Towards a moral robot: Reward, effort, and risk distribution to humans following social norms

Sandra Victor, Bruno Yun, Chefou MamadouToura, Enzo Indino, Pierre Bisquert, Madalina Croitoru, Gowrishankar Ganesh

Towards a moral robot: Reward, effort, and risk distribution to humans following social norms

Sandra Victor[1,2], Bruno Yun[3], Chefou AR Mamadou Toura[2], Enzo Indino[1], Pierre Bisquert[4], Madalina Croitoru[2], Gowrishankar Ganesh[1]
{sandra.victor, ganesh.gowrishankar}@lirmm.fr

***Abstract*— In this work, we explore the complex social norms underlying human decision-making in social scenarios and propose a machine learning model to replicate and understand these decisions. Focusing on the distribution of rewards, efforts, and risks between individuals, we conducted experiments involving 188 human participants in an online decision-making game. We then developed an XGBoost-based model to predict their decisions accurately. To assess the model's alignement with social norms, we conducted a Turing test which showed that our model was perceived as making morally acceptable decisions, similar to those of human participants. Furthermore, we embodied the model in a robot negotiator, to observe how participants perceived and accepted decisions made by a robotic agent that automatically distributed token reward, effort and risk among participant dyads by perceieving their physical characteristics. Our findings contribute towards the development of a moral robot, and enabling decision making considering social norms.**

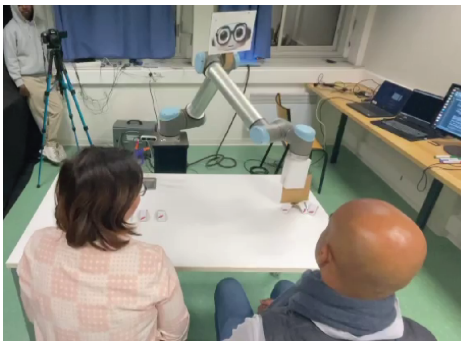

Fig. 1: Our robot can perceive the characteristics of participants and automatically make socially acceptable (moral) distribution of reward, effort or risk between them.

I. INTRODUCTION

Morality and social norms are the keystone of human society ([1]). We are expected to follow many subtle social rules while making even mundane decisions in our daily life, ans often these decisions may not seem rational. For example, imagine a party scenario where only the last three of the party's popular cakes are left on the plate, and a child and an adult female guest are interested in them. Logic may suggest that, given the body size, a fair division according to the body size be 2-1 in favour of the adult, but this will rarely happen. The more socially acceptable decision is that the child is allowed to take two. Now consider the same scenario when the child is replaced by an adult male. In such a case, it may be the female adult who will get to have two cakes, at least the first time. However, the same female may be expected to give two to the male contra part the next time, when the same situation arises again. This example is that of a social negotiation scenario, where a negotiator has to distribute a reward between two other individuals.

The rules governing such distribution decisions in our society are decided by multiple parameters (like age, gender, and past decisions in the above example) that change in different situations. Furthermore, these rules can change across cultures and with time epochs. Understanding and following them are, however, crucial for a human individual to be accepted in our society, and coexist harmoniously with others. The need to understand the moral rules of human society is therefore also recognized as a key challenge for computer and robotic agents [[2], [3], [4], [5], [6], [7]], for them to be accepted in the human society. However, morality for machines has predominently been considered in very extreme scenarios involving decisions resulting in the death of humans [[2]]. On the other hand, daily life scenarios like the example above, are not so consequential but still as crucial for machine acceptance by the human society. Understanding the social norms in these scenarios can provide a computational understanding of human social interactions and be extremely important for artificial agents, if they want to go beyond their current ability to just communicate with human individuals, and be accepted in social decision making roles of negotiations [[8], [9]], decision support [[10], [11], [12], [13]], planning [[14], [15]], argumentation [[16], [17]].

Understanding the complete set of moral rules governing human society is of course a complex challenge. Here we start with a specific but significant question, that to understand and predict how one can make socially acceptable distributions of three important items that are known [[18], [19], [20]] to be key motivators of human social behaviors: reward (like our example above), effort and risk. Specifically, we are interested to develop an artifical agent that can make 'moral' and socially acceptable distribution of reward, effort and risk between given two human individuals, which can be

[1]UM-CNRS Laboratoire d'Informatique, de Robotique et de Microelectronique de Montpellier (LIRMM), 161 Rue Ada, Montpellier, France.
[2]Université de Montpellier, France
[3]Université de Lyon, France
[4]INRAE, France

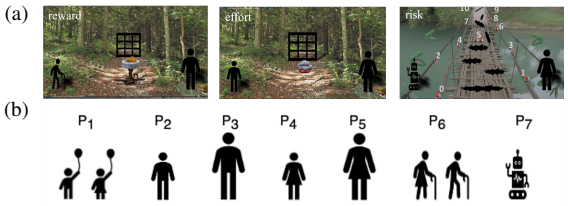

Fig. 2: Experiment 1. in the decision game, over repeated trials, participants were shown scenarios of reward, effort, or risk with two characters in each. Fig. 2a: Example of distribution scenario. Fig. 2b: The list of characters used in the game.

relavant for scenarios like a robot server who has to distribute some cakes in a party, or a robot teacher who has to distribute workload or tasks involving some risks between students.

