Vulnerable robots positively shape human conversational dynamics in a human–robot team

Margaret L. Traeger[a,b], Sarah Strohkorb Sebo[c], Malte Jung[d], Brian Scassellati[c], and Nicholas A. Christakis[a,b,e,f,1]

[a]Yale Institute for Network Science, Yale University, New Haven, CT 06520; [b]Department of Sociology, Yale University, New Haven, CT 06520; [c]Department of Computer Science, Yale University, New Haven, CT 06520; [d]Department of Information Science, Cornell University, Ithaca, NY 14853; [e]Department of Biomedical Engineering, Yale University, New Haven, CT 06520; and [f]Department of Statistics and Data Science, Yale University, New Haven, CT 06520

Edited by Cesar A. Hidalgo, Massachusetts Institute of Technology, Cambridge, MA and accepted by Editorial Board Member Mary C. Waters January 28, 2020 (received for review June 21, 2019)

Social robots are becoming increasingly influential in shaping the behavior of humans with whom they interact. Here, we examine how the actions of a social robot can influence human-to-human communication, and not just robot–human communication, using groups of three humans and one robot playing 30 rounds of a collaborative game (n = 51 groups). We find that people in groups with a robot making vulnerable statements converse substantially more with each other, distribute their conversation somewhat more equally, and perceive their groups more positively compared to control groups with a robot that either makes neutral statements or no statements at the end of each round. Shifts in robot speech have the power not only to affect how people interact with robots, but also how people interact with each other, offering the prospect for modifying social interactions via the introduction of artificial agents into hybrid systems of humans and machines.

human–robot interaction | groups and teams | conversational dynamics

Human behavior is very sensitive to the influence of other people in their groups. Verbal behavior, in particular, plays a special role in establishing successful social interactions and cooperation (1). Additionally, successful social interactions are remarkably dependent on group dynamics, especially when contrasting a one-on-one (dyadic) interaction with an interaction occurring across several members of a group (2, 3). Social influence across social ties can enable the spread of phenomena as diverse as emotions, voting, and cooperation, in both natural and laboratory settings (4–7).

Just as humans are capable of influencing the behaviors of the people with whom they interact, robots may exert similar effects. For example, robots can alter how well human participants complete a task (8–12) and how humans respond to requests (13–15) in human–robot interactions. Studying such one-on-one interactions between humans and robots is important to understanding how robots can affect direct human–robot relationships. For example, algorithms programmed with simple communication capacities, such as "cheap talk," are able to effectively improve one-on-one human–robot cooperation to a level on par with human–human cooperation in online games (16). However, analyzing interactions in groups with several humans and one robot, compared to one-on-one interactions, can illuminate other aspects of how robots influence human behavior between the humans themselves.

Research on group interactions is less common, but extant work has shed light on several novel aspects of human–robot interactions. For example, a social robot can shape the roles members of a group assume through gaze direction (17). Robots can also be programmed with algorithms that alter their behavior to emphasize different aspects of the group, such as increasing group cohesion, productivity, sense of peace, or engagement (18–21). Furthermore, in settings with many human participants, robots can guide interactions between humans by adjusting their speech, focus, and actions (9, 22–25). This effect is especially pronounced when the content of robot speech is based on group emotions, which can lead to a more positive view of the robot (26). As robots increasingly populate our homes and workplaces, the ubiquitous presence of such groups composed of humans and machines could alter how humans relate to both robots and each other (27–29).

For a group to sustain and achieve collective outcomes, it is extremely helpful if its members socially engage with one another and are willing to communicate. Groups without social interaction are less able to learn from each other and work together (30, 31). One way to establish and promote social engagement within groups is through vulnerable expressions. Vulnerability focuses individuals on others, which encourages interpersonal connection (32). In interpersonal interactions, vulnerability is often expressed as self-disclosure or personal stories (which increase solidarity) and humor (which alleviates tension) (33–36). But this can also be extended to nonhuman agents, like robots, by programming them to exhibit these behaviors (37, 38). For instance, when robots disclose emotions in a collaborative task, humans are more likely to project feelings of companionship onto the robot (37). Expressions of vulnerability by a robot have also been shown to positively influence interpersonal behaviors (e.g., laughing with

Significance

Prior work has demonstrated that a robot's social behavior has the ability to shape people's trust toward, responses to, and impressions of a robot within human–robot interactions. However, when the context changes to interactions within a group involving one robot and multiple people, the influence of the robot on group behavior is less well understood. In this work, we explore how a social robot influences team engagement using an experimental design where a group of three humans and one robot plays a collaborative game. Our analysis shows that a robot's social behavior influences the conversational dynamics between human members of the human–robot group, demonstrating the ability of a robot to significantly shape human–human interaction.

