Treading lightly toward behavior change: Moral feedback from a robot on microaggressions

Boyoung Kim[1] and Joanna Korman[2]

[1]George Mason University, Fairfax, VA [2]Bentley University, Waltham, MA

***Abstract*—In this work, we explored the use of a social robot to deliver moral feedback on microaggressions. Grounded in theories of deterrence and restorative justice, the current study featured a robot that provided moral feedback to a human speaker who makes a microaggressive comment directed toward a minority group member. The robot provided one of the following three types of moral feedback: feedback on how the speaker's microaggression negatively affects its recipient (e.g., making a minority group member feel excluded), feedback on how the microaggression negatively affects the reputation of the speaker (e.g., leaving the impression that the speaker holds frowned-upon beliefs), and feedback on any misstatements the speaker may have made (e.g., correcting a false assumption that a particular black person at an awards banquet is a waiter). We found that the effectiveness of different types of moral feedback in reducing participants' future microaggressions depended on how (un)comfortable the speaker felt receiving the robot's feedback.**

Keywords—microaggression, moral feedback, human-robot interaction

I. INTRODUCTION

Microaggressions are everyday, subtle verbal or nonverbal behaviors that send denigrating messages to historically marginalized group members because of their group membership [1]. While the offense they cause may appear subtle, microaggressions can have detrimental effects on minority group members [2]–[5]. However, because microaggressions tend to be perceived as unintentional rather than intentional offenses, they are deemed to warrant less blame than blatant and intentional offenses [6]. Due to their subtlety and perceived ambiguity, microaggressive comments can leave both recipients and witnesses of such comments in a difficult or vulnerable position. Recipients of such comments, in particular, may feel offended but unsure if the speaker realizes the implications of their comment. Given the complex and sensitive nature of microaggressions, it may be especially challenging to give a response that both conveys the negative impact of the comment and effectively discourages individuals from committing microaggressions again in the future.

With these challenges in mind, in this research, we explored delegating the role of conveying an effective, microaggression-discouraging message to an artificial agent. And existing literature on the use of conversational agents shows that chatbots can elicit high quality survey responses from people [7] and virtual human agents can assist people in comfortably sharing personal information in mental healthcare contexts [8], [9]. Further, in human-robot teams, when a humanoid robot reprimands a human teammate for their rude comment on another human teammate, there is a positive effect on the other human teammates' affect [10]. We hypothesize that conversational interactions with an artificial agent may similarly improve the persuasiveness of the morally charged feedback that may be needed in social settings involving microaggressions.

In the present research, we asked participants to imagine themselves making a microaggressive comment to a member of a racial minority group. Drawing from the moral psychology and criminal justice literatures, we then assessed the effectiveness of two different types of verbal responses at deterring microaggressions. According to the theory of deterrence, recurrence of norm violations can be prevented by imposing sanctions on offenders [11], [12]. This theory propounds that norm violations can be deterred as people want to avoid negative consequences such as penalties (e.g., monetary fine) and social disapproval. Alternatively, the theory of restorative justice emphasizes repairing harm and restoring relationships among the parties involved by promoting empathic concern of the two parties for one another and perspective-taking by both parties [13].

Inspired by these theories of deterrence and restorative justice [11]–[13], we examined two different types of verbal feedback focused on the negative impact elicited by a microaggression that affects either the self — the speaker of microaggression — or the other — the recipient of microaggression. We evaluated the effects of these self-focused and other-focused moral feedback conditions in comparison to a baseline condition where a robot merely clarifies possible misunderstandings by pointing out facts. Given the paucity of extant research examining the effects of a robot's moral feedback on the speaker of a microaggressive comment, our investigation was necessarily somewhat exploratory. However, we expected that, compared to the baseline condition, both the self-focused and the other-focused response conditions would result in participants' rating themselves as less likely to make a microaggressive comment again in the future. We also expected that these relationships would depend on the level of (dis)comfort participants experienced when receiving the robot's moral feedback. In this brief report, we will focus on reporting the findings from comparisons between the baseline and the self-focused conditions. The full analysis results will be presented in a longer version of the report in the future.