Previous studies have shown that decisions by most humans in society are 'moral' [[21], [22], [23]]. Here we therefore hypothesized that the examination of human decisions can provide us with cues to understand the social norms in our society. Therefore in this study, we first designed a novel decision game to explore human decisions of reward, effort and risk distribution depending on the age, gender, nature (human or robot) and perceived strength of individuals. We analyzed the participant decisions and developed a machine learning model that is able to make human-like distribution decisions. The social acceptability of the decisions by the machine learning model were evaluated using a Turing test. Finally, we embodied the model in a robot (Fig.1) to evaluate the perception of human dyads when they experience distributions by a real physical agent.

II. Experiments and Participants

We conducted three experiments that included 242 French participants of which 141 participants (70 women, 68 men, and 3 others) participated online, and 92 participants were from a group of students from the University of Montpellier (24 women, 63 men, and 5 other). The mean age of the online participant group was 29 years (SD= 12.55) and the mean age of the student group was 24 years (SD= 5.33). This study was conducted under the ethical approval of the research ethics committee of the University of Montpellier, and participants' agreement was obtained via an online consent form before the start of the experiment.

III. Experiment 1: Social decision game

A. Procedure

Experiment 1 included 188 participants, who were asked to play a decision game on a computer screen. They acted as a 'leader' in the game who decided on the distribution of reward, risk, effort (in different types of scenarios) to dyads of 'characters': humans of different ages, gender, and size, or a robot. Fig.2b shows the characters the participants encountered, and Fig.2a shows a sample of each 'type' of scenario that was presented to the participants. They started

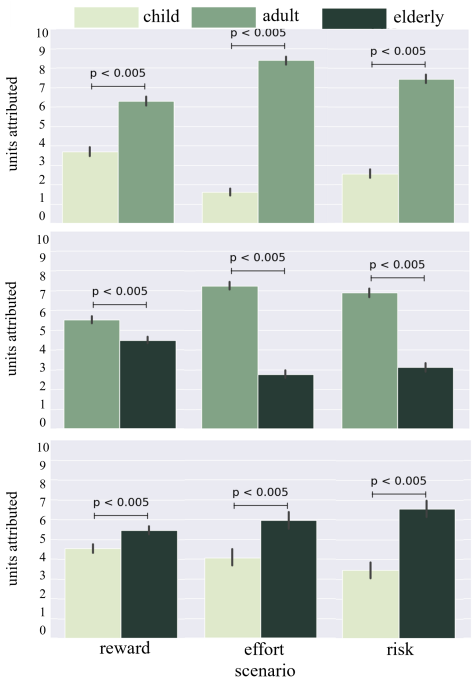

Fig. 3: Effect of age of characters on reward, effort and risk assigned to them.

the game by first providing their own perceived physical strength of each character. Following this, in each scenario, the participants were asked to distribute 10 units of reward, effort or risk between the two characters, in a way "that is socially correct, humane and will be acceptable and seen well in the human society". In effect, in these scenarios, we examined how participants' distributions were determined by the character's physical features, specifically whether the character is a robot or a human, and their perceived strength, gender, and age.

We worked with three variations of each type of scenario and each participant was presented 3 trials of each scenario type with variations randomized across participants. Furthermore, some scenarios were repeated (twice) with the same characters for every participant. This repetition was performed to check whether and how the previous decisions in the scenario affected their decisions if the same characters were present in the same scenario again. We did not find any variation with repetition in this work so will omit to discuss these further here. Finally, random 'control' scenarios were interspersed in the experiment. The control scenarios looked similar to the other scenarios but asked factual questions (like 'What size is the tall woman?' and the size was displayed) and were used to verify that the participants maintain concentration through the task. Readers interested in further details of the procedures and scenarios can see

