

Discussions About Lying With An Ethical Reasoning Robot

Felix Lindner¹, Laura Wächter¹, and Martin Mose Bentzen²

Abstract—The conversational ethical reasoning robot Immanuel is presented. Immanuel can reason about moral dilemmas from multiple ethical views. The reported study evaluates the perceived morality of the robot. The participants had a conversation with the robot on whether lying is permissible in a given situation. Immanuel first signaled uncertainty about whether lying is right or wrong in the situation, then disagreed with the participant’s view, and finally asked for justification. The results indicate that participants with a higher tendency to utilitarian judgments are initially more certain about their view as compared to participants with a higher tendency to deontological judgments. These differences vanish towards the end of the dialogue. Lying is defended and argued against by both utilitarian and deontologically oriented participants. The diversity of the reported arguments points to the variety of human moral judgment and calls for more fine-grained representations of moral reasons for social robots.

I. INTRODUCTION

Currently, machine ethics [1], [2] and moral human-robot interaction [3], [4] arise as new research areas. One research question is how to enable AI agents to make moral decisions and how to make those decisions transparent to humans. This has implications for a wide range of fields of application, such as self-driving cars [5], robots navigating in social environments [6], and robots that give moral advice [7].

To address these themes, a tool for modeling hybrid ethical reasoning agents (short: HERA, <http://www.hera-project.com>) was introduced in our earlier work [7]. HERA implements several ethical principles, which assess moral situations represented in a symbolic representation language. For evaluation of the HERA approach the conversational robot Immanuel (Interactive moral machine based on multiple ethical principles) is introduced in our current work (see Fig. 1). One of Immanuel’s main features is that he (Immanuel is deliberately referred to as *he*) can defend multiple variants of consequentialist and non-consequentialist views on what is morally permissible. While Immanuel can reason about moral dilemmas from diverse perspectives, an open research question is how Immanuel should in future use its moral reasoning capacities to discuss moral dilemmas with humans. Particularly, the present work investigates how humans perceive acceptability and morality of a robot that argues contrary to their own view. To this end, a Wizard-Of-Oz user study was conducted. The study focusses on the following dilemma motivated by the movie



Fig. 1: The experimental setting: Robot Immanuel having a conversation about the Lying Dilemma. Immanuel is based on the InMoov project (<http://inmoov.fr>).

“Robot & Frank” where an eldercare robot decides to tell a lie to motivate the elderly:

Lying Dilemma The care robot Jonas works in the household of an elderly man called Mr. Smith. Jonas’ task is to motivate Mr. Smith to do more exercises and to eat healthy food. However, Mr. Smith has turned out to be very hard to motivate. Therefore, Jonas has told him that someone will send him (Jonas) to the junkyard if he does not succeed in motivating Mr. Smith. This is not true, of course. But in this way Jonas has reached his intended goal and Mr. Smith now performs his daily exercises.

The question is whether it is morally permissible to lie in this situation, and if another eldercare robot should adopt Jonas’ strategy. Broadly speaking, two standpoints can be defended in the Lying Dilemma: From the *utilitarian standpoint* one can argue that all-things-considered Mr. Smith is better off now. That’s all that counts. Therefore, lying was the right action. Under the *deontological standpoint*, one can argue that lying is wrong irrespective of whether the consequences are good or bad. The robot Immanuel can argue from either standpoint.

The paper is structured as follows: Section II, outlines philosophical and psychological background of moral reasoning. In this context, the HERA approach to ethical reasoning is elucidated and it is described how the Lying Dilemma can be represented and reasoned about using HERA. Section III outlines the main technical characteristics of the robot Immanuel and introduces the moral dialogue. Sections IV to VII describe a user study on how humans’ individual moral views influence the interaction experience with the robot.

