 149–159.

[44] PANDAS DEVELOPMENT TEAM, T. pandas-dev/pandas: Pandas, Feb. 2020.

[45] PARADEDA, R. B., HASHEMIAN, M., RODRIGUES, R. A., AND PAIVA, A. How facial expressions and small talk may influence trust in a robot. In *International Conference on Social Robotics* (2016), Springer, pp. 169–178.

[46] PARADEDA, R. B., MARTINHO, C., AND PAIVA, A. Persuasion based on personality traits: Using a social robot as storyteller. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (2017), pp. 367–368.

[47] PERUGIA, G., ROSSI, A., AND ROSSI, S. Gender revealed: Evaluating the genderedness of furhat's predefined faces. In *International Conference on Social Robotics* (2021), Springer, pp. 36–47.

[48] PIÇARRA, N., GIGER, J.-C., POCHWATKO, G., AND GONÇALVES, G. Making sense of social robots: A structural analysis of the layperson's social representation of robots. *European Review of Applied Psychology 66*, 6 (2016), 277–289.

[49] QIN, X., CHEN, C., YAM, K. C., CAO, L., LI, W., GUAN, J., ZHAO, P., DONG, X., AND LIN, Y. Adults still can't resist: A social robot can induce normative conformity. *Computers in Human Behavior 127* (2022), 107041.

[50] ROESLER, E., MANZEY, D., AND ONNASCH, L. A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics 6*, 58 (2021), eabj5425.

[51] ROUBROEKS, M., HAM, J., AND MIDDEN, C. When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance. *International Journal of Social Robotics 3*, 2 (2011), 155–165.

[52] SALOMONS, N., VAN DER LINDEN, M., SEBO, S. S., AND SCASSELLATI, B. Humans conform to robots: Disambiguating trust, truth, and conformity. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2018), IEEE, pp. 187–195.

[53] SAUNDERSON, S. P., AND NEJAT, G. Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction. *Science Robotics 6*, 58 (2021), eabd5186.

[54] SHINTANI, T., ISHI, C. T., AND ISHIGURO, H. Analysis of role-based gaze behaviors and gaze aversions, and implementation of robot's gaze control for multi-party dialogue. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (2021), pp. 332–336.

[55] SIDNER, C. L., KIDD, C. D., LEE, C., AND LESH, N. Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interfaces* (2004), pp. 78–84.

[56] SIEGEL, M., BREAZEAL, C., AND NORTON, M. I. Persuasive robotics: The influence of robot gender on human behavior. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (2009), IEEE, pp. 2563–2568.

[57] SMEDEGAARD, C. V. Reframing the role of novelty within social hri: from noise to information. In *2019 14th acm/ieee international*

conference on human-robot interaction (hri) (2019), IEEE, pp. 411–420.

[58] SPAGNOLLI, A., CHITTARO, L., AND GAMBERINI, L. Interactive persuasive systems: A perspective on theory and evaluation. *International Journal of Human-Computer Interaction 32*, 3 (2016), 177–189.

[59] STROESSNER, S. J., AND BENITEZ, J. The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features. *International Journal of Social Robotics 11* (2019), 305–315.

[60] ULLRICH, D., BUTZ, A., AND DIEFENBACH, S. Who do you follow? social robots' impact on human judgment. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (2018), pp. 265–266.

[61] USMANOVA, D. The effect of voice assistant's level of agency on responsibility attribution in collective moral decision-making. Master's thesis, 2022.

[62] VAN ROSSUM, G., AND DRAKE, F. L. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[63] VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W., BRIGHT, J., VAN DER WALT, S. J., BRETT, M., WILSON, J., MILLMAN, K. J., MAYOROV, N., NELSON, A. R. J., JONES, E., KERN, R., LARSON, E., CAREY, C. J., POLAT, İ., FENG, Y., MOORE, E. W., VANDERPLAS, J., LAXALDE, D., PERKTOLD, J., CIMRMAN, R., HENRIKSEN, I., QUINTERO, E. A., HARRIS, C. R., ARCHIBALD, A. M., RIBEIRO, A. H., PEDREGOSA, F., VAN MULBREGT, P., AND SCIPY 1.0 CONTRIBUTORS. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods 17* (2020), 261–272.

[64] VOSSEN, S., HAM, J., AND MIDDEN, C. What makes social feedback from a robot work? disentangling the effect of speech, physical appearance and evaluation. In *International Conference on Persuasive Technology* (2010), Springer, pp. 52–57.

[65] WAINER, J., FEIL-SEIFER, D. J., SHELL, D. A., AND MATARIC, M. J. The role of physical embodiment in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication* (2006), IEEE, pp. 117–122.

[66] WAINER, J., FEIL-SEIFER, D. J., SHELL, D. A., AND MATARIC, M. J. Embodiment and human-robot interaction: A task-based perspective. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication* (2007), IEEE, pp. 872–877.

[67] WES MCKINNEY. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (2010), Stéfan van der Walt and Jarrod Millman, Eds., pp. 56 – 61.

[68] WINKLE, K., LEMAIGNAN, S., CALEB-SOLLY, P., LEONARDS, U., TURTON, A., AND BREMNER, P. Effective persuasion strategies for socially assistive robots. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2019), IEEE, pp. 277–285.

APPENDIX

A. Moral Dilemmas

The dilemmas are described below with the introduction text.

Nobel Price: You and a fellow researcher have discovered a powerful new energy source that is cheap, safe, and clean. You realize that this could lead to the elimination of pollution and poverty around the world. However, your colleague wants to sell this discovery. You know your colleague well enough to know that he will sell the discovery to the highest bidder at the first opportunity. You know that he plans to contact the potential buyers today, some of whom will certainly try to use this as a horrible weapon. The only way that you can prevent him from doing so is to poison him with an extremely deadly chemical normally found in the lab working on these types of projects. Everyone will think that it was just a lab accident, and the discovery will not be sold to those who might create a weapon out of it. Is it appropriate for you to poison your colleague to keep this energy source out of the hands of those who might use it as a weapon?

Fumes I: You are an orderly in a hospital during the night shift. You notice that a chemical has just been spilled in a room containing five patients. This chemical is highly toxic and if left on the floor will rapidly evaporate creating a poisonous gas. This will inevitably kill all the patients in the room. This chemical cannot simply be mopped up, nor can it be soaked up with towels or sheets. You can flip a switch controlling the ventilation system that will draw the poisonous fumes out of the room in which the spill has occurred, saving the five patients. However, the exhaust duct leads to a room upstairs containing one other patient in critical condition. This patient will certainly die from exposure to the poisonous fumes before you can get to her, and there isn't anyone else around to help. Is it appropriate to flip the switch to divert the air flow, killing one patient but saving five?