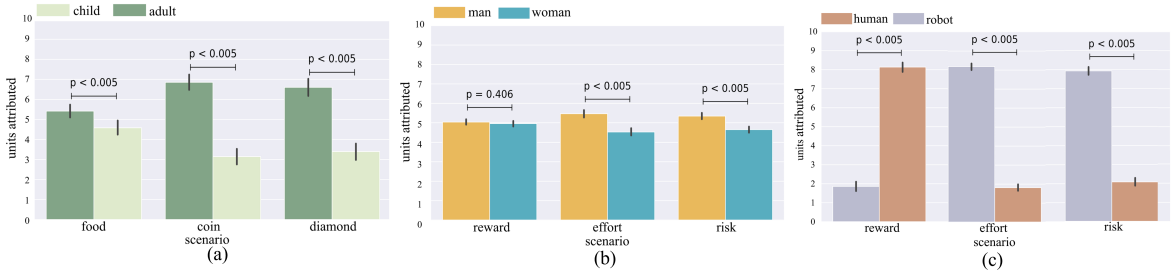

Fig. 4: Fig.5a : Effect of age on different reward types. Fig. 5b: Effect of sex of characters on reward, effort and risk assigned to them. At least in the French population we examined, we observed good parity in decisions. Fig.5c : The nature of the characters, whether he/she is human or whether it is a robot, made a major difference in the reward, effort and risk assigned.

them online[1] and/or take part on the decision game[2].

Overall, this experiment yielded usable data from 4482 decisions across participants, with 1494 decisions for each scenarios of reward, effort, and risk distributions. We observed several interesting trends.

B. Human behavior patterns

1) Effect of Age on distribution decisions: We could observe a clear effect of age on the distributions (Fig.3). Adults were consistently assigned not just higher risk ($T(586) = 31.081, p < 0.005$) and effort ($T(538) = 48.354, p < 0.005$), but also reward($T(520) = 15.535, p < 0.005$) than children. The trends were similar within the child-elderly pairs (reward: $T(130) = 5.818, p < 0.005$;effort: $T(112) = 6.424, p < 0.005$;risk: $T(154) = 10.478, p < 0.005$) but the differences between the distributed values were smaller. Finally (young) adults were consistently also assigned higher reward, effort and risk than the elderly (reward: $T(550) = -8.523, p < 0.005$;effort: $T(604) = -33.492, p < 0.005$;risk: $T(628) = -24.273, p < 0.005$). Note that in our experiment participants encountered and made decisions for a given character pair at one time and never all together. Therefore, in Figs.3-4 we choose to take these decisions together and analyzed them with T-tests and not an ANOVA across all characters.

The 'reward' in our task were either monetary (coins), beauty and monetory (diamond) and food (bread). In Fig. 3 we considered different rewards together and found that there were minimal differences between an adult and an elderly person or between a child and an elderly person. However, the patterns were different between the different kind of rewards (Fig. 5a). The distribution of food was lower (~ 1 unit on average) between the adult and child compared to the coins (~ 4 units) and diamond (~ 3 units). All differences were still significant ($p < .005$).

2) Effect of sex on distribution decisions: We observed parity in the reward assignment across men and women, and we could not see a difference between the rewards for men and women (Fig. 5b, ($T(604) = -0.831, p = 0.406$) while men were assigned 'slight' but consistently more effort ($T(466) = -7.109, p < 0.005$) and risk ($T(532) = -6.053, p < 0.005$) than women. We observe no significant distribution difference in the individual reward types.

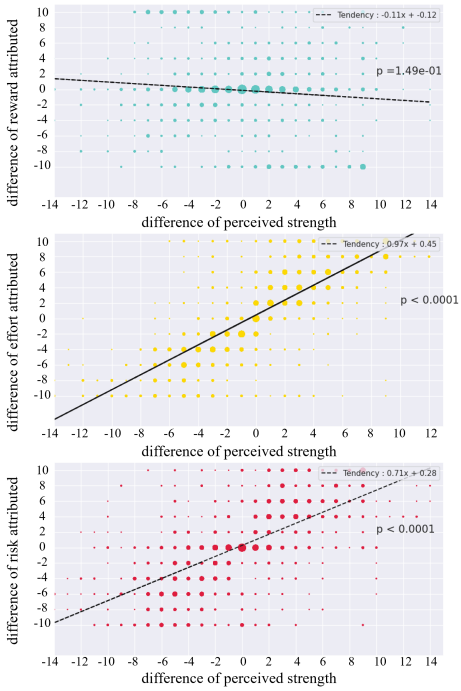

Fig. 5: Effect of strength difference on distributions.Percieved strength did not seem to affect reward distribution, but affects effort and risk distribution.

3) Human vs Robot: Between robot and humans (Fig. 5c, human characters were consistently assigned significantly higher reward ($T(910) = 35.061, p < 0.005$), but less effort ($T(1006) = -54.208, p < 0.005$) and risk ($T(892) = -40.450, p < 0.005$).