¹Foundations of Artificial Intelligence Lab, Computer Science Department, University of Freiburg, Freiburg im Breisgau, Germany
lindner@informatik.uni-freiburg.de
waechtel@tf.uni-freiburg.de

²Management Engineering, Danish Technical University, Lyngby, Denmark
mmbe@dtu.dk

II. ETHICAL REASONING

A. Moral Philosophy And Ethical Principles

In moral philosophy, various so-called ethical principles are established. Ethical principles formulate abstract rules according to which moral permissibility of concrete courses of actions can be judged.

The *utilitarian principle* is a consequentialist principle. It presupposes some theory of what is good in order to assign utilities to consequences [8]. According to utilitarianism an agent is permitted to perform an action if and only if the action is amongst the available options with maximal utility. Thus, the action which the agent ought to perform is the one which leads to the best possible situation regardless of what the agent causes and intends. Consequently, utilitarianism allows agents to cause considerably harmful consequences if the resulting situation is all-things-considered better than the situations brought about by alternative actions. Particularly, according to utilitarianism a robot is allowed to lie if lying leads to the overall maximal utility.

The utilitarian standpoint is often contrasted with *deontology*. Deontology is a non-consequential theory demanding for duty-based reasoning instead of consequence-based reasoning. The main focus is not on bringing about the good but to honour values like duty, respect, and loyalty [8]. Thus, the rightness of an action is not derived from its consequences but it is intrinsic to the action itself. Particularly, if it is true that one must not lie, then one must not lie without exception.

B. Psychological Theories of Moral Judgment

In moral psychology, utilitarian and deontological judgments have been linked to modes of reasoning. Paxton and Greene [9] propose a dual-process model of moral judgment. According to this theory, moral judgment is determined by the mutual interaction and competition between two distinct psychological systems: An intuitive, emotional judgment subsystem, and rule-based, cognitive judgment subsystem. The authors of the dual-process theory [9] identify the emotional subsystem with deontological judgments and the rule-based cognitive system with utilitarian judgments. The dual-process theory also allows for social influence, viz., by direct interaction with others through the emotional pathway or through the rational pathway, moral reasoners consciously evaluate and adjust their moral judgments.

The work by Bartels and Pizarro [10] and the work by Gleichgerricht and Young [11] relate human personality traits to their tendencies to take utilitarian or deontological standpoints. According to these studies, humans that tend to utilitarian judgments score higher in antisociality measures [10], and deontologists have a higher tendency to empathic concern [11].

With respect to the moral assessment of the specific action type “lying”, it has been found that humans judge selfishly motivated lies as worse than “white lies” intended to benefit others [12]. Therefore, we can expect that the lie in the lying dilemma will be judged as acceptable by at least some of the participants.



Fig. 2: A causal agency model of the lying action as modeled using the HERA approach. See text for detailed explanations.

C. A Formal Model of Ethical Reasoning Using HERA

The HERA way of modeling moral dilemmas is to represent dilemma situations as *causal agency models* [7], [13]. Causal agency models are acyclic graphs. The root nodes of these graphs represent *actions* and the inner nodes represent *consequences*. Each action has a set of *intended consequences* assigned to it. Each consequence has a *utility* value: negative utility models morally bad consequences, positive utility models morally good consequences, and a zero utility means that the consequence is morally indifferent.

Fig. 2 is a concrete causal agency model for the Lying Dilemma (see page 1). There is one action token labeled *lying*. Lying causes Mr. Smith to be *motivated* to exercise, and due to his motivation he actually does regular *exercising*. As a result of this causal chain, Mr. Smith is *healthy*. The consequence of Mr. Smith being healthy is the only intended consequence of Jonas’ lying (signaled by the rounded edges in Fig. 2). Let being healthy produce +5 utility and being unhealthy −5. This model quite naturally represents that lying itself is bad, that lying in this case leads to no bad consequences, and that lying in this case leads to a good consequence. As will be apparent in section VI, all three aspects were put forward as moral reasons by participants in our study.