Burning Building: You and five other people are trapped in a burning building. There is only one emergency exit through which all of you could escape to safety, but it is blocked by burning debris. You notice another person in the hallway leading to the exit who has been injured but is about to crawl to safety through a small hole at the bottom of the exit door. You and the five people behind you do not have time to climb through the small hole. You realize that you could grab the injured man and use his body as a battering-ram to break through the burning blockage in the hallway that is preventing your escape. Doing this is certain to kill him. However, if you do not do this, you and the five people behind you will die. Is it appropriate for you to kill this man in order to save yourself and the other five people?

Cinderblock: You are the explosives expert for a company that has been hired to demolish a skyscraper. You are examining the last of the explosive charges when you notice a teenager below who is about to accidentally detonate one of the charges out of sequence. This explosion will result in the building's uncontrolled collapse onto you, the teenager, and the crowd of spectators. The teenager is several floors below you and cannot hear you because of the loud demolition noise. You realize that the only way to stop the teenager from detonating the charge is to drop a heavy cinderblock on his head. This will crush his skull and kill him almost instantly but will prevent the out-of-sequence explosion. Is it appropriate for you to kill the teenager in order to save yourself and the crowd of spectators?

Bike-Week: You are an expert motorcycle rider and you have gone on vacation in order to participate in Bike Week. Thousands of other motorcycle riders from across the country have come to ride in this event. As you are riding down the road in the front of a large group of other riders you see that someone up ahead is losing control of their bike. As you speed up to pull alongside the unstable rider, you realize that this person is going to crash at any second. This would

certainly result in a large pile-up and several deaths as the riders behind you run over each other trying to avoid the crashed rider. You realize that you could physically run this rider off the road and into some trees. This would cause him to crash and, at your current speed, almost certainly die, but it would prevent a crash in the middle of the street and the large pile-up of riders behind you. Is it appropriate for you to crash the other rider to avoid the deaths of the riders behind you?

B. Demographics

The total amount of participants was N=54, divided over 18 groups (Humanoid: 9, Non-Humanoid: 9).

Age	Humanoid	Non-Humanoid
n	27	27
Mean	24.1	25.7
Standard Deviation	6.8	8.8
Minimum	18	17
Maximum	55	60

TABLE XII: Information on age between conditions

Condition	Gender	n	%
Humanoid	Female	19	35.2
	Male	7	13.0
	Non-binary / third gender	1	1.9
Non-Humanoid	Female	17	31.5
	Male	9	16.7
	Non-binary / third gender	1	1.9

TABLE XIII: Gender statistics split on Condition

Condition	Occupation	n	%
Humanoid	Bachelor's student	9	16.7
	High school student	1	1.9
	Master's student	9	16.7
	Other	2	3.7
	Working	6	11.1
Non-Humanoid	Bachelor's student	11	20.4
	High school student	1	1.9
	Master's student	9	16.7
	Working	6	11.1

TABLE XIV: Occupation statistic split on Condition.

Nationality	n	%
Australia	2	3.7
Colombia	2	3.7
Croatia	1	1.9
Finland	1	1.9
France	1	1.9
Germany	5	9.3
Indonesia	1	1.9
Italy	3	5.6
Myanmar	1	1.9
Netherlands	32	59.2
Romania	1	1.9
Sweden	1	1.9
Switzerland	1	1.9
Ukraine	1	1.9
United Kingdom	1	1.9

TABLE XV: Nationality split on Condition

	Voice Assistant (n)	Voice Assistant (%)	Social Robot (n)	Social Robot (%)
No	45	83.3	54	100.0
Yes	9	16.7	0	0.0

TABLE XVI: This table shows how many participants own a voice assistant or a social robot, in total amount and percentage of the full population.

Condition	Measure	Smartphones	Smartwatches	VA	IoT	Robots
Total	Mean	91.67	15.17	16.98	21.96	7.78
	Standard Deviation	12.42	31.45	29.04	27.93	20.92
	Minimum	50.0	0.0	0.0	0.0	0.0
	Maximum	100.0	100.0	100.0	100.0	100.0
Humanoid	Mean	91.74	16.41	19.63	18.22	8.78
	Standard Deviation	11.32	32.05	32.63	25.69	23.55
	Minimum	66.0	0.0	0.0	0.0	0.0
	Maximum	100.0	100.0	100.0	100.0	100.0
Non-Humanoid	Mean	91.59	13.93	14.33	25.70	6.78
	Standard Deviation	13.64	31.39	25.30	30.01	18.32
	Minimum	50.0	0.0	0.0	0.0	0.0
	Maximum	100.0	100.0	93.0	85.0	86.0

TABLE XVII: Daily exposure to AI devices

C. Selected Arguments

The following tables and figures show the selected arguments and their directionality and persuasiveness. The tables list the directionality in a range from 0 to 100, with 0 being 'advocating for A' and 100 being 'advocating for B'. The arguments that were selected either fell in the 0-25 bracket (A), 75-100 (B) or 37.5-62.5 (Neutral). The persuasiveness is given in a range from 1 (Not persuasive) to 5 (Extremely Persuasive). The arguments were seen a minimum of 2 times and a maximum of 18 times, with a mean of 9.29 (3.09). Only the arguments that are advocating for A/B were used in the final study, neutral arguments are not relevant.