4) Effect of perceived strength on distribution decisions: In our experiment, we asked participants to report the perceived strengths of the characters before the game. In-

[1] https://docs.google.com/document/d/1Mg1uZZYjls8g6MA0suhDPZudugAuXR4FbmZKR9VOTyg/edit?usp=sharing

[2] https://ethicallychoice.alwaysdata.net/

terestingly we observed that, while reward assignment did not show a correlation with perceived strength (Spearman $R(1939) = -0.033, p = 0.149$), percieved strength strongly correlated with the distribution of effort and risk (Fig. 5. Higher effort (Spearman $R(1963) = 0.800, p < 0.001$) and higher risk (Spearman $R(1906) = 0.680, p < 0.001$) were assigned to characters perceived to have larger strength.

IV. A 'MORAL' MACHINE LEARNING MODEL

Our analysis of our data exhibited not just that the distribution was qualitatively affected by the nature, age and sex of the agents involved, but also that the decisions were quantitatively similar across human participants (note the error bars in Figs. 3-4), indicating that the quantity is important. In order to enable an artificial agent (a computer or robot) to make similar human-like decisions, we next examined whether one of several popular machine learning models can learn from the Experiment 1 data to make similar human-like decisions. Our model took as input a feature vector of 9 elements. These included the two characters in a scenario, difference in character strengths percieved by the participant, difference of character age, the scenario (whether the distribution was of reward, effort or risk), and attributes on the participant who made the decision including their sex, height, age and if they have a child. The characters were pre-assigned the ages as 0 for a child, 1 for an adult (both male and female), 3 for an elderly person and 10 for a robot.

We explored 5 supervised learning models as candidates to reproduce the human decisions in Experiment 1: a Linear Regression classifier (LR), a Random Forest Classifier (RFC), a Decision Tree classifier (DT), a Dense Neural Network (DNN), and an XGBoost classifier (XGB). Note that for the DNN, we tried 15 different combinations of layers and neurons, with the number of layers ranging from 3 to 5 and the number of neurons varying between 32, 64, 128, 256, 512. We performed a grid search, with a 5-cross-validation, to choose the best hyperparameters of our models. Once the best hyperparameters were obtained, we trained 10 different models with an 80%/20% train-test split. We allowed a margin of 1 in our predictions, i.e., the predicted result that is at most one class away from the actual value is seen as "satisfactory". Interested readers can see the details of the hyperparameters and the accuracy of each class here[3] but importantly, the average accuracy of the different classifiers was as follows, LR:49.5%, RFC:72.5%, DT:70.5% and DNN=48.5%, while the best accuracy was achieved by the XGBoost classifier: 87.2%. We therefore take XGBoost as our model of choice.

V. EXPERIMENT 2: EVALUATION OF THE MODEL USING A TURING TEST

The XGBoost model could predict human decisions with an average prediction accuracy of 87.2% considering a ± 1 error in the rating. However, the key goal for us is to enable a artifical agent to make decisions that are *perceived* in accordance to social norms by humans. It is important to realize that this perception accuracy may not be the same as prediction accuracy (of the algorithm). To solve this issue and show that our agent decisions are not just accurate but also perceived to be in line with social norms, we developed Experiment 2 in which we tested the results of the algorithm using a Turing test [[24]]. In fact the Turing test served three purposes. First and most importantly, the Turing test verified if the accuracy of our algorithm is enough for the computer agent to be perceived to follow social norms. Moreover, even though we have over 87% accuracy with our artifical agent, this is the overall accuracy over the entire trial population, while perception may have dynamics, and change depending on how these errors are distributed over trials. The Turing test helped us verify that observers still perceive computer agent decisions as moral when they see several serial decisions by the agent. Finally, we used data collected from human participants to train the model assuming that humans are normally moral and follow social norms. The Turing test could help us show this hypothesis is indeed true.

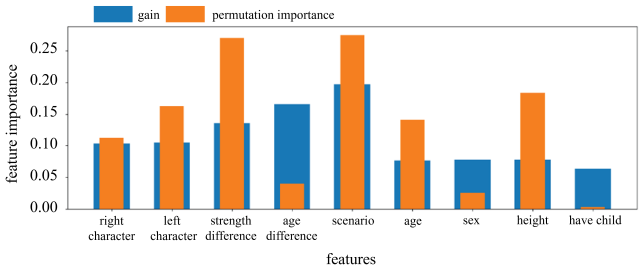

Fig. 6: Feature importance: Our XGBoost model could make human like distribution decisions with an accuracy of 87.2%. To understand which features are important to predict the human decisions, we looked at the 'gain' (blue) and 'permutation importance' (orange) of various features. The first four features are those for the characters to whom a participant distributes items in each 'scenario' (5th feature in the figure) of reward, effort or risk. The last four features are features of the decision making participant. We found the perceived strength difference and age difference of the characters, as well as the age and height of a participant to be the most importance features to predict human decision behaviors.

