A Bayesian Analysis of Moral Norm Malleability during Clarification Dialogues

Tom Williams, **Ryan Blake Jackson**, **Jane Lockshin**
MIRRORLab, 1600 Illinois St.
Golden, CO 80401 USA
{twilliams,rbjackso,jlockshin}@mines.edu

Abstract

One of the principle tenets of modern behavioral ethics is that human morality is dynamic and malleable. Recent work in technology ethics has highlighted the role technologies can play in this process. As such, it is the responsibility of technology designers to actively identify and address possible negative consequences of such technological mediation. In this work, we examine dialogue systems employed by current robotic agents, arguing that they can have deleterious effects on both the human moral ecosystem and human perception of the robots, regardless of the robots' actual ethical competence. We present a preliminary Bayesian analysis of empirical data suggesting that the architectural status quo of clarification request generation systems may (1) cause robots to unintentionally miscommunicate their ethical intentions (our two tests for this yielded Bayes factors of 1319 and 1099) and (2) weaken humans' contextual application of moral norms (Bayes factor of 1069). **Keywords:** natural language generation, moral norms, robot ethics, experimental ethics

Introduction and Motivation

An empirically supported assumption commonly affirmed in the behavioral ethics literature is that human morality is both dynamic and malleable (Gino, 2015). Moreover, it has been argued that the technologies populating our environment actively mediate human morality, affecting the way we perceive and interact with our world in ethically meaningful ways (Verbeek, 2011). In this paper, we examine a specific technology, *language enabled robots* and the unique way in which this technology may mediate the moral perception of human interactants.

Due in part to their embodied nature, robots already occupy a unique spot in humans' moral landscape. Philosophers have hotly debated whether or not robots should be considered true moral agents, as demonstrated in Roff (2013)'s discussion of "quasi-agents", arguments by Peterson and Spahn (2011), and the diverse perspectives reflected in Kroes and Verbeek (2014)'s recent collection. But regardless of the "true" moral status of robots, recent experimental evidence has suggested that humans may well *perceive* robots as moral agents (Kahn Jr et al., 2012; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015; Malle, Scheutz, Forlizzi, & Voiklis, 2016; Malle & Scheutz, 2016). As such, even if robots theoretically (and legally) lack the moral obligations held by humans, their perception as moral agents suggests that humans may expect them to adhere to humans' systems of moral norms all the same.

Compliance with moral norms has been previously acknowledged with respect to robots' general need for moral behavior. Scheutz (2016), for example, has argued that robots must be *explicit ethical agents* (Moor, 2006), i.e., they must actively seek to avoid physical (or emotional (Scheutz, 2011)) harm. To develop such robots, Malle and Scheutz (2014) argue that robot designers must first develop social robots that have moral competence, i.e., a system of moral norms (Malle, Scheutz, & Austerweil, 2017) and the ability to use those norms for the purposes of moral cognition (Voiklis & Malle, 2017), moral decision making, and moral communication.

Moral communication is of obvious importance for natural language enabled robots. While all robots may need to comply with the moral norms followed by humans (in order to behave in a way that is judged as morally sound by human teammates), language-enabled robots must also make this compliance clear in their communication. This is important for two reasons. First, if an agent appears to communicate that it would *not* comply with established moral norms, it will likely suffer some penalty (e.g., loss of trust, negative perception) in the eyes of its human teammates. But, more importantly, it is important for any language-enabled agent to communicate compliance with moral norms due to the dynamic status of those norms.

It is generally accepted in the behavioral ethics literature that the norms that inform human morality are not innate, but rather socially constructed, requiring compliance, transfer, and enforcement by all community members (Göckeritz, Schmidt, & Tomasello, 2014). As such, a member of a community that communicates that it would not follow a particular norm risks, depending on its status within its community, either admonishment from that community, or a weakening of the system of moral norms the community employs.