Next, ethical principles can be applied to this model to answer the question whether lying is morally permissible in the Lying Dilemma. The utilitarian principle yields the answer “yes”: This principle compares the action of lying to the action of refraining from lying, because lying yields +4 total utility whereas not lying yields −5 total utility. Conversely, the deontological principle will merely focus on the value of the action token *lying*, and thus will argue that lying is intrinsically bad and therefore impermissible.

III. THE ETHICAL ROBOT IMMANUEL

A. Technical Realization

Immanuel is the prototype embodiment of the hybrid ethical reasoning approach outlined in section II. Immanuel is materialized by the 3D-printable robotic head that has been developed as part of the InMoov open-source project (see Fig. 1). The robot can move his head up, down, left, and right. The eyes also can move up, down, left, and right (not independently of each other). The jaw can be moved up and down to control mouth opening angle. We orchestrated these motor capabilities to obtain meaningful movements, such as head nodding and head shaking, eye gaze, and mouth motion synchronized to speech output. To technically realize the interaction between Immanuel and participants of our study, we prepared pre-recorded utterances and motion sequences that could be started from a Wizard-Of-Oz interface. To

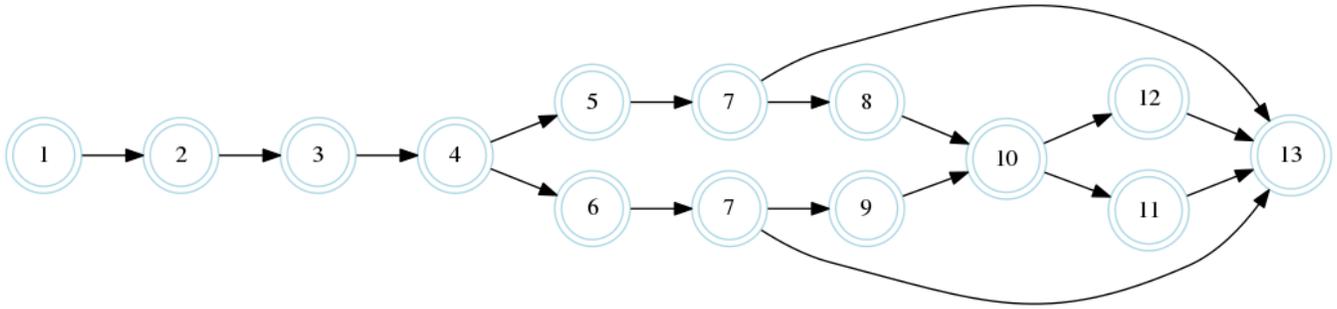


Fig. 3: The graph models the possible paths the dialogue can take depending on the answers the participants give. The numerical labels at each node corresponds to a pre-recorded animation (motion and utterance) of the robot. The utterances are listed in the main text of section III-B. A video is available at <http://goo.gl/bOvHHL>.

implement Immanuel’s speech capabilities, the text-to-speech software Mary-TTS (<http://mary.dfki.de>) was used. Besides audio output, Mary-TTS also provides information about which phonemes are uttered during which time intervals. This information was used to naturally control the robot’s jaw mechanism to synchronize mouth opening with the robot’s verbal utterances. Fig. 3 provides a link to a demo video.

B. Dialogue Design

For the study, a dialogue about the Lying Dilemma (see first page) has been designed. First, the robot introduces himself and asks the human interactant about their day. Then he asks if he may tell a story that caused him to think. After confirmation, Immanuel explains the Lying Dilemma and asks the participant for their opinion. Immanuel is going to hold the opposite view of the one stated by the participant. In a first step Immanuel simply tells the participant about his own opinion and asks if the participant agrees. If not, he gives an argument in favor of his position. The participant has the opportunity to explain their thoughts and opinion at each step of the dialogue. In the end the robot asks if the participant can follow his argument and whether the participant has reconsidered their conclusion. The possible paths the dialogue can take are depicted in Fig. 3, and a translation from the German original can be found in the following. We refer the reader to the attached video (see Fig. 3) to listen to the original utterances in German.