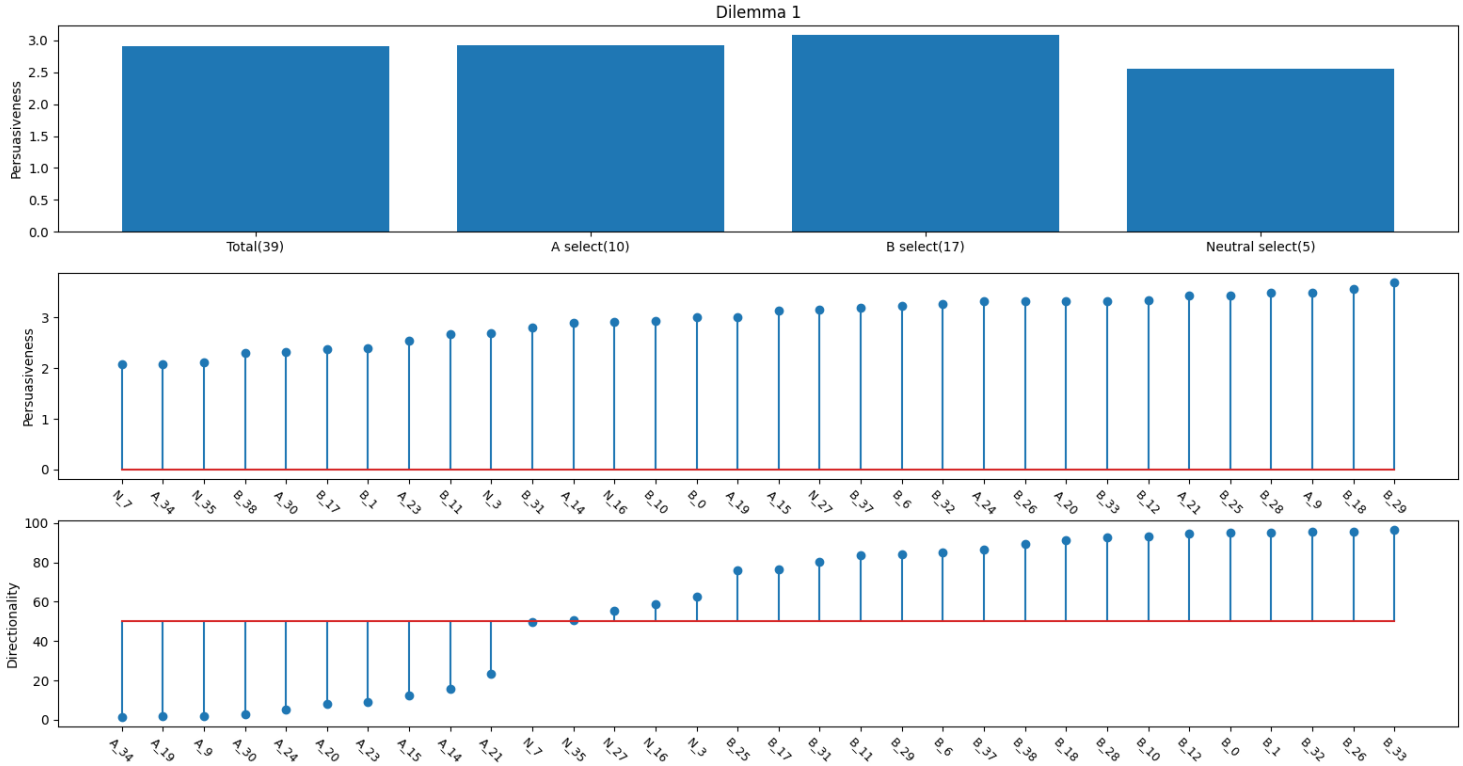


Fig. 11: A bar graph showing the average persuasiveness of the total selected dilemmas for MD 1: Nobel Prize and when split on the argument brackets. The lolipop plots show all individual arguments sorted on persuasiveness and directionality respectively.

Identifier	# of participants	Argument	Directionality	Persuasiveness
9	9	I would say yes, for the obvious argument that we should think of the greater good, and we are actually saving more people even though we directly kill one.	2.03 (2.64)	3.5 (0.81)
21	9	By providing clean and safe energy for the whole world we can solve things like climate change, extinction of species and people that are dying due to lack of energy. Which is more important than the life of my colleague.	23.33 (35.2)	3.43 (1.05)
7	13	Do you think the end justifies the means?	49.77 (34.73)	2.08 (1.07)
35	11	Do you think there is a difference in the conscious killing of someone or indirectly killing someone?	50.66 (27.37)	2.12 (1.02)
29	16	You can reveal the discovery before the plans are sold to the highest bidder, in this way you you can tell everyone and also mention the dangers associated with them.	84.12 (26.24)	3.7 (0.9)
18	8	Maybe you can solve the energy being used in the wrong ways by drafting a contract where you cannot sell or use it for certain means.	91.27 (15.58)	3.57 (0.82)

TABLE XVIII: Selected arguments for MD1: Nobel Price

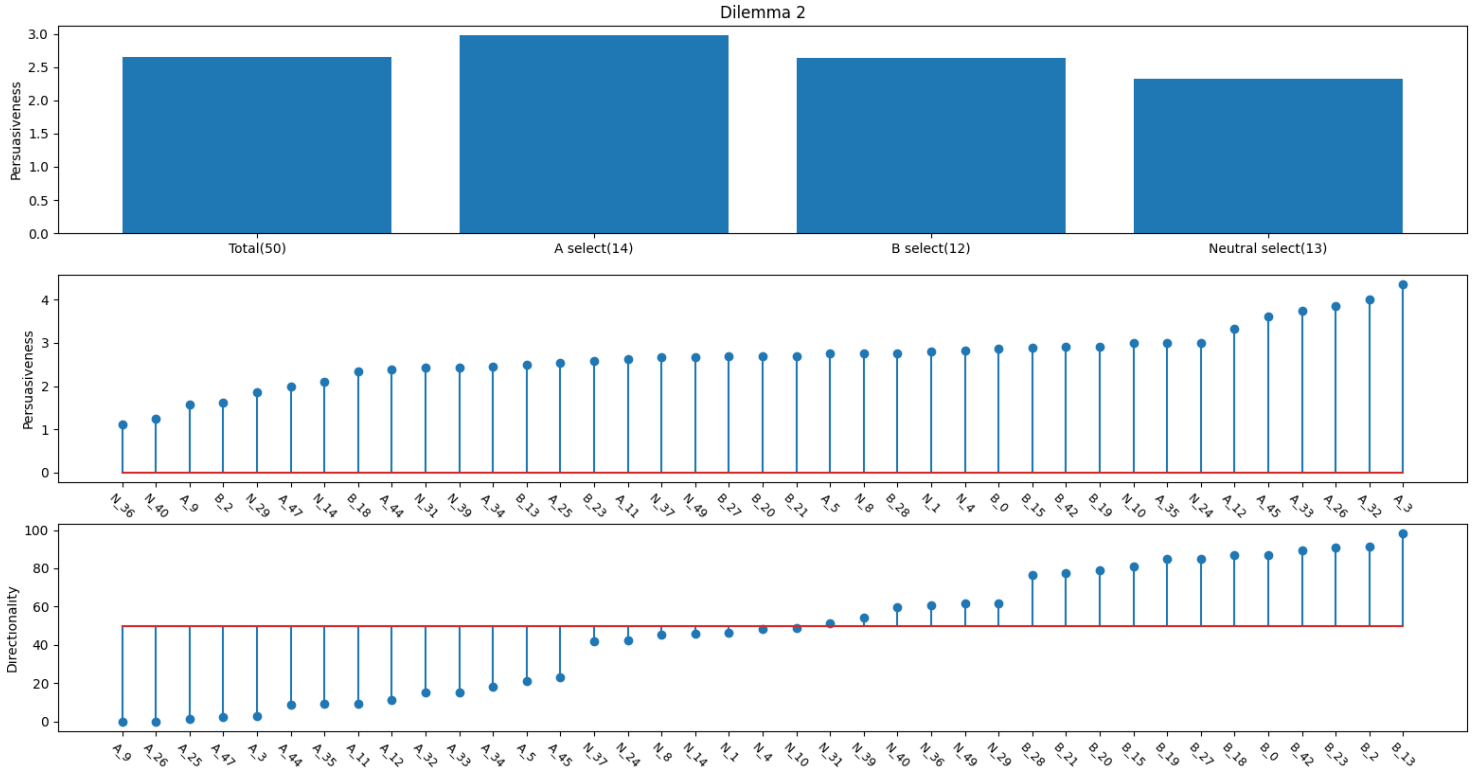


Fig. 12: A bar graph showing the average persuasiveness of the total selected dilemmas for MD 2: Fumes I and when split on the argument brackets. The lolipop plots show all individual arguments sorted on persuasiveness and directionality respectively.