Author contributions: M.L.T., S.S.S., M.J., B.S., and N.A.C. designed research; M.L.T. and S.S.S. performed research; M.L.T. analyzed data; M.J., B.S. and N.A.C. supervised the research; B.S. and N.A.C. provided resources and funding; and M.L.T., S.S.S., M.J., B.S., and N.A.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. C.A.H. is a guest editor invited by the Editorial Board.

Data deposition: Data and reproducible R scripts are available on the Human Nature Lab website at http://humannaturelab.net/publications/vulnerable-robots.

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: nicholas.christakis@yale.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1910402117/-/DCSupplemental.

First published March 9, 2020.

team members and liking the robot more) within a human–robot group (39, 40).

While this prior work offers evidence of a robot's capacity to influence how human group members interact in the presence of a robot, this work has not examined whether these changes are sustained over time or if the influence of robot utterances extends to the dynamics of the discussion between the humans themselves. Here, we focus on the ability of a robot to improve human discussion in a collaborative task. We test whether a robot programmed to utter vulnerable statements throughout a sustained group interaction can influence the conversational dynamics between human members of the group.

In this experiment, 153 participants were assigned to 51 groups consisting of 3 human participants and 1 robot. The groups worked collaboratively, with the social robot, over 30 consecutive rounds, to play our custom-built, tablet-based game (39). This game was developed so that the human members of the group and the robot were perceived to all be part of the same team (see *SI Appendix* for more information). A limited version of this paradigm has been shown to increase engagement with robots themselves as well as how likely human participants are to explain their mistake and console their teammates in the game (39). Groups were assigned to one of three conditions: the vulnerable condition, the neutral condition, or the silent condition. In the vulnerable condition, the robot made a vulnerable comment at the end of every round; in the neutral condition the robot made a task-related comment at the end of every round (see *SI Appendix* for more information); and in the silent condition the robot did not speak at the end of every round.

Results

We used multilevel modeling to account for the clustered structure of the data (participants in groups) and, when needed, the longitudinal aspects of the data (rounds within participants within groups). Further details on the analysis can be found in *SI Appendix*.

Total Talking Time. There was a substantial difference between conditions in the total amount of time spent talking by participants, where those in the vulnerable condition spoke twice as much over the course of the game ($\overline{x_{V_i}} = 253.60$ s, SD = 184.41 s) compared to those in the neutral condition ($\overline{x_{N_i}} = 124.23$ s, SD = 78.78 s) and the silent condition ($\overline{x_{S_i}} = 119.86$ s, SD = 148.17 s). This difference was statistically significant between the vulnerable and neutral conditions (c = 140.68, P = 0.001) and the vulnerable and silent conditions (c = 124.52, P = 0.004), even after adjustment for age, extraversion, gender, and familiarity, using regression models, but there was no significant difference between the silent and neutral conditions (c = 16.15, P = 0.70) (see Fig. 1*A* and *SI Appendix*, Table S2).

In Fig. 1*B*, we show the total time participants spent talking to each of the other human participants and the robot, represented by the line width of the connections in the group network. The vulnerable robot condition enhanced interhuman conversation. In the neutral condition, across all groups, participants spoke to their human teammates for 83.22 min (4,993 s) and to the robot for 10.27 min (616 s) over the course of the game. In the vulnerable condition, participants spoke to their human teammates and the robot more than twice as much [178.38 min (10,703 s) and 24.02 min (1,441 s), respectively]. In the silent condition, participants spoke to their human teammates for similar amounts of time to the neutral condition [83.07 min (4,984 s)] but spoke to the robot very little [5.61 min (336 s)].