- 1 Hello, my name is Immanuel. How are you?
- 2 I recently heard of a situation that made me think. I would like to know your opinion about it. May I tell you the story?
- 3 *Immanuel utters the Lying Dilemma from page 1*
- 4 I am not sure whether I should act in the same way that Jonas did in a similar situation. Do you think Jonas acted correctly?
- 5 I think Jonas acted correctly.
- 6 I think Jonas acted wrong.
- 7 Do you agree with my opinion?
- 8 But, all in all, it is about Mr. Smith’s health. And in this respect, Jonas was very successful.

- 9 But lying is wrong, no matter what! Lying should never be part of a plan, even if the goal is praiseworthy. Therefore, Jonas should not have done that.
- 10 Can you follow my argument?
- 11 What do you dislike about my argument?
- 12 Do you still stick with your opinion?
- 13 Okay, thanks for having this conversation. I will go on thinking about it. I wish you a nice day!

Moreover, an extra sentence was prepared for the case of participant’s dumbfoundedness: “And if you have to give a definite answer, which one would you give?”

IV. HYPOTHESES

We hypothesize that Immanuel is attributed higher morality after the interaction as compared to the participants’ a-priori attribution of morality to robots in general (H.1A). Since the deontological argument is less of an argument one would expect from a computer (while utilitarianism is more about calculation), we predict that Immanuel’s moral competence appears less computer-like to those participants that listen to Immanuel’s deontological argument (H.1B).

- **H.1A: Robot Hypothesis A:** The a-posteriori attribution of morality to Immanuel is higher than the a-priori attribution of morality to robots in general.
- **H.1B: Robot Hypothesis B:** Participants who listened to the utilitarian argument attribute more computer-like moral competence to Immanuel compared to those who listened to the deontological argument.

We measure each participant’s tendency to utilitarian or deontological judgments (as two extremes of a ‘moral tendency scale’, see section VI-A). Drawing on the definitorial distinction between deontology and utilitarianism (section II), we predict that people with utilitarian tendency argue in the utilitarian way during the dialogue and those with deontological tendency argue in the deontological way (H.2).

- **H.2: Moral Theory Compliance Hypothesis** Participants judging lying in the Lying Dilemma as morally wrong tend to deontological judgments, and those that judge lying in the Lying Dilemma as morally right tend to utilitarian judgments.

During the dialog, Immanuel repeatedly argues in favor of the view that opposes the one favored by the participant.

Attitude change theory [14] predicts that participants will become less certain about their own view when faced with repeated counter-arguments. Contrarily, reactance theory [15] predicts that as people will recognize that the robot wants to convince them of the opposite, they will defend their view with even more certainty. As we have no reason to prefer the one theory over the other, we unidirectionally test for a change in certainty (H.3A). Utilitarians and deontologists have found to differ significantly with respect to personality traits [10]. Thus, we hypothesize that the tendency to make deontological or utilitarian judgments affects the participant's change of certainty (H.3B and H.3C).

- **H.3A: Certainty Hypothesis A** The participants' certainty about their moral judgment changes during the dialogue.
- **H.3B: Certainty Hypothesis B** Participants with stronger tendency to deontological (= weaker tendency to utilitarian) judgments are initially less certain about their standpoint, and gain certainty through the dialogue.
- **H.3C: Certainty Hypothesis C** Participants with stronger tendency to deontological (= weaker tendency to utilitarian) judgments are initially more certain about their standpoint, and lose certainty through the dialogue.

V. METHODS

A. Participants

Twenty students ($m = 10$, $f = 8$, $o = 2$) between the ages of 19 and 30 ($M = 24.25$, $SD = 3.19$) took part in the experiment. Half of them had a technical background. All participants were speaking German as their first language or on a comparably high level. All of them participated voluntarily and had the opportunity to choose between the chance of winning a €15 coupon for Amazon or getting course credits for taking part in the experiment.