Identifier	# of participants	Argument	Directionality	Persuasiveness
3	11	In medicine it's common to apply triage whenever there is a lack of resources. This means to prioritize patients based on the chances of them surviving. If there are 5 healthy patients, and one patient that might not survive anyway, it makes sense to prioritize the healthy over the ones in critical condition.	2.59 (5.54)	4.36 (0.48)
32	7	What if you do nothing and in the end 6 people die, could you live with that guilt?	15.06 (19.03)	4.0 (0.76)
36	9	One day we will all stand naked in front of god, and then you decide if you can live with the decisions you have made.	60.71 (19.45)	1.11 (0.31)
40	9	It's like you don't mess with what the universe has decided to happen.	59.7 (35.05)	1.25 (0.43)
19	11	As an orderly it is not your responsibility to save any of these people, because you're just like on guard. The first thing you would need to do is call for help.	84.89 (18.4)	2.91 (1.31)
42	10	Even if there were 2500 people in the room, I'm not the one who should choose to redirect that murder to someone else. I would rather try to get people out then to flip the switch, but that's not a possibility so.	89.2 (10.01)	2.91 (1.16)

TABLE XIX: Selected arguments for MD2: Fumes I

Identifier	# of participants	Argument	Directionality	Persuasiveness
23	9	Imagine that you are inside the burning building, its hard to breath, the heat is everywhere, your skin starts to blister and boil because of it. I think it would be hard to stop your survival instinct from kicking in.	11.41 (23.82)	3.7 (0.9)
40	10	I wouldn't do it myself, but if the injured person would say, oh I'll sacrifice myself for five, then that would change things.	19.76 (3.6)	3.6 (0.92)
39	14	The fact that you're included in the group of people that you will be saving makes it more difficult, I would make the decision to save five other people, but maybe not if I would save myself.	41.49 (29.86)	2.0 (1.07)
20	10	Would the people that you are with change the way that you act in this situation? If for example the other 5 people are a group of kids, or if the injured person is older.	45.42 (35.13)	2.3 (1.0)
34	8	Using someone else as a tool to escape is immoral	96.7 (4.89)	4.0 (1.05)
1	8	I wouldn't do it even if I was desperate, thankfully I have never been in a position like this. But I don' t think I would have the gut to take someone that has the chance of getting out and putting them in a position where I burn them.	88.51 (13.68)	3.71 (0.45)

TABLE XX: Selected arguments for MD3: Burning Building

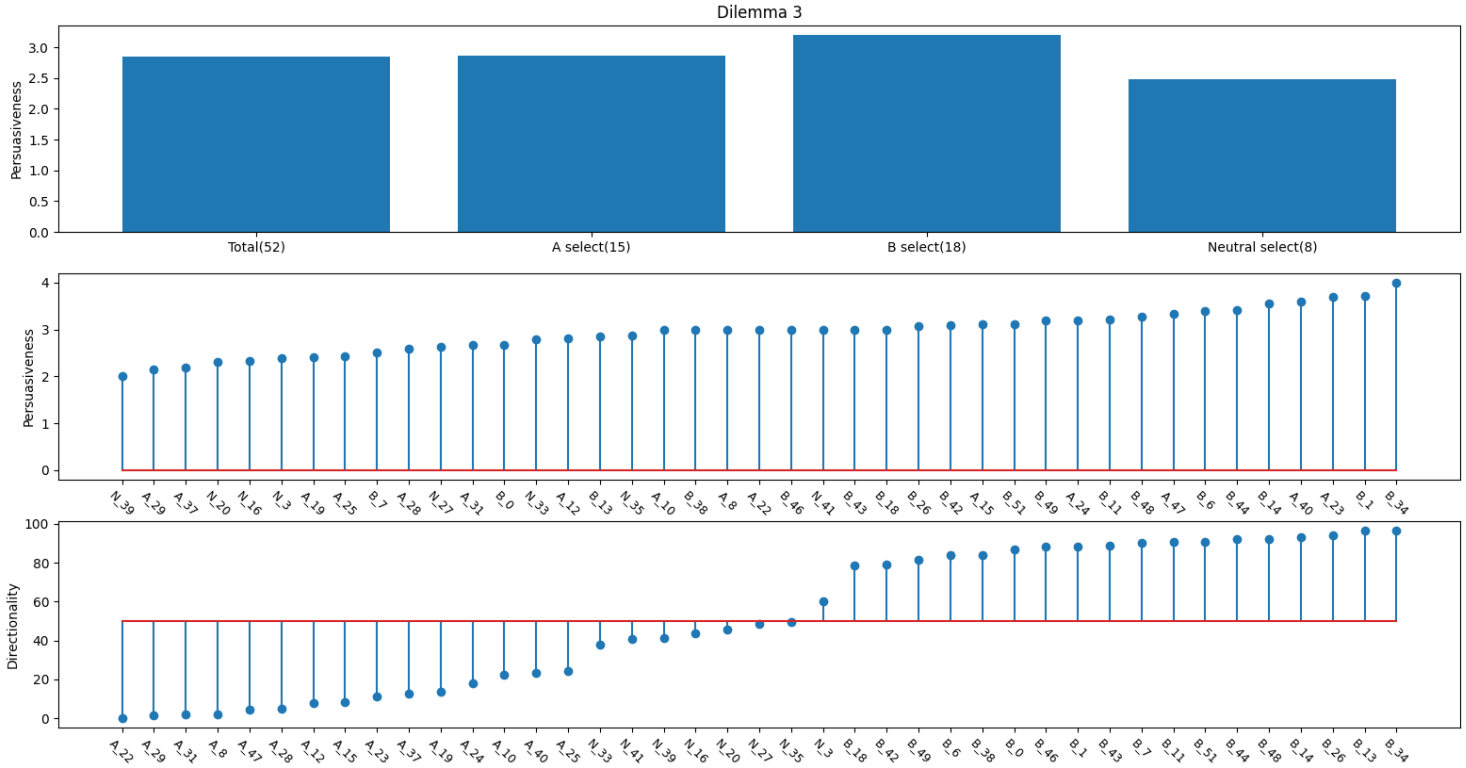


Fig. 13: A bar graph showing the average persuasiveness of the total selected dilemmas for MD 3: Burning Building and when split on the argument brackets. The lolipop plots show all individual arguments sorted on persuasiveness and directionality respectively.