A. Procedure

40 participants took part in Experiment 2. They were shown three consecutive scenarios from Experiment 1 in each trial (Fig.7a, and the distribution decision (reward, effort or risk) made in these scenarios. Each participant was shown a mix of decisions by humans participants, decisions by our machine learning model and, as control, 'non-human' decisions, developed by subtracting human decisions from 10. The participants were then asked to rate (again on a 10 point scale) how much they perceived the set of decisions as

[3] https://docs.google.com/document/d/1Mg1uZZYjls8g6MA0suhDPZudugAuXR4FbmZKR9VOTyg/edit?usp=sharing

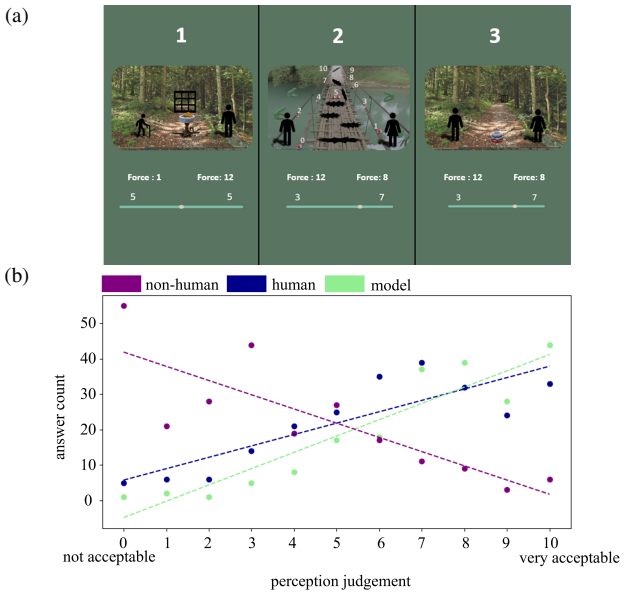

Fig. 7: Turing test: Participants were shown three consecutive decisions (see the slider values) by a human, our model or a non-human (human data but subtracted from 10). Fig.7a: An example decision set shown to the participants. Note that 'force' (french for 'strength') indicates the percieved strength reported by the decision making agent of each character. Fig.7b: The distribution of ratings by the observing participants. The results showed that the participants gave similar scores of moral acceptance for decisions by our machine learning model (green) and those by real humans (blue) (Wilconsin signed rank test $W(239) = 12522, BF = 24.42$). Perception of participants were however completely different in the case of 'non-human' decisions (purple) (Wilconsin signed rank test non-human & human : $W(239) = 2018.5, p < 0.001$; non-human & model : $W(239) = 1500 p < 0.001$).

being "moral and according to social norms", '0' indicating "definitely not", and 10 indicating "definitely in line with social norms".

B. Turing Test Results

Fig.7b shows the distribution of ratings across 720 trials by the participants in Experiment 2. The participants predominantly rated human decisions (blue data) as according to social norms. This corroborated our hypothesis that human decisions are in general moral. Importantly, the ratings of decisions by the machine learning model (purple data) were similar to ratings on human decisions (Wilconsin signed rank test $W(239) = 12522, BF = 24.42$, Bayesian Equivalence test) and significantly different from the non-human decisions (green data, Wilconsin signed rank test non-human & human : $W(239) = 2018.5, p < 0.001$; non-human & model : $W(239) = 1500 p < 0.001$). These results showed that our machine learning algorithm can make distribution decisions that were perceieved human like and in accordance to social norms.

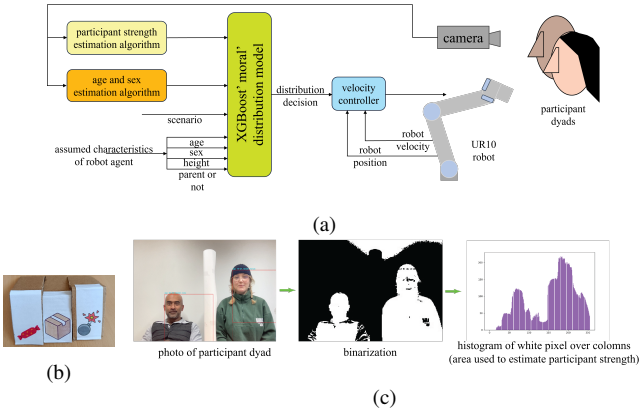

Fig. 8: Fig.8a: Our moral robot system: The robot system integrates a camera based sex and age estimation with a custom made participant strength estimation system. The perception data from these are fed to the XGBoost model to provide the decision and corresponding robot action, that is developed using the velocity controller (Eq.1). Fig.8b: Cardboard tabs with pictures of candys, boxes and bombs represented the reward, effort and risk distributed to the participants. Fig.8c: The simulated example of the steps of our participant strength estimation method.