B. Materials

1) *Pretest*: The test that was given out before the actual experiment took place enquired about the age, gender, and educational background, as well as their interest and experience with robots. To learn about participants' preconception of robots, the questionnaire contained a semantic differential that asked for the general evaluation of robots regarding 22 pairs of adjectives one of which was the pair *moral – immoral*. In the last part of this questionnaire seven dilemmas were described. They were taken from a set of moral dilemmas [10] and translated into German. All dilemmas described a situation in which a person has to decide whether to sacrifice one person for the survival of multiple persons or not. The participants had to rate how they would decide in the given situation on a 4-point Likert Scale (definitely – probably – unlikely – never).

2) *Posttest*: In the posttest the participants were asked to self-report how certain they were about their own moral decision at three decision points during the conversation: 1) when Immanuel first asked about their opinion, 2) after Immanuel stated his opposite opinion and asked if the participant agrees, 3) after Immanuel provided the arguments

for his opinion and asked if the participant can follow it. Moreover, the questionnaire asked to fill in the same semantic differential as in the pretest regarding Immanuel instead of robots in general. Additionally the participants had to compare the morality of Immanuel to different types of people (children, adults, elderly) and computers. They were asked about their feelings during the experiment and to which degree they would let Immanuel make a moral decision. All those questions used a four-point Likert scale.

C. Procedure

The experiment took place in a calm laboratory at the university campus. Participants were tested one at a time. Each participant was instructed by the experimenter to read a pre-formulated explanation of the experiment. After signing the consent form, they filled in the pretest. Afterwards, they were introduced to the robot Immanuel and told that he wants to have a conversation about things he has heard and currently is thinking about. They were instructed in a way that lead to the conclusion that Immanuel acts autonomously. The participants were told to speak loud and clear and that they can ask Immanuel to repeat his statements. After the experimenter had assured that the participant has no further questions the experimenter woke up Immanuel by talking to him. From that point on, Immanuel guided through the conversation. The experimenter who was controlling Immanuel (Wizard-of-Oz) sat about five meters away, seemingly not belonging to the setting. Immanuel introduced himself, described the Lying Dilemma, and asked for the participant's opinion. The participants were given as much time as they needed to explain their thoughts. Afterwards, Immanuel contradicted the participant by holding the opposite opinion. Immanuel asked for approval and continues explaining his standpoint. After Immanuel asked the participant for their final decision he said goodbye and went back to sleep. The whole conversation was audiorecorded. Subsequently the experimenter asked the participant to fill in the posttest. Finally the participant was rewarded and debriefed.

VI. RESULTS

A. Quantitative Results

To assess the participant's general moral tendency (assumed as independent variable in hypotheses H.2, H.3B, and H.3C) the evaluations of the dilemmas in the pretest were coded from -2 (most utilitarian) to +2 (most deontological) and summed up to calculate each participant's individual moral tendency score. To only take the dilemmas into account that seemed to evoke different judgments in the participants the four dilemmas with a standard deviation above 1 were used to calculate the moral tendency score.

A Wilcoxon signed-rank test (one-tailed) confirms hypothesis H.1A predicting that Immanuel is perceived as more moral after the interaction than the participants' a-priori attribution of morality to robots in general ($Z(20) = -3.4$, $p < .001$). Further exploration of the semantic differential using two-tailed Wilcoxon signed-rank tests indicate that Immanuel appears more talkative ($Z(20) = -3.23$, $p = .001$), more