Additionally, those in the vulnerable condition spoke progressively more over time (across rounds in the game) ($\overline{x_{V_i}} = 8.45$ s, SD = 8.33 s) compared to those in the neutral condition ($\overline{x_{N_i}} = 4.14$ s, SD = 4.99 s) as demonstrated by the significant interaction effect between round and experimental condition with

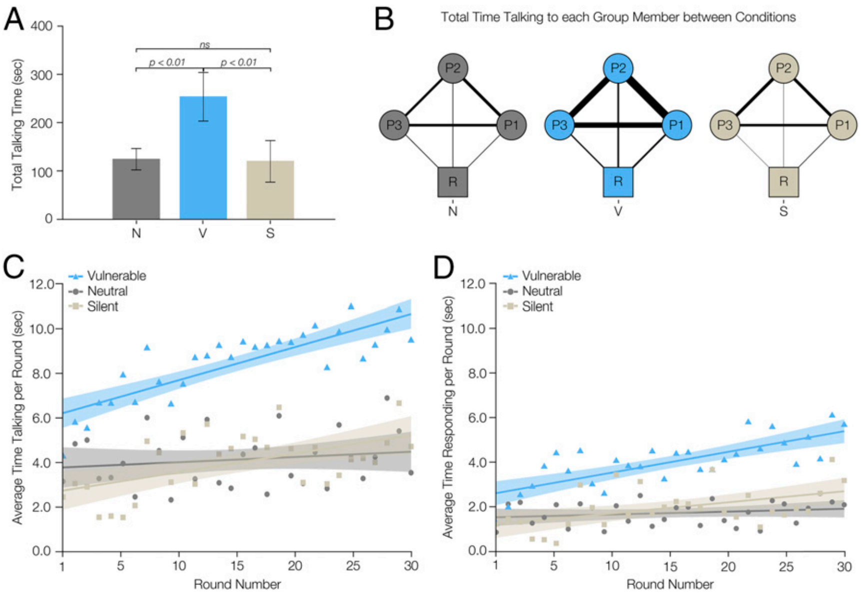

Fig. 1. Total talking time by condition. Compared to the neutral and silent conditions, human participants in the vulnerable condition spoke more, in total, to the other participants in their group, and increasingly across game rounds. In *A*, we see that participants in the vulnerable condition (V) spoke significantly more than participants in either the neutral (N) or silent (S) condition (n = 153 participants). In *B*, the line widths represent the amount of talking by human participants toward their teammates who are connected by the line, in seconds (summed across all groups within a condition (n = 153 participants)). R = robot; P1, P2, and P3 = human participants, in their relative positions around the table. In *C* and *D*, the shaded area around each line represents a 95% confidence interval (n = 4,590 rounds); the dots represent the condition average for that round. In *C*, the vulnerable condition has more talking in every round, and the slope (i.e., the rate of increase in talking per round, across rounds) is higher than the neutral condition (but indistinguishable from the silent condition). In *D* we see that, compared to the neutral condition, those in the vulnerable condition respond more over time to their fellow human group members (n = 4,590 rounds).

respect to the vulnerable and neutral conditions (c = 0.13, P = 0.03), although there is no significant difference between the silent condition ($\overline{x_{S_i}} = 4.00$ s, SD = 6.73 s) and neutral condition (c = 0.06, P = 0.32) or the vulnerable and silent condition (c = 0.07, P = 0.27) (see Fig. 1*C* and *SI Appendix*, Table S3).

We further find that the difference in the amount of talking by those in the vulnerable condition was primarily driven by one type of utterance, namely, the communications between the human players themselves, with an increase in responses to other humans over time. In other words, participants in the vulnerable condition ($\overline{x_{V_i}} = 3.99$ s, SD = 5.16 s) increased the amount of time they spent responding to the utterances of their other human group members as the game progressed (Fig. 1*D*) compared to those in the neutral condition ($\overline{x_{N_i}} = 1.72$ s, SD = 2.76 s) (c = 0.08, P = 0.04), although there was no difference between the silent condition ($\overline{x_{S_i}} = 1.96$ s, SD = 4.01 s) and the neutral condition (c = 0.04, P = 0.36) or the vulnerable condition and the silent condition (c = 0.04, P = 0.27), in terms of the increase over rounds. No other utterance category (see *SI Appendix* for details) was statistically significant across rounds of the game (*SI Appendix*, Table S4).