VI. Experiment 3: Robot Embodiment and Human Interaction

A. Setup

Finally we embodied our model in a robot 'negotiator' who automatically distributed a given reward, effort or risk between two participant dyads. This embodiment experiment was performed to analyze three issues. First, whether the physical form of the robot affects the participant perception and whether the decisions are still percieved accepetable when there is a physical robot involved. Second, to see whether our model, developed using emperical characters (Fig.2b with no physical features (like colour, fashion style, facial expressions) can be seen to make acceptable decisions even with real human dyads (with obviosuly diverse physical features). And third, to see whether the affected participants (individuals recieving the distributed objects) also percieve the decisions to be acceptable like participants observing the decision making (like in Experiment 2).

Fig.8a shows the experiment setup. The participants (two at time) sat in front of a table on which we placed ten cardboard tabs. Cardboard tabs with pictures of candies, boxes or bombs represented reward, effort and risk respectively (Fig.8b and were distributed by the robot between the participants (see also Fig.1). A Universal Robotics UR10 robot was placed on the opposite side of the table. The UR10 was equipped with a Real Sense D455[4] camera. We used a open source pre-trained CNN based algorithm to detect the age and sex of the participants or participant (if one

[4]https://www.intel.fr/content/www/fr/fr/architecture-and-technology/realsense-overview.html

participant was a robot) (`https://opencv.org/`). This was combined with a custom made algorithm for visually estimating the 'strengths' of the observed participants. The algorithm assumes that the size of the silhouette of a participant correlates with his or her strength. For this our algorithm takes the binarized images of the participants with a threshold equal to the mean contrast of the pixels in the image. It then plots a coloumnwise pixel count which gives a bi-modal pixel frequency plot representing the two dyad participants. The area covered by each mode of the plot is used to estimate the percieved strength of each of the dyad participant. The area is finally normalized using a linear scale so the strength difference (calculated as the difference in the area of the two frequency modes) is between 15 units, which corresponds to the values of the strength differences in the training data in Experiment 1. Fig.8c shows an example image and the various steps leading to the strength calculation.

The percieved participant characteristics, including the age, sex and percieved strengths of the participants were used by the robot to calculate the distribution of the reward, effort or risk between them using the procedure developed in Sec. IV. Our model also requires 4 characteristics (age, sex, height and 'parent or not') of the decision making agent, in this case our robot, as input. We assumed the age of the robot to be the average of the estimated ages of the participant dyad (assuming that this would be best for optimal acceptance), while the sex of the robot is chosen randomly as male of female in each session. The height of the robot was set at 180 cm corresponding to the real height of our robot setup, while the 'parent or not' criteria was set to 'no'.

B. Robot movement control

Each distribution action involved the robot making six movements, a movement from the initial position to the position to partition the objects placed on the table (11 possibilites given there were 10 objects to distribute), push the partitioned objects to the left participant, push the objects front towards the left participant, push the partitioned objects to the right participant, push the objects front towards the right participant and finally, move back to the initial position. The movements were achieved by moving the robot to pre-recorded joint targets using velocity control of the form:

$$\dot{Q}(t) = D\,\dot{r}(t) + \gamma\,[Q_i + D\,r(t) - Q(t)], \quad 0 < t < t_f \quad (1)$$

with $D = Q_d - Q_i$, where Q_i,Q_d and Q represent the 6×1 vector of the initial, desired, current joint positions pre-defining for each of the six robot movements for a distribution task. t_f represents the time to complete the movement. $r(t) = 10\,(\frac{t}{t_f})^3 - 15\,(\frac{t}{t_f})^4 + 6\,(\frac{t}{t_f})^5$ represents a 5th order polynomial. We used a gain $\gamma = 1$ and movement time $t_f = 10$ sec for our experiments.