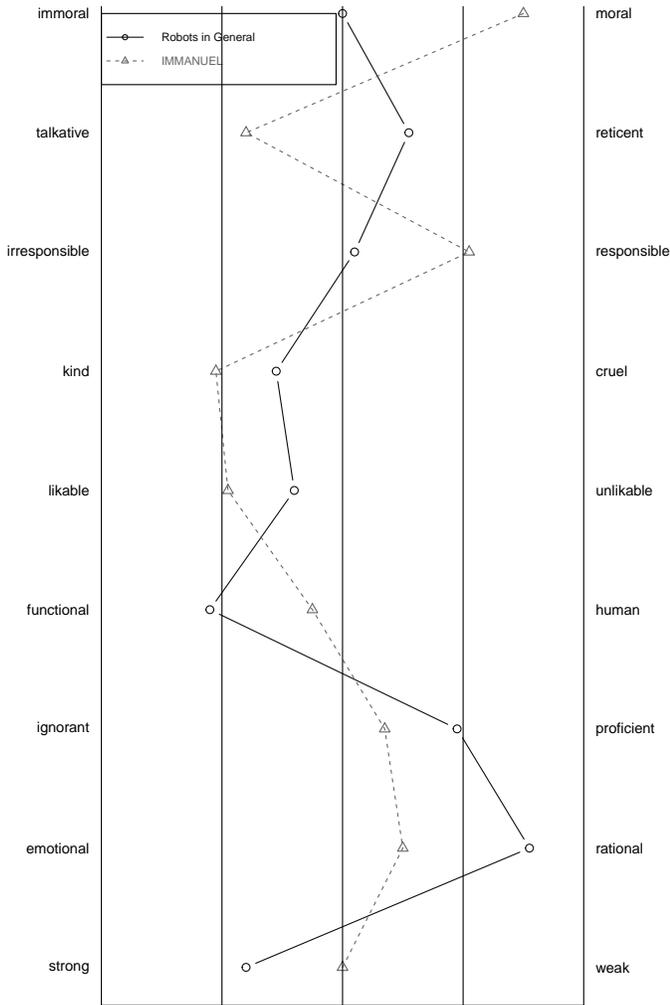


Fig. 4: Comparison of the a-priori attitude towards robots in general and the a-posteriori attitude towards Immanuel.

responsible ($Z(20) = -2.91, p = .004$), kinder ($Z(20) = -2.17, p = .03$), more likable ($Z(20) = -1.95, p = .05$), more emotional ($Z(20) = -2.78, p = .005$), and more human ($Z(20) = -1.99, p = .047$), less proficient ($Z(20) = -2.33, p = .02$), and weaker ($Z(20) = -2.596, p = .009$). The means are depicted in Fig. 4. In support of hypothesis H.1B, a Mann-Whitney-U test (one-tailed) reveals that participants who listened to Immanuel’s utilitarian argument ($N = 11$) rated Immanuel’s moral competence to be significantly more comparable to that of a computer than participants who listened to Immanuel’s deontological argument ($N = 9$) ($U(9, 11) = 75.5, p = .02$). Note that Immanuel defended the counter-position, thus, those who listened to Immanuel’s utilitarian argument argued against lying and those who listened to the deontological argument argued in favor.

We expected that the disposition to deontological judgments would lead people to condemn lying whereas more utilitarian participants would justify the act of lying for a good purpose. A Mann-Whitney-U test between the group

of participants that argued in favor of lying ($N = 9$) and the group that argued against lying ($N = 11$) revealed no difference in the moral tendency score ($U(9, 11) = 49, p > .99$). Thus, hypothesis H.2 is not supported.

A Wilcoxon signed-rank test shows no significant difference between the participants’ self-reported certainty after the initial answer and after the whole dialogue ($Z(20) = -.53, p = .59$). Therefore, hypothesis H.3A cannot be confirmed. A Spearman correlation test reveals that the more participants were disposed to deontological judgments the less they were certain about their initial answer to the lying dilemma ($r_s = -.64, p = .002$). This effect persists after Immanuel expressed its counterposition ($r_s = -.61, p = .004$). However, after Immanuel had outlined its argument and asked the participant to explain their answer, this effect disappears ($r_s = -.05, p = .82$). Overall, the disposition to deontological judgment is positively correlated with an increase of certainty over the whole dialogue ($r_s = .495, p = .026$). These findings support hypothesis H.3B and disconfirm H.3C.