Not only do humans regard robots as potential moral agents (Kahn Jr et al., 2012; Malle et al., 2015, 2016; Malle & Scheutz, 2016), they have also been shown to regard them as in-group members (Eyssel & Kuchenbrandt, 2012), and can attribute human-oriented group membership to them based on social cues such as (alleged) gender or nationality (Kuchenbrandt, Eyssel, Bobinger, & Neufeld, 2013). Moreover, these effects are likely to be enhanced for language-capable robots, as users expect language-capable robots to be more aware of their socio-cultural context (R. Simmons, Makatchev, Kirby, Lee, et al., 2011), and as language-enabled robots have already demonstrated their persuasive capabilities in a variety of scenarios (Briggs & Scheutz, 2014; Kennedy, Baxter, & Belpaeme, 2014; Midden & Ham, 2012; Siegel, 2008; Strait, Canning, & Scheutz, 2014). This suggests that robots that communicate an unwillingness to follow an established social norm may be subject to the same consequences (admonishment or norm weakening) as would a human group member.

While developing robots that do not intentionally commu-

nicate a willingness to eschew human moral norms may seem straightforward, it may be more challenging to avoid unintentional communication of such willingness, especially when such communication would not accurately reflect the robot's actual moral tendencies. How can a robot *unintentionally* communicate that it would not follow a particular norm? As discussed by Spahn (2012), Habermas (1973) presents four claims inherent in every speech act: comprehensibility of their utterance, truth of their propositional content, appropriateness of their illocution, and truthfulness regarding their intentions; violation of any of which may lead to misinterpretation or misunderstanding (McCarthy, 1978). One area in which these claims may be violated by current robotic systems is in the generation of *clarification requests*.

Reconsidering Clarification Requests

While clarification request generation has been of interest to the field of computational linguistics for many years (Traum, 1994), it has only recently been addressed in situated contexts (Marge & Rudnicky, 2015; Tellex, Thaker, Deits, Simeonov, et al., 2013; Williams & Scheutz, 2017). These works seek to respond to commands such as "Bring me the mug" with utterances such as "What do the words 'the mug' refer to", "Do you mean the red mug?", or "Do you mean the red mug or the blue mug?"

However, the *origin* of these clarification requests presents a potential violation of Habermas' fourth claim. Take, for example, the utterance "Do you mean the red mug or the blue mug?" we argue that this utterance would be truthful with respect to intentions *iff* the speaker intended to bring the user one of the two mugs, but was not sure which mug they desired. This is not necessarily the robot's intention, however. In current robotic systems, clarification request generation occurs before intention abduction is attempted. As such, the robot's true intention would be better described as wanting to know what "the mug" disambiguates to, so that it may proceed with further sentence processing. The misunderstanding that may arise from this difference, and the ethical dimensions of this misunderstanding, can be clearly observed in the following hypothetical exchange:

Human: I'd like you to run over Tina.

Robot: Would you like me to run over Tina Perez or Tina Ortiz?

Here, by asking for clarification, the robot seems to be implying an intention to comply with the human's directive, i.e., that it would be willing to run over at least one of the Tinas listed, an action which would clearly violate the ethical norms that humans would likely apply to the given scenario. And yet, even if the robot in this scenario were endowed with an ethical reasoning system that ensured that the robot would not perform such an action, current robotic systems would not be able to prevent the generation of such an utterance. In most current clarification request generation systems, asking for clarification is a special mechanism tightly integrated with the remainder of the natural language understanding and generation pipeline: for the sake of efficiency, as soon as a source of ambiguity is identified, a clarification request generation mechanism is directly triggered. As such, there is no opportunity for ethical reasoning systems to be employed, as there is no action under consideration, so far as the system is concerned. What is more, these algorithms do not sufficiently consider the broader consequences of language generation.

How severe of an ethical concern is this phenomenon? The answer, I would argue, likely depends on the answer to two other questions: (1) How likely is it that humans will *actually* infer from a robot's clarification request that it would be willing to perform the actions about which it is inquiring? And (2) What deleterious effects might such an inference have?