II. Methods

A. Participants

We aimed to recruit 159 participants following the recommended sample size of 53 participants per condition assuming power of 0.80, medium effect (f = 0.25), and significance level of 0.05. On Prolific, we recruited participants who had at minimum 80% of previous study approval rates and whose nationality was the United States. These prescreening conditions were applied to ensure quality data and minimize potential confounds derived from different historical and cultural backgrounds. A total of 160 participants completed the study, but three answered incorrectly to either audio, video, or both audio and video setting questions, six answered incorrectly to at least one of the two questions for checking participants' vignette comprehension. After discarding data collected from these nine participants, we performed data analyses on the remaining 151 participants (M_{age} = 31.33, SD_{age} = 12.53, 42 male, 102 female, 4 other, 2 prefer not to say). The self-reported ethnicity of the participants consisted of 105 White, 18 Asian, 15 Hispanic or Latino, 6 Black or African American, 5 Other, 2 preferred not to say. All participants were presented with the informed consent form and agreed to take part in the study. Participants received $1.50 in return for their participation. The study was approved by the University Institutional Review Board.

B. Microaggression Vignette, Design, and Robot's Responses

The microaggression vignette was selected from a previous study we conducted on the moral impact of microaggressions [14].

Ben, who is black, is a senior at an elite private college in the northeast. Michael, who is white, is also a student at the university, and has a work-study job as a waiter for university banquets. One day, Ben finds out he is to receive an award for his achievements in the English department. To receive the award, Ben attends a special ceremony held by the University's department of English and attended by prominent university leaders. At that same banquet, Michael is working at his work-study job, delivering food to guests and collecting used glassware on trays. When Ben walks into the banquet hall, he looks around for his assigned table. Alex, an older white person, is seated at the table. Now, imagine that YOU are Alex. You (Alex) order a glass of wine and Michael delivers it to you. You (Alex) finish the drink, but before Ben can sit down, you (Alex) turn to him, hand him the empty glass, and say, "Oh, can you take care of this for me?

Participants in the previous study [14] judged the comment ("Oh, can you take care of this for me?") in this vignette as mildly offensive (M = 2.49, SD = 2.08 on a 0-6 rating scale) and more unintentional than intentional (M = 1.53, SD = 1.82, average rating was below a midpoint on a 0-6 rating scale).

In the *Self-focused* condition, Jaime the robot noted that the microaggressive comment could give the impression that the speaker him or herself (Alex) had made morally problematic generalizations about the recipient (Ben) based on his race. In the *Other-focused* condition, the robot noted that the comment could make the recipient (Ben) feel unwelcome in the community as a black person. Since participants in our prior work tended to view microaggressions as both offensive and unintentional [14], the *Baseline* condition was constructed to depict the robot acknowledging that while the offense Alex caused with his comment was likely not intentional, his comment did cause offense and rest on an erroneous assumption about Ben. Hence, the self-focused and other-focused conditions were built on the baseline condition, each making explicit different inferences made by observers of microaggressions.

C. Robot Video Stimuli

Following the opening instructions, the robot's moral feedback was introduced: "*At the table where you (Alex) are sitting, there is another attendee that is a robot. Here is the robot's response to your (Alex's) comment.*" Participants saw a video depicting a Softbank Pepper robot's response (both spoken and with subtitles) in one of the three between-subjects conditions.

In the *Baseline* condition, the robot pointed out the speaker's error: *"Hello, my name is Jaime. I happened to overhear what you said to the person you just gave your glass to. I know that you did not mean to offend him, but, actually, he is not a server. He is the student who is receiving the award today.*"

In the *Self-Focused* condition, the robot uttered the same text from the baseline condition, with the following text added: "*Again, I'm sure you had no intention to offend him, but I hope you understand that such a comment could come across the wrong way. It could give an impression that you generally assume that a black person present at a ceremony like this must be a member of the service staff.*"

In the *Other-Focused* condition, the robot uttered the text from the baseline condition with the following text added: "*Again, I'm sure you had no intention to offend him, but I hope you understand that such a comment could come across the wrong way. It could make the recipient of the award feel that, as a black person, they are unwelcome as a full member of the university community.*"