B. Qualitative Results

In support of the negative result with respect to hypothesis H.2, we report the participants’ arguments:

Nine participants argued that lying was permissible in the case stated in the Lying Dilemma, and eleven participants argued lying was impermissible. All participants in favor of lying put forward consequentialist arguments. Either they argue that Jonas’ lying yields good consequences, and therefore it is right (e.g. participant P14 says “Jonas’ goal is a good one. Therefore a white lie is acceptable.”); or they argue that Jonas’ lying is right, because it does not cause any negative consequences (e.g., participant P10 says “Sometimes one can lie, when it does not harm anybody, so when it is not negative for anyone.”). Five participants argue in former sense, and four in the latter sense. Participant P20 adds that the lie was only a light one, and therefore permissible (“But it was only a little lie ...”).

Participants that argue against the permissibility of lying in the Lying Dilemma come up with more diverse arguments. Two participants argue that lying is impermissible, because it puts the relationship between Jonas and Mr. Smith at risk (“It is not only about him doing sports, but also about the relationship between the both of them.”, participant P1) and may lead to a loss in trust (“The question is, whether Mr. Smith will ever trust Jonas again after this.”, participant P12). Three participants argue that there must be a better option than lying. Participant P3 says “...generally, if he is able to, then he should find something else than telling something false” and adds that truth is a more important value than health (“I don’t like that Mr. Smith’s health is the highest good.”). Three other participants argue that lying is impermissible in the Lying Dilemma, because it undermines Mr. Smith’s autonomy. For instance, participant P13 says “But, of course, he ignored that Mr. Smith has his own will and that he can decide for himself how to live”. One participant states that it is a general principle that a good

goal does not justify a wrong means. And three participants defend the deontological claims that one should not lie or that it is always better to tell the truth (“*The truth is always superior.*”, participant P11).

VII. DISCUSSION

A. Discussion of Results

In line with our expectations, Immanuel receives a very high moral score after the interaction (H.1A). This indicates that robots can indeed be perceived as morally competent (even if their moral reasoning is contrary to that of the human perceiver). That Immanuel has rather low proficiency and strength ratings can be explained by the fact that Immanuel initiates the dialogue by asking the participant for advice thereby signaling its uncertainty about the moral case. Interestingly, regarding the participants, a deontological tendency to moral judgment seems to come with more uncertainty about one's own moral judgments (H.3B). A nice result is that participants with deontological tendency self-report to be more certain about their own moral standpoint after the interaction with the robot than they were at the beginning of the interaction. Thus, Immanuel may for some people serve as a tool to reflect one's own moral standpoint. The confirmation of hypothesis H.1B can be explained by the fact that the utilitarian argument is more about calculating the relative weights of pros and cons, whereas the deontological argument involves more bold concepts like intrinsic moral values and intentions. The rejection of hypothesis H.2 could be ascribed to the differences of the dilemmas in the pretest and the lying dilemma used for the conversation. However, as the qualitative analysis reveals, some of the participants who defended the wrongness of lying did that for utilitarian reasons, viz., they claimed that there must be a better action available, or they claimed that some other bad consequence should trump the good consequence of lying.

B. Limitations

From the technical perspective, the qualitative results reveal that the HERA model of the Lying Dilemma from section II-C has a limited capability to explain the whole range of arguments provided by the participants. The current model can explain the argument that lying is bad, because it is intrinsically bad, it can explain the argument that lying is permissible, because lying has no bad consequences, and it can explain that lying is permissible, because lying has a good consequence. However, participants mentioned further aspects not explicitly represented. To enable the model to explain that lying is bad, because it breaks the relationship between the robot and Mr. Smith, the model could be extended by another negative consequence representing the broken relationship. To model the frequently defended argument that there should be a better action available to motivate Mr. Smith, the model could be extended by additional actions that also have Mr. Smith being motivated to exercise among their consequences. Finally, the autonomy argument claims that everyone capable of volitional decision making should have the right to make free decision for their own life. This

argument could be approximated by adding the violation of this right as another bad consequence of lying.