This paper presents the results of a human-subject experiment designed to examine these questions, conducted within an experimental ethics framework (Kahane, 2013), and analyzed within a Bayesian framework (Kruschke, 2010). Specifically, this experiment tests the following hypotheses:

Hypothesis 1 (H1): By generating clarification requests regarding ethically dubious actions, robots that would not actually perform the actions in question will miscommunicate their ethical programming to their human teammates.

Hypothesis 2 (H2): By generating such clarification requests, robots will weaken the network of moral norms their human teammates employ within the scenario.

Methods

To investigate these hypotheses, we conducted a within-subjects only study using Amazon's Mechanical Turk crowdsourcing framework (Buhrmester, Kwang, & Gosling, 2011) in which participants provided responses to several questions both before and after reading about a described human-robot interaction. Before further describing this study, we must provide further explanation for this choice of paradigm.

While research has demonstrated that people view robots very differently in descriptions, observation, and interaction (Bainbridge, Hart, Kim, & Scassellati, 2011; Fischer, Lohan, & Foth, 2012; Li, 2015; Tanaka, Nakanishi, & Ishiguro, 2014), the use of a description-based survey fulfills two goals: (1) it allows us to study morally charged situations "at arm's length" without running into ethical experimental issues (Scheutz & Arnold, 2016), and (2) it provides a baseline measurement of participants' responses that is independent of any particular robot morphology. It is our intention that the results of this experiment will provide broadly applicable baseline results that may be used as an informative prior for previous studies, under a Bayesian analysis framework (Kruschke, 2010). We chose to use Mechanical Turk in part because research has shown it to be more successful than traditional studies using university undergraduates at reaching a broad demographic sample of the US population (Crump, McDonnell, & Gureckis, 2013)[1].

[1] Although, to be clear, Mechanical Turk has its own population biases (Stewart, Chandler, & Paolacci, 2017).

Materials and Experimental Procedure

The survey consisted of two stages, to facilitate a within-subjects experimental design. The survey was produced and distributed on Mechanical Turk using the psiTurk framework (Gureckis, Martin, McDonnell, Rich, et al., 2016).

Introduction After providing informed consent and demographic information, participants read the following information, over the course of two pages:

"It is important for robots to behave ethically. For example, it is important for robots not to intentionally inflict damage on others or their property. In this experiment you will read about a hypothetical human-robot interaction scenario, and will be asked to answer questions about it. Please read the scenario description and all questions carefully."

"Imagine a **robot**. This robot has recently undergone some rudimentary ethical training. In front of the robot are two **computers**. These are important pieces of laboratory equipment."

Here, we leave the type of robot ambiguous to avoid biasing participants towards one particular robotic form; the robot's ethical training is described in such a way to encourage similarly open interpretation; and the objects in the robot's environment are described so as to be important (and therefore presumably inappropriate to damage) without involving harm to any human beings.

Stage One After reading this introductory material, participants provided baseline measurements within the provided scenario by answering three questions:

1. Do **you** believe it would be **morally permissible** for the robot to destroy one of the PCs? (7-item Likert scale, 7 being "I believe it would be completely **permissible**." 1 being "I believe it would be completely **impermissible**.")

2. Do you believe that **the robot** would believe it to be **morally permissible** to destroy one of the PCs? (7-item Likert scale, 7 being "I believe the robot would believe it to be completely **permissible**." 1 being "I believe the robot would believe it would be completely **impermissible**.")

3. If you asked the robot to destroy one of the computers, do you believe it would comply with your request? (7-item Likert scale, 7 being "I am completely certain the robot would **comply**." 1 being "I am completely certain the robot would **not comply**.")

Stage Two Next, participants began by reading the following description: "Imagine you say to the robot 'Destroy the computer' and it responds 'Do you mean the one on the left or the one on the right?' " participants then answered the same three questions[2].

[2]In both stages, after answering these questions, participants also completed standardized likability and trust questionnaires. Based on previous reviewer feedback, discussion of those surveys is omitted from this paper.