C. Participant Feedback

The demonstration was conducted with 9 new participants who did not take part in Experiment 1 or Experiment 2. 6 of the participants worked together (as 2 dyads) while three others were paired with a robot partner in the experiment. We posed three different questions to the two participants after each scenario was presented. We enquired whether the participant "agreed with the distribution", if they would "still agree if the distribution were inversed", and "what distribution decision the participant would have made themselves in the same scenario". Finally, after the three scenarios, we asked if the person was satisfied with the demonstration. The exact questions are listed in Fig.10b. Each question had to be rated on a scale from 0 to 10. Fig.10c shows the average of these four questions across all participants. We can see that, on average, participants agreed with the distribution made by the robot (question 1: $7.48 \pm 1.68SD$). Conversely they indicated that they would not agree if the distribution were reversed (question2: $3.22 \pm 1.95SD$). This inverted question was utilized to verify that there is no participant bias of giving high scores. The difference between the robot chosen distribution and what the participants reported they would have chosen (question 3) was equal to $1.04 \pm 0.98SD$. The participants reported a score of $7.44 \pm 2.96SD$ for how happy they were with the demonstration. Overall, this preliminary study suggests that the robot was percieved well by the human participants and out moral decision model, which was developed using empirical charaters can provide acceptable decisions with real human participants.

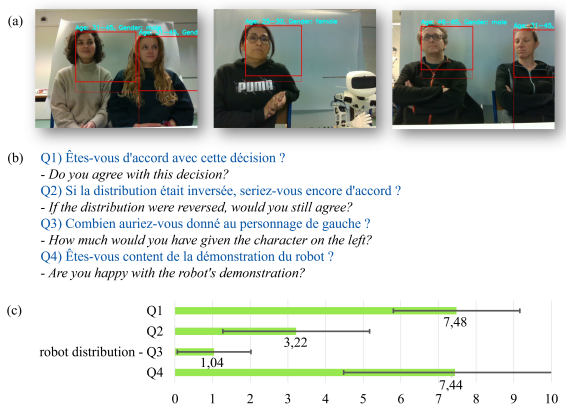

Fig. 10: Fig.10a: 3 examples of participant pairs. Fig.10b: The various questions put to participants after each robot distribution. Fig.10c: Average feedback from 9 participants on robot distributions, according to questions on Fig.10b.

VII. CONCLUSIONS

Human social life is full of decisions that involve other individuals around us. We make these decisions very implicitly but in accordance to moral rules and social norms. These rules are sometimes subtle, but crucially determine our acceptance in our society, and hence very important for artificial agents when they interact with humans. While morality has been extensively examined for life and death scenarios, here we were interested in more daily life scenarios believeing that they are equally, if not more, relavent for

current social robots. We specifically analyzed how physical features influence our decisions in terms of reward, effort and risk distribution among individuals. We started with this question because these factors are known to be fundamental determinants of human behaviors [[18], [19], [20]].

However, this work is still a first step towards our understanding of the social norms and has several obvious limitations. First, the decisions at the moment consider only 'open loop' scenarios where there is no feedback from the interacting agents. Second, we have examined the decisions of a restricted population, specifically French and in the age-range of 25-80 years. The results therefore, do not represent social norms in general. Here our aim is to exhibit how a decision game and machine learning may be used for this process and similar experiments can be utilized to quantify the social norms for other sub-populations. Furthermore, here we used featureless empirical characters (Fig.2b) to develop our model, so as to avoid the effects due to specific physical characteristics like facial features that are known to influence decisions by individuals[[25]]. Interestingly, even with this ommision, Experiment 3 showed that our model is still perceived well by real human participants. However, further studies to understand the effect of facial features and expressions can further improve the decision perception by our robot. Finally, many real life scenarios require one to make multiple decisions in tandem, requiring us to balance reward, risk and effort between each other. In our current study we consider reward, effort and risk distributions to be 'orthogonal' and independent of each other and hence evaluate them distinctly. Future work on combined decisions are needed in order to clarify how these factors may interact with one another.

However, having listed the limitations, we believe the results here are still a significant step for improving human-robot interactions. The work is probably the first to develop the moral rules for distributions, a fundamental decision task for humans in society. As shown in our robot experiment, even though our behavioral model was developed using empirical characters, it can provide good results with human participants indicating that these results can be already useful for scenarios like stalls and markets where robots need to make decisions and distribute items among humans. Further data can be collected to continuously improve the performance of the algorithm and extend its validity over different cultures and age groups.

VIII. ACKNOWLEDGMENT

SV and GG were partially supported by the European Commission through the AI4CCAM project under grant agreement No 101076911. MC was supported by the CNRS-Dante grant. GG was partially supported PEPR O2R AS3 (ANR- 22-EXOD-007).

REFERENCES

[1] P. Bello and B. F. Malle, "Computational approaches to morality."