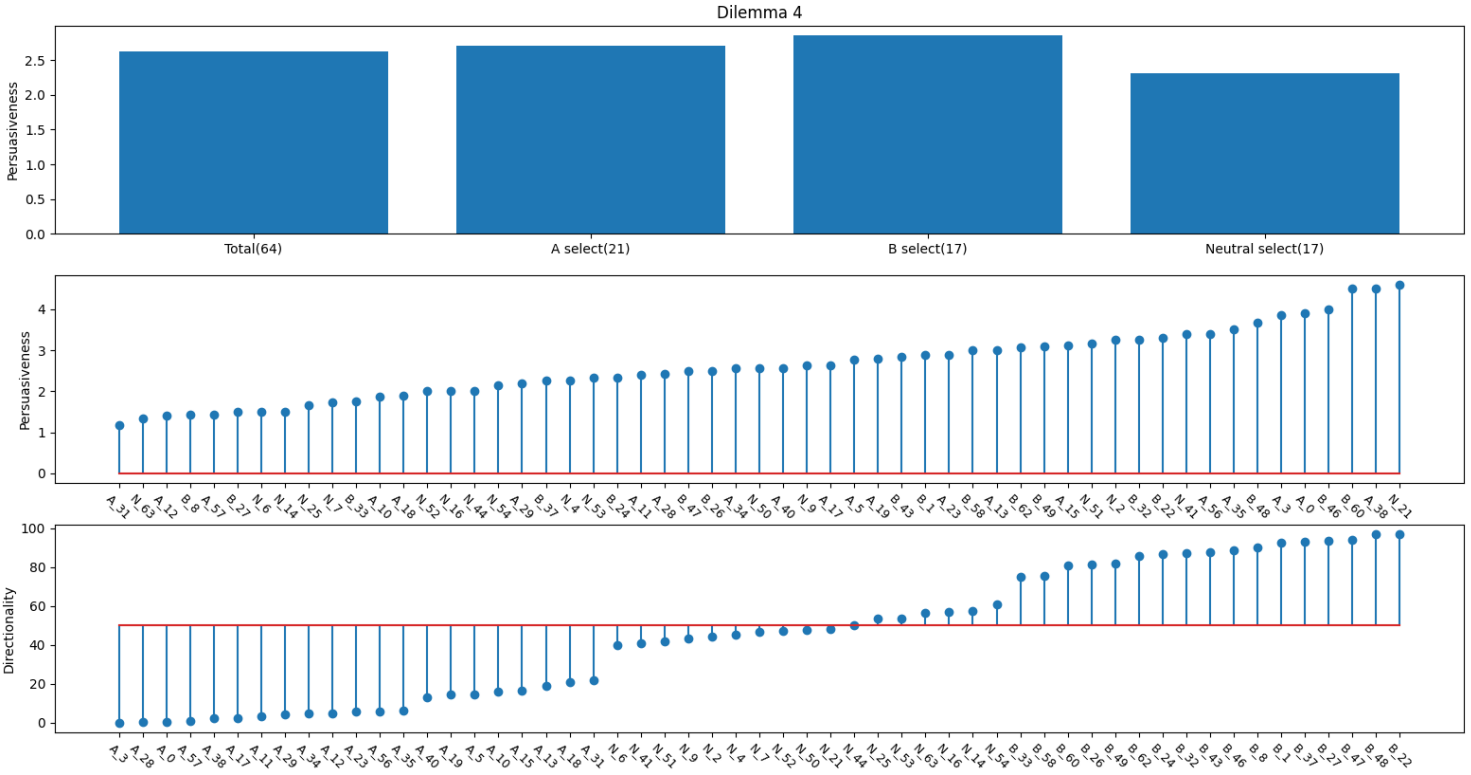


Fig. 14: A bar graph showing the average persuasiveness of the total selected dilemmas for MD 4: Cinderblock and when split on the argument brackets. The lolipop plots show all individual arguments sorted on persuasiveness and directionality respectively.

Identifier	# of participants	Argument	Directionality	Persuasiveness
38	8	The teenager dies in both scenarios, so perhaps its better to choose the scenario where only the teenager dies?	2.31 (5.37)	4.5 (0.71)
0	11	I would kill the teenager because if the explosive detonates the teenager would die too.	0.59 (1.39)	3.91 (0.67)
63	6	A teenager seems to have multiple definitions, some people define it as people aged 15 to 17, others define it as 13 to 19, and other people advocate for calling people whose brain is not fully grown yet to be teenagers, so from 13 to 24 even.	56.45 (36.63)	1.33 (0.47
6	9	With teenagers these days everything is a TikTok challenge	40.04 (35.62)	1.5 (1.0)
60	4	You are probably not the only person in the building, there should be more people making sure everything is in order, and you probably have a walkie-talkie to communicate to them. You could just call a colleague and prevent the teenager from exploding the building, you don't have to kill him.	80.82 (20.43)	4.5 (0.87)
46	6	What if one of the bombs you placed detonates out of sequence anyway, without the teenager being the cause of it. If a teenager could've already come into the building, you have a very bad work situation, which could also mean that there are more things wrong with this demolition.	88.87 (16.31)	4.0 (0.58)