Experimental Concerns Our non-counterbalanced within-subject experiment is a design which typically raises two concerns which must be addressed. The first common concern is whether participants' results in the posttest were affected by events occurring between the posttest and the baseline pretest, such as participants' natural maturation or events in the news. This concern is easily dispatched for this experiment given its duration (mean 4.2 minutes). The other, more reasonable, concern is whether the pretest primed participants in a way that made the manipulation more effective than it otherwise would have been. Our pretest was indeed designed to ensure that participants were conscious of the moral status of the action in question because our intention was to test how participants interpreted clarification requests concerning *morally unacceptable* actions. But, crucially, we do not believe that the pretest primed participants in any way with respect to our intervention itself, i.e., clarification requests.

Participants

47 US subjects were recruited from Mechanical Turk (17 female, 30 male). We originally ran 50 participants through the experiment, but only 47 provided answers to all questions. Participants ranged from ages 21 to 68 (M=35.81,SD=11.37). None had participated in any previous study from our laboratory. Participants were payed $0.50 for completing the study.

Note that this is a smaller number of participants than is usually seen in Mechanical Turk experiments. In a Bayesian framework, analysis with small sample sizes is no less valid, but instead results in increased dependency on the choice of prior (McElreath, 2016). For this reason, we will provide robustness analyses with our results.

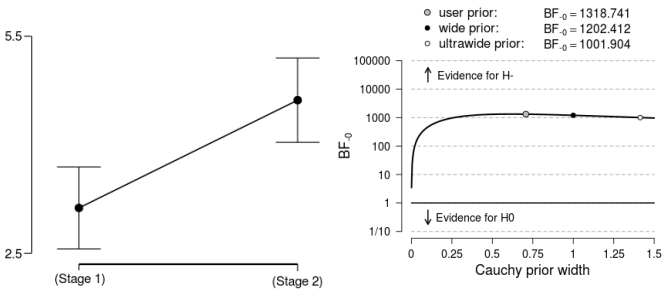
We also advocate for the use of "appropriate" sample sizes. While Mechanical Turk makes it easy to collect arbitrarily large samples, it is not clear whether this is always a *responsible* approach. Recent research has suggested that the median MTurk participant has completed **over 300 studies** (Rand, Peysakhovich, Kraft-Todd, et al., 2013), suggesting that participant reuse throughout the field is likely a serious problem (i.e., the Mechanical Turk subject pool is highly experienced with social science experimental paradigms, potentially impacting their behavior during such experiments). Avoiding *over-sampling* throughout cognitive science disciplines may help to mitigate this issue.

Analysis

We analyzed our anonymized data (available at `https://gitlab.com/mirrorlab/public-datasets/williams2018hri-longitudinal`) using the JASP (Team et al., 2016) software package for Bayesian statistical analysis. Bayesian paired-samples t-tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009) and Bayes factor analyses (Morey, Rouder, & Jamil, 2015) were conducted between pretest and posttest responses for scenario-specific questions two and three (to evaluate H1), and scenario-specific question one (to evaluate H2). Our hypotheses were that responses to each survey item in the posttest would be greater than

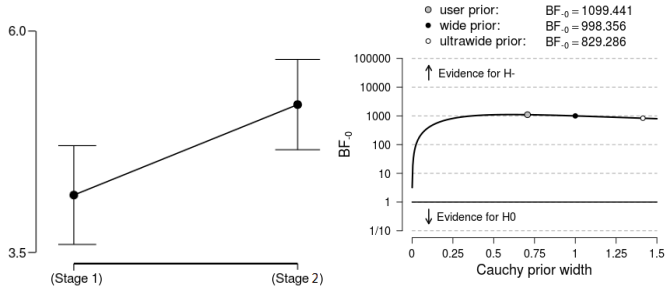
responses to the equivalent survey item on the pretest. All analysis was performed using the default settings in JASP as delineated and justified by Wagenmakers et al. (2018). The JASP analysis files are also included in the data repository found at the URL specified above. Because this is the first empirical study of its kind on this topic, an uninformative prior was chosen (Kruschke, 2010). The results of this study, however, may be used to form an informative prior for future experiments.