Equality in Conversation. In addition to examining the amount of participants' speech, we also explored how equally participants' speech durations were within a group as well as how evenly participants distributed their speech to the two other human members in the group. To quantify the former, we used the following "equality in talking time" (E_{TT}) metric:

$$E_{TT_i} = c\left|\frac{\tau_i}{\sum_1^n \tau_i} - \frac{1}{n}\right|,$$

where τ_i represents the total amount of time participant i spoke during the game, n is the number of human participants (3 in this case), $\sum_1^n \tau_i$ is the total amount of time participant i's group spoke during the game, and c is a normalizing constant, causing E_{TT_i} to have a range of [0, 1]. E_{TT_i} takes on a value of 0 when a participant speaks for a third of the total amount of time their group speaks and a value of 1 when a participant speaks and their group members did not speak at all. We found that equality of time speaking did not differ between the vulnerable robot condition ($\overline{x_{E_{TT_{V_i}}}} = 0.14$, SD = 0.10) and the neutral robot condition ($\overline{x_{E_{TT_{N_i}}}} = 0.14$, SD = 0.11) ($c$ = –0.03, P = 0.88, see *SI Appendix*, Table S5), but there is a significant difference between the neutral and silent conditions ($\overline{x_{E_{TT_{S_i}}}} = 0.25$, SD = 0.17) ($c$ = –0.63, P = 0.001) and the vulnerable and silent conditions (c = –0.66, P = 0.0004). In other words, the distribution of speech by participants in the vulnerable condition did not differ from that in the neutral condition, but participants in the silent condition had the least equal distribution of talking time, as seen in Fig. 2*A*. Thus, the mere presence of a robot that communicates may enhance the equality of talking time in conversation among humans in a group.

To examine how evenly distributed each participant's utterances were toward their fellow human teammates in the human–robot group, we created an "equality in talking partners" (E_{TP}) metric as follows:

$$E_{TP_i} = \frac{\left|\tau_{(Pi,Pj)} - \tau_{(Pi,Pk)}\right|}{\tau_{(Pi,Pj)} + \tau_{(Pi,Pk)}},$$

where $\tau_{(Pi,Pj)}$ represents the talking time of participant i's speech specifically directed at participant j during the game and $\tau_{(Pi,Pk)}$ represents the talking time of participant i's speech specifically directed at participant k during the game. In other words, this measures how balanced a participant's speech is toward the two other human members of their group over the whole game. If a participant directs all of their speech to one participant and none to the other, that participant gets a value of 1. If a participant speaks for the exact same amount of time to each of the other two participants, that participant will receive a value of 0. In other words, values of 1 represent perfect inequality and 0 represents perfect equality. Every human–human pairwise comparison is made for each participant in each group. We found no evidence that participants in the vulnerable condition ($\overline{x_{E_{TP_{V_i}}}} = 0.38$, SD = 0.26) distributed their speech more equally between their fellow human group members than those in the neutral condition ($\overline{x_{E_{TP_{N_i}}}} = 0.43$, SD = 0.29) ($c$ = –0.38, P = 0.16) or between the silent condition ($\overline{x_{E_{TP_{S_i}}}} = 0.57$, SD = 0.34) and neutral condition (c = 0.36, P = 0.25). However, the vulnerable condition is significantly more equal than the silent condition (c = –0.74, P = 0.01), as shown in Fig. 2*B* (*SI Appendix*, Table S6). In other words, speech seems to be more balanced in the conditions with a vulnerable robot compared to a silent robot.