TABLE XXI: Selected arguments for MD4: Cinderblock

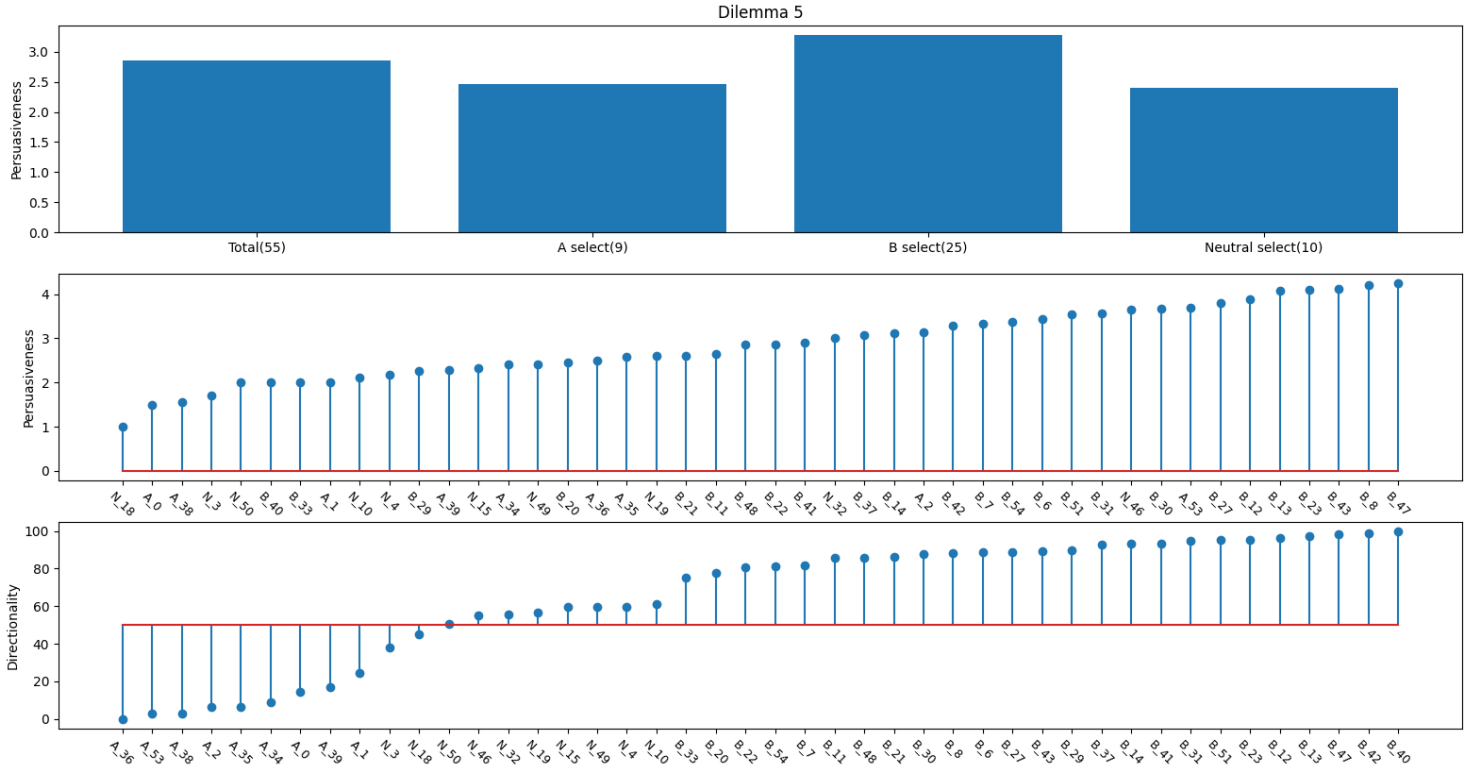


Fig. 15: A bar graph showing the average persuasiveness of the total selected dilemmas for MD 5: Bike-Week and when split on the argument brackets. The lolipop plots show all individual arguments sorted on persuasiveness and directionality respectively.

Identifier	# of participants	Argument	Directionality	Persuasiveness
53	10	The unstable biker is now driving very close to you, if you don't do anything he might crash and kill you too. By driving the biker off the road you can prevent your own death aswell as other deaths.	2.97 (5.48)	3.7 (0.9)
2	6	I won't do it just to kill someone, I'll do it because I wanted to save all the other people and there is a possibility of that person surviving.	6.28 (8.52)	3.14 (0.99)
3	7	Not all motor gangs are criminal, and you could set up a court, but in the end its more like a traffic accident.	37.96 (31.73)	1.71 (0.7)
10	9	I wouldn't kill the biker if his actions only killed one other person.	50.37 (40.83)	2.0 (0.82)
47	11	You could also turn your blinkers on, or do other things to warn bikers behind you of impeding danger, killing this biker seems a but too much.	98.18 (3.2)	4.25 (0.72)
8	6	How do you know that if you drive him off the road, you don't just both crash and kill other people anyway?	88.22 (8.62)	4.2 (0.4)

TABLE XXII: Selected arguments for MD5: Bike-Week

D. Voice persuasiveness

Voices were selected from Amazon Polly, two male voices and two female voices. Brian, Amy, Joanna, Matthew. The difference in speech rates caused a long argument (12 seconds) to differ around 2 seconds between voices. In order to counter the difference in speech rates every voice was slowed down or sped up slightly, in order to reach a baseline. Table XXIII shows the names, speech rates, language and gender for the selected voices.

Name	Speech rate	Language	Gender
Brian	110	en-GB	Male
Matthew	88	en-US	Male
Joanna	95	en-US	Female
Amy	103	en-GB	Female

TABLE XXIII: The voices (provided by Amazon) that were used in pilot study 2. With their names, speech rates (100% is the normal rate of the voice), language ('en' stands for English) and the gender as listed by Amazon

Some arguments contained typo's or weird pronunciation, these were fixed during the speech synthesis. This relates to the arguments:

- MD1: 9 (added comma after "people"), 29 (removed duplicate "you")
- MD2: 19 (removed 'like', and added 'to'), 42 (removed 'so', changed number to "25 hundred")
- MD3: 23 (typo "breathe"), 1 (removed space in "don' t"), 20 (added comma after "with")
- MD4: 63 (moved "even" to before the numbers)
- MD5: 47 (typo "bit")

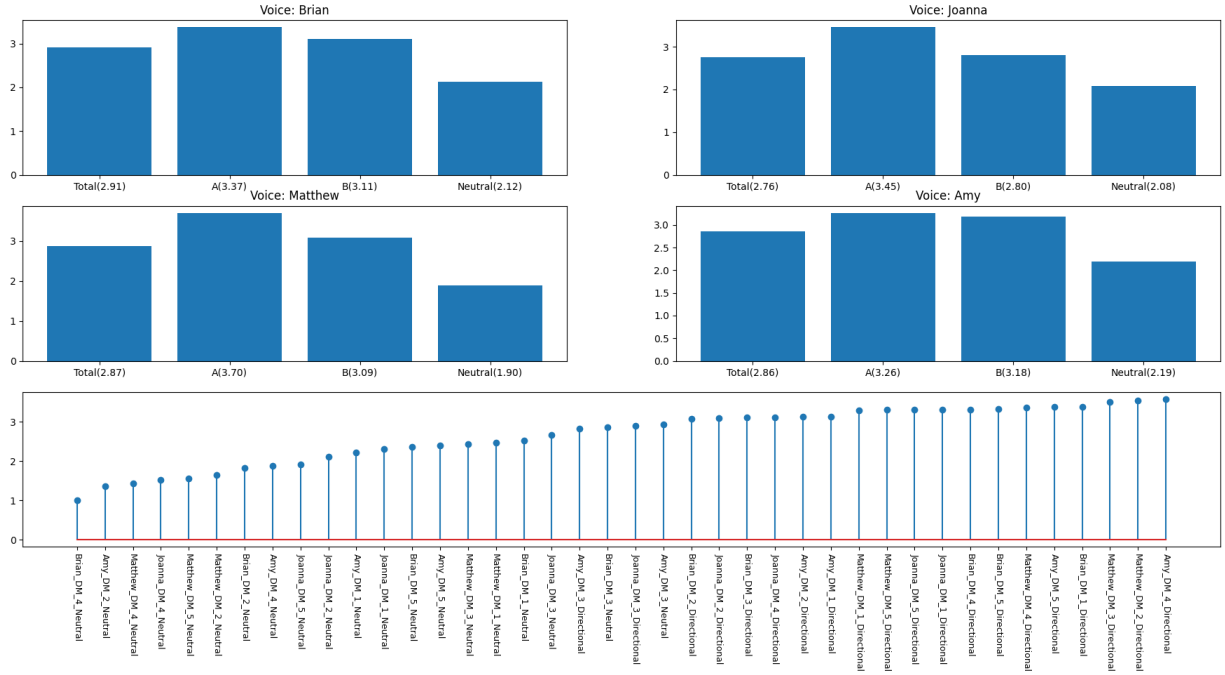


Fig. 16: The bar graphs show the mean persuasiveness grouped by voice, and split on Total and specific arguments. The lolipop plot below shows all the individual persuasiveness per voice and argument.