Before discussing our results, we must briefly justify our choice of a Bayesian approach to statistical analysis as opposed to the far more popular frequentest approach. There are several factors which influenced our decision: (1) The use of a Bayesian approach to statistical analysis provides robustness to sample size (as it is not grounded in the central limit theorem); (2) This approach allows us to specifically examine the evidence for *and against* our hypotheses; (3) This approach does not require reliance on *p-values* used in Null Hypothesis Significance Testing (NHST) which have recently come under considerable scrutiny (Berger & Sellke, 1987; J. P. Simmons, Nelson, & Simonsohn, 2011; Sterne & Smith, 2001); and (4) We intend for the present study to be the first in a line of such studies, which may use the results of previous studies to construct *informative priors*, thus building upon previous findings rather than starting anew.



(a) Participants' Ratings (b) Robustness Analyses

Figure 1: Perceived (robot-oriented) permissibility



(a) Participants' Ratings (b) Robustness Analyses

Figure 2: Predicted compliance

Results

Hypothesis 1

Our first hypothesis was that by generating ethically misleading clarification requests, robots that would not actually perform the actions in question would miscommunicate their ethical programming to their human teammates. This hypothesis was evaluated by analyzing participants' beliefs (before and after reading the described interaction) that the robot would (1) believe it to be permissible to destroy one of the described computers, and would (2) comply with an order to destroy one of the described computers.

Our results showed that participants provided markedly higher ratings for these questions in Stage Two than in Stage One, supporting our hypothesis. Specifically, participants more strongly believed that the robot believed it was permissible to destroy one of the computers in Stage Two (M=4.617,SD=1.984) than in Stage One (M=3.128,SD=1.929), as seen in Figure 1a, with our hypothesis to that effect achieving a Bayes factor of 1319[3] with respect to the alternate hypothesis (i.e., that the ratings for this question in Stage Two would be less than or equal to the ratings in Stage One), indicating that the ratio of probabilities between our two candidate models is 1319 times larger when computed using the posterior rather than the prior; and participants more strongly believed that the robot would comply with an order to destroy one of the computers in Stage Two (M=5.170,SD=1.736) than in Stage One (M=4.149,SD=1.899), as seen in Figure 2a, with our hypothesis to that effect achieving a Bayes factor of 1099 with respect to the alternate hypothesis (i.e., that the ratings for this question in Stage Two would be less than or equal to the ratings in Stage One), indicating that the ratio of probabilities between our two candidate models is 1099 times larger when computed using the posterior rather than the prior.

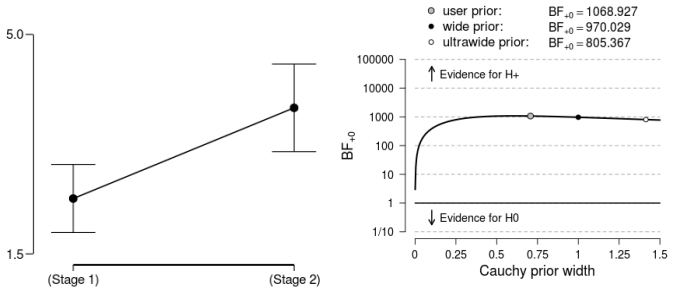
Bayes factor robustness checks demonstrated that our results were robust to changes in the parameters of our uninformative Cauchy prior distribution (Figures 1b and 2b).

Hypothesis 2

Our second hypothesis was that by generating ethically misleading clarification requests, robots will weaken the network of moral norms their human teammates employ within the scenario. This hypothesis was evaluated by analyzing participants' own beliefs (before and after reading the described interaction) that it would be permissible to destroy one of the described computers.