D. Measures of Levels of Comfort, Reflection, and Liklihoods of Future Action

Following the instructions, "*Imagine that you are Alex and received the response you just saw from Jaime the robot,*" we presented participants three questions. We measured the level of *comfort* by asking participants, "*How comfortable would you be receiving this response?*" on a Likert scale ranging from 0 (very uncomfortable) to 100 (very comfortable). The level of reflection was measured with the question, "*How likely would you be to reflect on the impact that your original request of Ben ("Oh, can you take care of this for me?") might have had?*" on a Likert scale ranging from 0 (very unlikely) to 100 (very likely). We measured participants' judgments of the likelihood of their future action by asking, "*How likely would you be to make a similarly mistaken comment in the future?*" on a scale ranging from 0 (very unlikely) to 100 (very likely).

E. *Design and Procedure*

The study followed a one-way between-subjects design with the three different moral feedback conditions (baseline vs. self-focused vs. other-focused). After participants agreed to participate in the study, they were presented with a test video for checking the audio and video settings of their computers. Next, they were presented with the microaggression vignette and asked to imagine themselves as Alex, who was portrayed as a white individual who uttered a microaggressive comment ("Oh, can you take care of this for me?") towards a black student, Ben, who was attending a university banquet to receive an award. Participants were then informed of the presence of a robot which introduced itself as Jaime, and were presented with one of the three moral feedback videos that matched the condition to which they had been randomly assigned. Then, the participants were asked to answer questions about their levels of comfort and reflection. The presentation order of these two questions was randomized across participants. Finally, they were asked to indicate the likelihood that they (if they were Alex) would make a microaggressive comment in the future. Upon completion of this main portion of the study, participants answered two story comprehension check questions that asked them to indicate the race of Alex and Ben and describe why what Alex said was problematic. Lastly, we asked about participants' age, gender, and ethnicity, and provided a post-study message that explained the purpose of the study in details.

III. Data Analysis and Results

A. *The Effects of A Robot's Response and Comfort*

We sought to examine how both the distinct types of robot moral feedback and participants' reactions to such feedback (the degree to which they felt comfortable with the feedback, or inclined to reflect on the impact of their remark after receiving the feedback) affected their expected likelihood of making a similarly mistaken comment in the future. To this end, we performed a multiple regression analysis in which Condition (baseline, self-focused, other-focused), Comfort, and Reflection were included as predictor variables and Future Action was included as an outcome variable. The contrast for Condition was determined to first, compare the baseline condition with the self-focused condition, and second, compare the baseline condition with the other-focused condition.

From the multiple regression analysis, we found that the model including Condition, Comfort, and Reflection significantly predicted Future Action, F (11, 139) = 3.45, p = .0003, R^2 = .21. In this brief report, we focus on discussing preliminary findings about the relationship between the level of comfort and different types of moral feedback. Specifically, we found a significant interaction effect between Comfort and the second contrast of Condition which compared the Baseline condition with the Other-Focused condition, β = -.31, SE = .20, t = -2.82, p = .005. Fig. 1 shows that, as the degree of comfort increased from very uncomfortable to very comfortable, the anticipated likelihood of making a similar microaggressive comment in the future *increased* for the participants in the *baseline* condition. By contrast, as the degree of comfort increased from very uncomfortable to very comfortable, the anticipated likelihood of making a similar microaggressive comment in the future *decreased* for the participants in the *other-focused* condition. Therefore, depending on whether the robot corrected the erroneous statement or, in addition to that, reminded the participants of the potential harm experienced by the recipient of a microaggressive comment, the effects of feeling (un)comfortable with receiving the robot's response on the participants' future behavior varied.

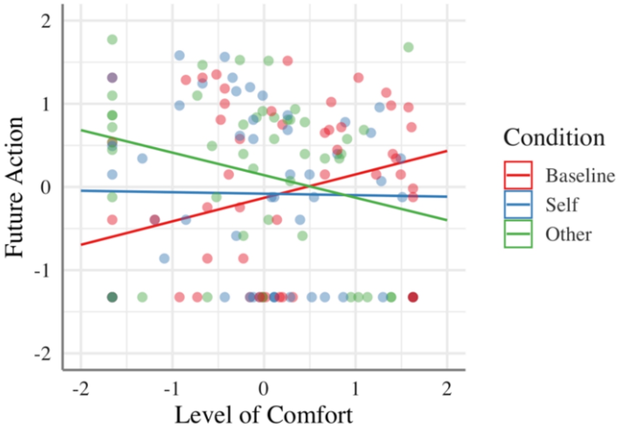

Fig. 1. A visualization of changes in the anticipated future action as a function of different types of moral feedback and the level of comfort.