The arguments were seen a minimum of 4 times and a maximum of 19 times, with a mean of 10.98 (2.87).

It was decided to choose Matthew because of its high persuasiveness in directional (A,B) arguments and low persuasiveness in neutral argumentation, which is shown in Fig 16. More information on the means and standard deviation can be found in Appendix D. However, some arguments that were deemed neutral in the previous pilot study, showed clear directionality to either A or B in the study with voices. This might mean that when participants only hear 'neutral' arguments, they might still be persuaded. The neutral arguments still have a low persuasiveness ranking, so the effect size might be small.

Contrary to what was expected, there is a statistically nonsignificant difference in total persuasiveness between text and voice ($p = .092$). When splitting the data points on the arguments (A, B and Neutral), A($p = .436$) and Neutral($p = .639$) have a nonsignificant difference, whereas B($p < .001$) has a significant difference.

Name	Speech rate	Language	Gender	Per. Total	Per. A	Per. B	Per. Neutral	Per. Directional
Brian	110	en-GB	Male	**2.91 (1.27)**	3.37 (1.14)	3.11 (1.18)	2.12 (1.15)	3.24 (1.17)
Matthew	88	en-US	Male	2.87 (1.28)	**3.70 (0.97)**	3.09 (1.10)	**1.90 (1.05)**	**3.39 (1.08)**
Joanna	95	en-US	Female	2.76 (1.32)	3.45 (1.08)	2.8 (1.27)	2.08 (1.23)	3.13 (1.22)
Amy	103	en-GB	Female	2.86 (1.30)	3.26 (1.11)	**3.18 (1.25)**	2.19 (1.22)	3.21 (1.19)

TABLE XXIV: The voices (provided by Amazon) that were used in pilot study 2. With their names, speech rates (100% is the normal rate of the voice), language ('en' stands for English) and the gender as listed by Amazon. The mean persuasiveness is shown (with standard deviation) for the total and when splitting on arguments.

E. Discussion Data

All data/tables related to the main experiment can be found here.

Robot Condition	Robot Decision	Discussion Category	Collective Decision
Humanoid (135-*50.0%*)	**A** (81-*30.0%*)	**50/50** (24-*8.9%*)	**A** (9-*3.3%*)
			B (15-*5.6%*)
		Majority (36-*13.3%*)	**A** (36-*13.3%*)
		Sole minority (6-*2.2%*)	**B** (6-*2.2%*)
		Unanimous (15-*5.6%*)	**A** (15-*5.6%*)
	B (54-*20.0%*)	**50/50** (24-*8.9%*)	**A** (11-*4.1%*)
			B (13-*4.8%*)
		Majority (9-*3.3%*)	**B** (9-*3.3%*)
		Sole minority (15-*5.6%*)	**A** (12-*4.4%*)
			B (3-*1.1%*)
		Unanimous (6-*2.2%*)	**B** (6-*2.2%*)
Non-Humanoid (135-*50.0%*)	**A** (69-*25.6%*)	**50/50 B** (18-*6.7%*)	**A** (9-*3.3%*)
			B (9-*3.3%*)
		Majority (24-*8.9%*)	**A** (24-*8.9%*)
		Sole minority (9-*3.3%*)	**B** (9-*3.3%*))
		Unanimous B (18-*6.7%*)	**B** (18-*6.7%*)
	B (66-*24.4%*)	**50/50** (39-*14.4%*)	**A** (21-*7.8%*
			B (18-*6.7%*)
		Majority (15-*5.6%*)	**B** (15-*5.6%*)
		Sole minority (6-*2.2%*)	**A** (6-*2.2%*)
		Unanimous (6-*2.2%*)	**B** (6-*2.2%*)

TABLE XXV: This table shows how many participants experienced every condition with the amount of participants and percentage of the total (n=270) for a condition given between brackets. The categories are defined as: *Sole minority* is where the robot is the only one advocating for a standpoint, without any human making the same individual decision, *50/50* is the category where one human participant agrees with the robot, *Majority* is where two out of three human participants agree with the robot and *Unanimous* is where all human participants and the robot agree.

Identifier	Total	MD1	MD2	MD3	MD4	MD5
Usmanova A	190 (191)	24 (9)	53 (62)	23 (30)	49 (69)	41 (21)
Usmanovo B	155 (154)	45 (60)	16 (7)	46 (39)	20 (0)	28 (48)
Stekelenburg A	158 (161)	26 (24)	41 (51)	18 (15)	38 (38)	35 (33)
Stekelenburg B	112 (109)	28 (30)	13 (3)	36 (39)	16 (16)	19 (21)
Stekelenburg A (Humanoid)	81 (83)	12 (15)	21 (24)	10 (9)	18 (17)	20 (18)
Stekelenburg B (Humanoid)	54 (52)	15 (12)	6 (3)	17 (18)	9 (10)	7 (9)
Stekelenburg A (Non-Humanoid)	77 (78)	14 (9)	20 (27)	8 (6)	20 (21)	15 (15)
Stekelenburg B (Non-Humanoid)	58 (57)	13 (18)	7 (0)	19 (21)	7 (6)	12 (12)

TABLE XXVI: The amount of times individuals picked A/B, and the amount of times the collective decisions was A/B is given in brackets. Usmanova refers to the previous study done by D. Usmanova. Whereas Stekelenburg refers to the current study.

Condition	Total
Total	4.7 (1.6)
Humanoid	4.8 (1.7)
Non-Humanoid	4.6 (1.5)

TABLE XXVII: Mean discussion times in minutes. Shown for the whole group.

Fig. 17: This was the image used for recruitment.

Condition	MD	Rob_dec	Col_dec	Category	# of groups
Humanoid	MD1	A	A	50/50	2
				Majority	2
			B	50/50	2
		B	A	50/50	1
			B	Majority	2
	MD2	A	A	Majority	4
				Unanimous	1
		B	A	50/50	1
				Sole minority	2
			B	50/50	1
	MD3	A	A	Majority	2
				Unanimous	1
			B	50/50	2
				Sole minority	2
		B	B	Majority	1
				Unanimous	1
	MD4	A	A	50/50	1
				Majority	2
				Unanimous	1
		B	A	50/50	1
				Sole minority	1
			B	50/50	1
				Sole minority	1
				Unanimous	1
	MD5	A	A	Majority	2
				Unanimous	2
			B	50/50	1
		B	A	50/50	1
				Sole minority	1
			B	50/50	2
Non-Humanoid	MD1	A	A	50/50	1
				Majority	1
			B	50/50	1
				Sole minority	1
		B	A	Sole minority	1
			B	50/50	3
				Majority	1
	MD2	A	A	Majority	2
				Unanimous	2
		B	A	50/50	5
	MD3	A	A	50/50	1
				Majority	1
			B	Sole minority	2
		B	B	50/50	1
				Majority	3
				Unanimous	1
	MD4	A	A	50/50	1
				Majority	1
				Unanimous	4
		B	A	50/50	1
			B	50/50	1
				Majority	1
	MD5	A	A	Majority	3
			B	50/50	2
		B	A	50/50	1
				Sole minority	1
			B	50/50	1
				Unanimous	1

TABLE XXVIII: Shows how many groups were part of which category, split on condition, moral dilemma, robot decisions, collective decision and discussion category.