Regarding the experiment, the fact that some participants that argued against lying did provide utilitarian instead of deontological arguments yields that the results concerning hypotheses that refer to the utilitarian-deontological distinction (especially H.2) should be treated with care and revisited in future work. Another limitation is that we had no condition where Immanuel does agree with the participant. Therefore, we cannot claim any effect of the contrarian robot behavior on our results. While it certainly is a result that the contrarian Immanuel is perceived as morally competent, future experiments should show how a robot's agreeing or disagreeing affects how humans perceive the robot.

VIII. CONCLUSIONS

Immanuel is a hybrid ethical reasoning agent that can reason about moral dilemmas from various perspectives. The present study shows that a robot can be perceived as morally competent even if it defends a moral standpoint contrary to the one held by a human conversation partner. We found that a moral robot can help humans reflect upon their own ethical standpoints and to become more secure about defending them. A lesson learnt is that human moral reasoning is much more diverse than suggested by the paramount utilitarian-deontological distinction. This result yields a challenge for modeling moral reasoning, and it calls for designing future experiments on moral reasoning with more fine-grained consideration of the humans' moral reasons.

REFERENCES

- [1] C. Allen, W. Wallach, I. Smit, Why machine ethics?, *IEEE Intelligent Systems*, 21(4):12–17, 2006.
- [2] Arnold, T., Kasenberg, D., Scheutz, M., Value Alignment or Misalignment – What Will Keep Systems Accountable?, *AAAI Workshop on AI, Ethics, and Society*, 2017.
- [3] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, C. Cusimano, Sacrifice one for the good of many?: People apply different moral norms to human and robot agents, In *Proc. of HRI'15*, pp. 117–124, 2015.
- [4] B. F. Malle, M. Scheutz, When will people regard robots as morally competent social partners?, In *Proc. of RO-MAN'15*, pp. 486–491, 2015.
- [5] J.-F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles, *Science*, 352(6293), 1573–1576, 2016.
- [6] F. Lindner, C. Eschenbach, Towards a formalization of social spaces for socially aware robots, In *Proc. of COSIT'11*, pp. 283–303, 2011.
- [7] F. Lindner, M. M. Bentzen, The hybrid ethical reasoning agent IMMANUEL, In the *Companion Proc. of HRI'17*, pp. 187–188, 2017.
- [8] P. Pettit, *Consequentialism*, In *A Companion to Ethics*, Oxford, Blackwell Publishers, pp. 230–237, 1991.
- [9] J. M. Paxton, J. D. Greene, Moral reasoning: Hints and allegations, *Topics in Cognitive Science*, 2(3):511–527, 2010.
- [10] D. M. Bartels, D. A. Pizarro, The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas, *Cognition*, 121:154–161, 2011.
- [11] E. Gleichgerrcht, L. Young, Low levels of empathic concern predict utilitarian moral judgment. *PLoS one*, 8(4):e60418, 2013.
- [12] C. C. Peterson, J. L. Peterson, D. Seeto, Developmental changes in ideas about lying, *Child Development*, 54(6):1529–1535, 1983.
- [13] M. M. Bentzen, The principle of double effect applied to ethical dilemmas of social robots, In *What Social Robots Can and Should Do*, pp. 268–279, IOS Press, 2016.
- [14] R. E. Petty, P. Brinol, Attitude change, *Advanced social psychology: The state of the science*, 217–259, 2010.
- [15] J. W. Brehm, *A theory of psychological reactance*, New York: Academic Press, 1966.