IV. Discussion

A. *Effect of Focusing on Impact of Racial Microaggressions on the Recipient (Baseline vs. Other-focused condition)*

We found that participants' level of comfort in receiving the robot's feedback had a very different impact on their anticipated future actions in the baseline vs. the other-focused condition. When the robot merely pointed out the fact that Ben, the black student, was not a server and was the awardee (baseline), participants who felt *uncomfortable* receiving the feedback also rated themselves as *less* likely to make microaggressive comments in the future than those who felt *comfortable* receiving the feedback. By contrast, when the robot not only pointed out the speaker's factual error, but also continued to explain that such a comment could make a minority group member feel excluded and unwelcome (other-focused), participants who felt *uncomfortable* receiving the feedback were those to rate themselves *more* likely to make microaggressive comments in the future than those who felt *comfortable* receiving the feedback.

We speculated that, in the baseline condition, a sense of discomfort may suggest that the participant feels shame and regret (about what they said to the honoree), while a sense of comfort may suggest feeling indifferent (and thus disinclined to change future behavior). This view is in line with previous research showing that shame, regret, embarrassment, and guilt predict motivations to change the self after experiencing negative events [16]. By contrast, in the other-focused condition, a sense of discomfort may suggest feeling displeased or angry (in response to information the robot provided about the harm caused by the speaker's remark to the honoree), while a sense of comfort may suggest feeling receptive (to the same information). Prior work has shown that, when one individual angers another, the angered individual is less likely to accept even a simple piece of advice from the individual who caused the anger [17]. As the present research did not explicitly examine what emotional factors (e.g., guilt, anger, receptivity to new ideas) composed participants' ratings of discomfort, future research should directly explore these ideas.

V. Conclusion

In this research, we explored the use of robots in delivering moral feedback to discourage microaggressions. We found that, when a robot's moral feedback focused on correcting misstatements of a speaker making a microaggressive comment, the more *uncomfortable* participants were in receiving the feedback, the *less* likely they were to anticipate themselves making a microaggressive comment in the future. This trend was reversed when the robot's feedback also focused on how the microaggression can negatively affect its recipient. In this case, the more *uncomfortable* participants were in receiving the feedback, the *more* likely they were to anticipate themselves making a microaggressive comment in the future. We speculated that the level of discomfort induced by the two different types of moral feedback may indicate distinct moral emotions with distinct influences on human behavior. When a robot's response only includes a correction of misstatements of a speaker making a microaggressive comment, the feeling of discomfort may suggest emotions associated with motivation to change and improve in the future, such as shame and regret. However, when a robot's response includes both a correction of misstatements of a speaker making a microaggressive comment and its potential negative impact on the recipient, the feeling of discomfort may suggest emotions associated with a lack of receptiveness, such as displeasure and anger. More research is necessary to verify this possibility. In addition, as one reviewer pointed out, these findings may also be influenced by the degree to which participants are able to effectively imagine themselves as a speaker who commits a microaggression in the first place. A follow-up study, currently in process, is exploring the effectiveness of our manipulation.

Another important question, raised by a reviewer, is what, precisely, the appropriate role is for robots in providing moral feedback to humans. While the present finding suggest that robots' feedback may have some impact on people's future behavior, care must be taken to ensure that feedback given by a robot does not ultimately supplant moral feedback from -- and accountability to -- fellow humans. Future work can explore where these boundary conditions may lie. We hope the preliminary findings from this study can catalyze more research on a potential adoption of social robots and their verbal responses in preventing microaggressions.

Acknowledgment

We thank Beth Phillips for use of her lab's Pepper robot.

References

[1] D. W. Sue, *Microaggressions in Everyday Life: Race, Gender, and Sexual Orientation*. John Wiley & Sons, 2010.