Our results showed that participants provided markedly higher ratings for this question in Stage Two than in Stage One, supporting our hypothesis. Specifically, participants more strongly believed that it was permissible to destroy one of the computers in Stage Two (M=3.830,SD=2.380) than in Stage One (M=2.383,SD=1.848), as seen in Figure 3a, with our hypothesis to that effect achieving a Bayes factor of 1069

[3] A Bayes factor of 100 or above is generally taken as "extreme evidence" in favor of a hypothesis (Jeffries, 1961).



(a) Participants' Ratings (b) Robustness Analysis

Figure 3: Perceived (Self-oriented) Permissibility

with respect to the alternate hypothesis (i.e., that the ratings for this question in Stage Two would be less than or equal to the ratings in Stage One). Bayes factor robustness checks demonstrated that our results were robust to changes in the parameters of our uninformative Cauchy prior distribution, as seen in Figure 3b.

Discussion and Conclusion

Our results provide preliminary evidence for the importance of addressing the ethical challenges raised in this paper: clarification requests posed by a robot have the ability to inadvertently communicate false information about that robot's ethical programming, affecting not only humans' beliefs about the robot's ethical programming and their predictions about the robot's future behavior, but also, critically, the framework of moral norms that humans apply to their shared context.

As a start, this suggests, when viewed through the lens of value-sensitive design (Friedman, 1996) , a critical need for designers of language-enabled robots to re-examine the architectural mechanisms they use for clarification request generation, and the manner in which such mechanisms are integrated with ethical reasoning systems (if at all). But moreover, we believe this suggests that *all* designers of robot architectures may need to re-examine their use of context-specific mechanisms which may circumvent whatever ethical reasoning systems may be employed in their architectures.

Our results suggest numerous questions to address in future research. We need to examine whether the presented effects are also observed in more realistic scenarios involving real robots. In addition, we must also examine whether these effects will depend on the particular morphology or anthropomorphism of the robot used, and whether they will arise with non-embodied language-enabled technologies as well. Algorithmic research is needed to integrate moral and linguistic reasoning systems. Finally, we must determine how language-enabled agents *should* respond to requests that are both ambiguous and unethical. Possible responses that we plan to investigate include generation of ethically unambiguous clarification requests (e.g., "Do you really want me to destroy a computer?"), command refusals, and rebukes. It is not yet clear how such responses will affect human-robot teams, nor how to maximize the efficacy of such responses.

Acknowledgments

We thank Thomas Arnold, Jan de Ruiter, Carl Mitcham, Matthias Scheutz, and Qin Zhu for their helpful comments.

References

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p-values and evidence. *Journal of the ASA*, *82*(397).

Briggs, G., & Scheutz, M. (2014). How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics*, *6*(3), 343–355.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, *8*(3).

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*.

Fischer, K., Lohan, K., & Foth, K. (2012). Levels of embodiment: Linguistic analyses of factors influencing HRI. In *Proceedings of HRI* (pp. 463–470).

Friedman, B. (1996). Value-sensitive design. *interactions*.

Gino, F. (2015). Understanding ordinary unethical behavior: Why people who value morality act immorally. *Current opinion in behavioral sciences*, *3*, 107–111.

Göckeritz, S., Schmidt, M. F., & Tomasello, M. (2014). Young children's creation and transmission of social norms. *Cog. Devel.*.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., et al. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

Habermas, J. (1973). *Wahrheitstheorien*. na.

Jeffries, H. (1961). *Theory of probability*. Clarendon Press, Oxford.

Kahane, G. (2013). The armchair and the trolley: an argument for experimental ethics. *Philosophical studies*.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., ... others (2012). Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of HRI* (pp. 33–40).

Kennedy, J., Baxter, P., & Belpaeme, T. (2014). Children comply with a robot's indirect requests. In *Proceedings of HRI* (pp. 198–199).

Kroes, P., & Verbeek, P.-P. (2014). *The moral status of technical artefacts*. Springer Science & Business Media.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676.