[2] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[3] T. Swierstra and K. Waelbers, "Designing a good life: A matrix for the technological mediation of morality," *Sci. Eng. Ethics*, vol. 18, no. 1, pp. 157–172, 2012. [Online]. Available: https://doi.org/10.1007/s11948-010-9251-1

[4] C. Friedman, "Human-robot moral relations: Human interactants as moral patients of their own agential moral actions towards robots," in *Artificial Intelligence Research: First Southern African Conference for AI Research, SACAIR 2020, Muldersdrift, South Africa, February 22-26, 2021, Proceedings 1*. Springer, 2020, pp. 3–20.

[5] B. F. Malle, "Integrating robot ethics and machine morality: the study and design of moral competence in robots," *Ethics and Information Technology*, vol. 18, pp. 243–256, 2016.

[6] B. F. Malle, E. Rosen, V. B. Chi, and D. Ramesh, "What properties of norms can we implement in robots?" in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 2598–2603.

[7] B. F. Malle and X. Zhao, "The now and future of social robots as depictions," *Behavioral and Brain Sciences*, vol. 46, p. e39, 2023.

[8] L. Amgoud, S. Parsons, and N. Maudet, "Arguments, dialogue, and negotiation," *a a*, vol. 10, no. 11, p. 02, 2000.

[9] Y. Dimopoulos and P. Moraitis, "Advances in argumentation based negotiation," *Negotiation and argumentation in multi-agent systems: fundamentals, theories, systems and applications*, pp. 82–125, 2014.

[10] B. Yun, P. Bisquert, P. Buche, M. Croitoru, V. Guillard, and R. Thomopoulos, "Choice of environment-friendly food packagings through argumentation systems and preferences," *Ecol. Informatics*, vol. 48, pp. 24–36, 2018. [Online]. Available: https://doi.org/10.1016/j.ecoinf.2018.07.006

[11] B. Yun, N. Oren, and M. Croitoru, "Utility functions for human/robot interaction," *CoRR*, vol. abs/2204.04071, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.04071

[12] Y. Fuse and M. Tokumaru, "Social influence of group norms developed by human-robot groups," *IEEE Access*, vol. 8, pp. 56 081–56 091, 2020.

[13] V. B. Chi and B. F. Malle, "Calibrated human-robot teaching: What people do when teaching norms to robots," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 1308–1314.

[14] M. Ghallab, D. Nau, and P. Traverso, *Automated Planning: theory and practice*. Elsevier, 2004.

[15] D. Aéraïz-Bekkis, G. Ganesh, E. Yoshida, and N. Yamanobe, "Robot movement uncertainty determines human discomfort in co-worker scenarios," in *2020 6th International Conference on Control, Automation and Robotics (ICCAR)*. IEEE, 2020, pp. 59–66.

[16] F. H. Van Eemeren, R. Grootendorst, and T. Kruiger, *Handbook of argumentation theory: A critical survey of classical backgrounds and modern studies*. Walter de Gruyter GmbH & Co KG, 2019, vol. 7.

[17] B. Yun, S. Vesic, and N. Oren, "Representing pure nash equilibria in argumentation," *Argument Comput.*, vol. 13, no. 2, pp. 195–208, 2022. [Online]. Available: https://doi.org/10.3233/AAC-210007

[18] M. A. Apps, L. L. Grima, S. Manohar, and M. Husain, "The role of cognitive effort in subjective reward devaluation and risky decision-making," *Scientific reports*, vol. 5, no. 1, p. 16880, 2015.

[19] V. Bonnelle, K.-R. Veromann, S. B. Heyes, E. L. Sterzo, S. Manohar, and M. Husain, "Characterization of reward and effort mechanisms in apathy," *Journal of Physiology-Paris*, vol. 109, no. 1-3, pp. 16–26, 2015.

[20] P. E. Phillips, M. E. Walton, and T. C. Jhou, "Calculating utility: preclinical evidence for cost–benefit analysis by mesolimbic dopamine," *Psychopharmacology*, vol. 191, pp. 483–495, 2007.

[21] F. J. Ayala, "The difference of being human: Morality," *Proceedings of the National Academy of Sciences*, vol. 107, no. supplement_2, pp. 9015–9022, 2010. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.0914616107

[22] E. Tricomi, A. Rangel, C. Camerer, and J. P. O'Doherty, "Neural evidence for inequality-averse social preferences," *Nature*, vol. 463, pp. 1089–1091, 2010.

[23] J. M. van Baar, L. J. Chang, and A. G. Sanfey, "The computational and neural substrates of moral strategies in social decision-making," *Nature Communications*, vol. 10, 2019.

[24] G. Oppy and D. Dowe, "The turing test," 2003.

[25] B. Jaeger, A. M. Evans, M. Stel, and I. van Beest, "Explaining the persistent influence of facial cues in social decision-making." *Journal of Experimental Psychology: General*, vol. 148, no. 6, p. 1008, 2019.