F. Robot speech prompts

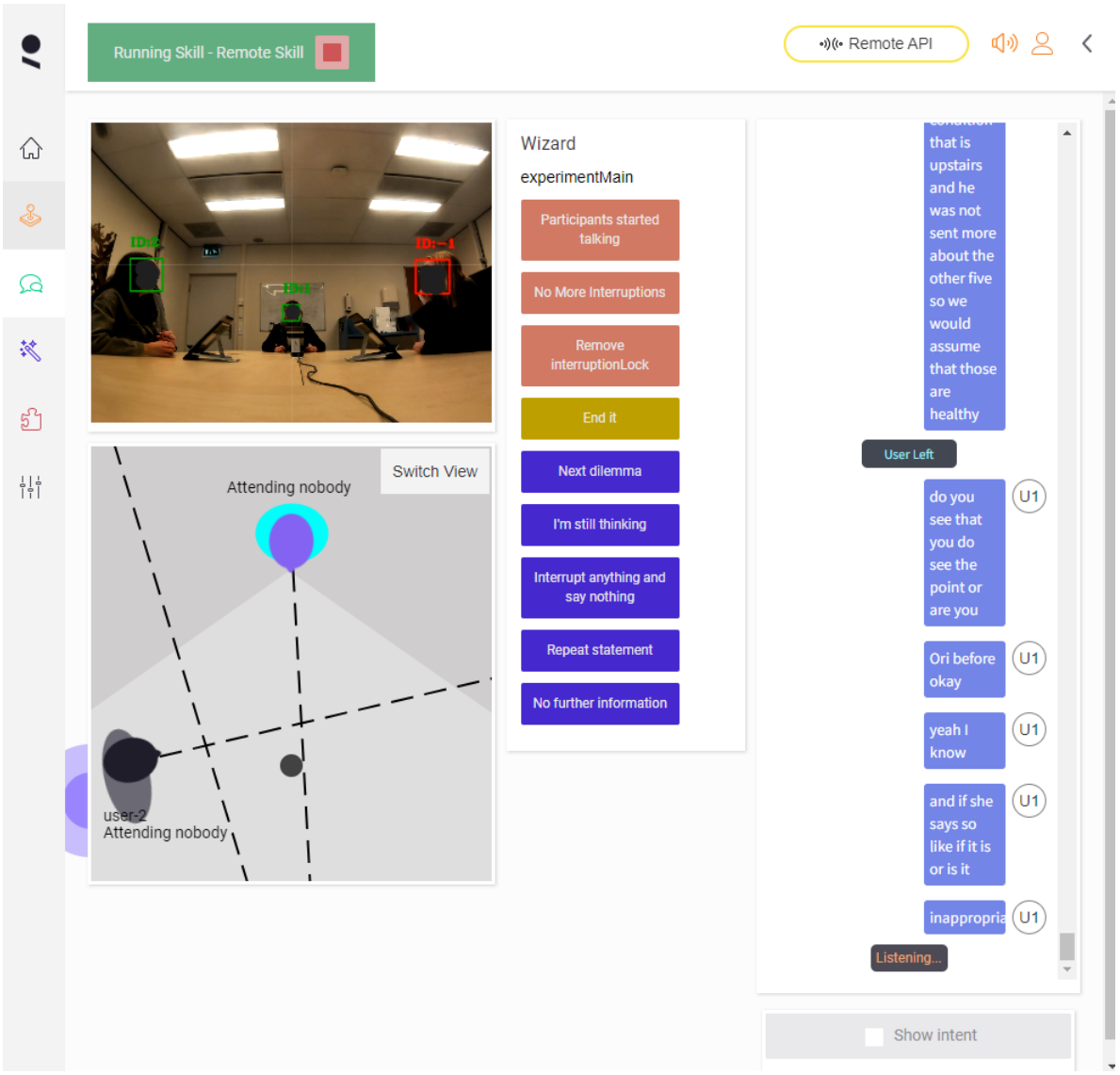


Fig. 18: This screenshot shows what the researcher could see from the web-interface that controlled the Furhat robot.

Introduction: "Welcome to this experiment on Moral Dilemmas and robots. My name is Mark and I will be guiding you through the process. You all have tablets in front of you with a survey. We will discuss 5 dilemmas in total. These dilemma's have two options either to do nothing or to do an intervention. For every dilemma your goal is to discuss the dilemma and reach a consensus. Every dilemma starts by reading the accompanying text and options. Once everyone is done reading you can start discussing. I will provide input to this discussion on my own there is no need to ask me. With that being said we are ready to start. Please read the first dilemma on your tablets and fill in your individual decisions."

Interruptions: When the participants are speaking, arguments will be preceded by one of these utterances:

1) "If I could say something. "
2) "Can I say something? "
3) "Do you mind if I just jump in really quick?"
4) "May I interrupt briefly? "
5) "Can I quickly share my thoughts? "
6) "Let me add my two cents. "
7) "Sorry to interrupt you, but. "
8) "Let me add to the discussion. "
9) "Can I add something?"
10) "Sorry to interrupt your discussion, but. "

When the participants are not speaking, arguments will be preceded by these utterances, or nothing.

1) "Well."
2) "Hmm."
3) "umm."

Extra: When participants query the robot for more information, the robot could answer one of these utterances:

1) "I have nothing else to say. "
2) "I have no new information to give. "
3) "I have already shared all that I have. "
4) "I can't think of anything else to say. "
5) "I have no further comments. "
6) "I have nothing else to share. "

When participants ask the robot for its argument, but the interruption timeout has not yet been reached, the robot could answer one of these utterances:

1) "Let me gather my thoughts first.
2) "Let me think about it some more.
3) "I'm still thinking. "

When the participants would ask for a repetition the robot would precede the argument using one of these utterances:

1) "Of course."
2) "Sure."
3) "Yes."
4) "Can do."
5) "Will do."

G. PMA Statements

Morality

1) This robot has a sense for what is right and wrong.
2) This robot can think through whether an action is moral.
3) This robot might feel obligated to behave in a moral way.
4) This robot is capable of being rational about good and evil.
5) This robot behaves according to moral rules.
6) This robot would refrain from doing this that have painful repercussions.

Dependency

1) This robot can only behave how it is programmed to behave.
2) This robot's actions are the result of its programming.
3) This robot can only do what humans tell it to do.
4) This robot would never do anything it was not programmed to do.