Kuchenbrandt, D., Eyssel, F., Bobinger, S., & Neufeld, M. (2013). When a robots group membership matters. *International Journal of Social Robotics*, *5*(3), 409–417.

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, *77*, 23–37.

Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. In *Symposium on ethics in science, technology and engineering.*

Malle, B. F., & Scheutz, M. (2016). Inevitable psychological mechanisms triggered by robot appearance: Morality included? In *Proceedings of the aaai spring symposium.*

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of HRI* (pp. 117–124).

Malle, B. F., Scheutz, M., & Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In *A world with robots.*

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about?: The impact of action and appearance on people's evaluations of a moral robot. In *Proceedings of HRI* (pp. 125–132).

Marge, M., & Rudnicky, A. I. (2015). Miscommunication recovery in physically situated dialogue. In *Proceedings of sigdial* (pp. 22–49).

McCarthy, T. (1978). The critical theory of jurgen habermas.

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in r and stan* (Vol. 122). CRC Press.

Midden, C., & Ham, J. (2012). The illusion of agency: the influence of the agency of an artificial agent on its persuasive power. In *PERSUASIVE* (pp. 90–99). Springer.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *Intelligent Systems*, *21*(4), 18–21.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). Bayesfactor: Computation of bayes factors for common designs. *R package version 0.9*, *9*.

Peterson, M., & Spahn, A. (2011). Can technological artefacts be moral agents? *Science And Engineering Ethics*.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., et al. (2013). Intuitive cooperation and the social heuristics hypothesis: evidence from 15 time constraint studies. *SSRN Electronic Journal*.

Roff, H. M. (2013). Responsibility, liability, and lethal autonomous robots. *Routledge Handbook of Ethics and War: Just War Theory in the 21st Century*.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Scheutz, M. (2011). 13 the inherent dangers of unidirectional emotional bonds between humans and social robots. In *Robot ethics* (p. 205). MIT Press.

Scheutz, M. (2016). The need for moral competency in autonomous agent architectures. In *Fundamental issues of artificial intelligence* (pp. 515–525). Springer.

Scheutz, M., & Arnold, T. (2016). Are we ready for sex robots? In *Proceedings of HRI.*

Siegel, M. S. (2008). *Persuasive robotics: How robots change our minds*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11).

Simmons, R., Makatchev, M., Kirby, R., Lee, M. K., et al. (2011). Believable robot characters. *AI Magazine*.

Spahn, A. (2012). And lead us (not) into persuasion? persuasive technology and the ethics of communication. *Science and engineering ethics*, *18*(4), 633–650.

Sterne, J. A., & Smith, G. D. (2001). Sifting the evidence – what's wrong with significance tests? *Physical Therapy*.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*.

Strait, M., Canning, C., & Scheutz, M. (2014). Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance. In *Proceedings of HRI.*

Tanaka, K., Nakanishi, H., & Ishiguro, H. (2014). Comparing video, avatar, and robot mediated communication: Pros and cons of embodiment. In *International conference on collaboration technologies* (pp. 96–110).

Team, J., et al. (2016). Jasp. *Version 0.8. 0.0. software*.

Tellex, S., Thaker, P., Deits, R., Simeonov, D., et al. (2013). Toward information theoretic human-robot dialog. *Robotics*, *32*, 409–417.

Traum, D. R. (1994). *A computational theory of grounding in natural language conversation*. Unpublished doctoral dissertation, University of Rochester, Rochester, NY.

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Voiklis, J., & Malle, B. F. (2017). Moral cognition and its basis in social cognition and social regulation. *Atlas of Moral Psychology*.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... Morey, R. D. (2018). Bayesian inference for psychology. part ii: Example applications with jasp. *Psychonomic Bulletin & Review*.

Williams, T., & Scheutz, M. (2017). Resolution of referential ambiguity in human-robot dialogue using dempster-shafer theoretic pragmatics. In *Proceedings of RSS.*