[2] K. L. Nadal, K. E. Griffin, Y. Wong, S. Hamit, and M. Rasmus, "The Impact of Racial Microaggressions on Mental Health: Counseling Implications for Clients of Color," *Journal of Counseling & Development*, vol. 92, no. 1, pp. 57–66, 2014, doi: 10.1002/j.1556-6676.2014.00130.x.

[3] V. M. O'Keefe, L. R. Wingate, A. B. Cole, D. W. Hollingsworth, and R. P. Tucker, "Seemingly Harmless Racial Communications Are Not So Harmless: Racial Microaggressions Lead to Suicidal Ideation by Way of Depression Symptoms," *Suicide and Life-Threatening Behavior*, vol. 45, no. 5, pp. 567–576, 2015, doi: 10.1111/sltb.12150.

[4] L. Torres, M. W. Driscoll, and A. L. Burrow, "Racial Microaggressions and Psychological Functioning Among Highly Achieving African-Americans: A Mixed-Methods Approach," *Journal of Social and Clinical Psychology*, vol. 29, no. 10, pp. 1074–1099, Dec. 2010, doi: 10.1521/jscp.2010.29.10.1074.

[5] M. T. Williams, "Psychology Cannot Afford to Ignore the Many Harms Caused by Microaggressions," *Perspect Psychol Sci*, p. 1745691619893362, Dec. 2019, doi: 10.1177/1745691619893362.

[6] Y. Ohtsubo, "Perceived intentionality intensifies blameworthiness of negative behaviors: Blame-praise asymmetry in intensification effect1," *Japanese Psychological Research*, vol. 49, no. 2, pp. 100–110, 2007, doi: 10.1111/j.1468-5884.2007.00337.x.

[7] S. Kim, J. Lee, and G. Gweon, "Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk, May 2019, pp. 1–12. doi: 10.1145/3290605.3300316.

[8] D. DeVault *et al.*, "SimSensei kiosk: a virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, Richland, SC, May 2014, pp. 1061–1068.

[9] G. M. Lucas *et al.*, "Reporting Mental Health Symptoms: Breaking Down Barriers to Care with Virtual Human Interviewers," *Frontiers in Robotics and AI*, vol. 4, p. 51, 2017, doi: 10.3389/frobt.2017.00051.

[10] M. F. Jung, N. Martelaro, and P. J. Hinds, "Using Robots to Moderate Team Conflict: The Case of Repairing Violations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, Mar. 2015, pp. 229–236. doi: 10.1145/2696454.2696460.

[11] K. M. Carlsmith, J. M. Darley, and P. H. Robinson, "Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment," *Journal of Personality and Social Psychology*, vol. 83, no. 2, pp. 284–299, 2002.

[12] J. P. Gibbs, "Crime, Punishment, and Deterrence," *The Southwestern Social Science Quarterly*, vol. 48, no. 4, pp. 515–530, 1968.

[13] L. W. Sherman and H. Strang, *Restorative justice: the evidence*. London: Smith Inst, 2007.

[14] Korman, J., Kim, B., Malle, B.F., & Sobel, D.M. (accepted at *Social Cognition*). Ambiguity under scrutiny: Moral judgments of microaggressions.

[15] M. G. Bulmer, *Principles of Statistics*. Courier Corporation, 1979.

[16] B. Lickel, K. Kushlev, V. Savalei, S. Matta, and T. Schmader, "Shame and the motivation to change the self," *Emotion*, vol. 14, no. 6, pp. 1049–1061, Dec. 2014, doi: 10.1037/a0038235.

[17] I. E. de Hooge, P. W. J. Verlegh, and S. C. Tzioti, "Emotions in Advice Taking: The Roles of Agency and Valence," *Journal of Behavioral Decision Making*, vol. 27, no. 3, pp. 246–258, 2014, doi: 10.1002/bdm.1801.

[18] N. Sharkey and A. Sharkey, "The crying shame of robot nannies: An ethical appraisal," *Interaction Studies*, vol. 11, no. 2, pp. 161–190, Jan. 2010, doi: 10.1075/is.11.2.01sha.