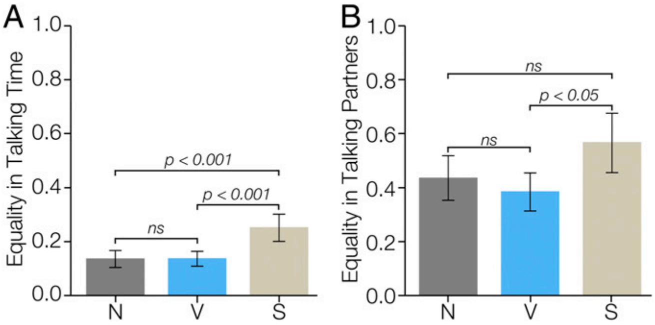

Fig. 2. Equality of conversation by condition. Although there was (*A*) no statistical difference between the vulnerable (V) and neutral (N) conditions in the equality in talking time (n = 150 participants; one group did not speak at all and was excluded), the silent (S) condition was less equal than either of the other two conditions. (*B*) Also, participants in the vulnerable condition directed their utterances more equally to each of their human group members than participants in the silent condition, as measured by the total amount of time spent talking to each participant's two human partners (n = 144 participants; participants who didn't speak at all or who did not make directed utterances were excluded).

Self-Reported Group Dynamics. Using comments that the participants provided in the postexperiment survey, we analyzed how participants perceived their own group's dynamics (*SI Appendix*, Table S7). Comments were reliably coded by two coders (see *SI Appendix* for details) into the following four categories: quiet, positive, supportive, and fun.

We found that those in the vulnerable condition thought of their groups as being less quiet than did those in the neutral condition (P = 0.03) and that the silent condition did not differ from the neutral condition (P = 0.29)—in accordance with the foregoing analyses of actual talk time. The vulnerable condition did not differ from the silent condition (P = 0.23). Those in the vulnerable condition viewed their groups as more positive than did those in the neutral condition (P = 0.04) and the silent condition (P = 0.01). There was no difference between the silent and neutral conditions (P = 0.54). Those in the vulnerable condition also viewed their groups as being more fun than did those in the neutral condition (P = 0.03) or the silent condition (P = 0.02), and the silent and neutral conditions did not differ (P = 0.71). There was no difference between the vulnerable and neutral conditions in how supportive participants found their groups (P = 0.77), nor was there a difference between the silent and neutral conditions (P = 0.13) or the vulnerable and silent conditions (P = 0.19). See *SI Appendix* for more details. In summary, participants in the vulnerable condition described their groups as more pleasant overall than those in either of the two other conditions.

Discussion

We experimentally evaluated human-to-human conversational dynamics in a group with three human group members and either a vulnerable, neutral, or silent robot group member. Participants in a group with a vulnerable robot spoke more overall than participants in a group with a neutral or silent robot and increasingly over the rounds of the game than participants in a group with a neutral robot. This difference is largely driven by responses to other human group members, which demonstrates an increasing social engagement effect in the conversation. It is likely that the vulnerable statements made by the robot helped participants feel more comfortable conversing with their group members. In short, it is not just the presence of a speaking robot in a hybrid system that changes human–human communication, but rather the nature of the robot's speech—specifically, speech that is vulnerable.

As the total amount of time spent talking may not fully capture the nature of the conversation, we also examined the equality of each member's speech amount, relative to their group members, as well as how evenly distributed each member's talking was between their fellow human group members. We did not find a significant difference in the equality in talking time measure in groups with the vulnerable robot and in groups with the neutral robot, but participants in the silent robot condition had significantly less equal talking times, as if the mere presence of a talking humanoid robot helped even out the total amount of speaking between each human. Additionally, we found that those in a group with the vulnerable robot distributed their talking more equally between the other two human members of their group than those in a group with the silent robot. In short, we found that the vulnerable utterances by the robot not only influenced groups to talk more, but also positively shaped the directionality of the utterances to be more evenly balanced between the other two human members of the group compared to the silent condition.

These results demonstrate that robots are able to positively shape group conversational dynamics and emphasize the influential role that robots can have on the way we converse and interact with each other in the context of human–robot groups. As robots and other forms of machine intelligence (such as digital assistants and online bots) become increasingly prevalent in our daily lives in what we have called "hybrid systems," (9) they will likely shape our actions, relationships, and conversations. They will critically affect our interactions with other people. Without an understanding of how machines may influence our connections with the people around us, we may encounter negative and unexpected consequences of interacting with these robots (41, 42).

For instance, there has been a great deal of concern about the capacity for physical robots and autonomous bots to damage human interactions (for instance by causing car accidents or spreading propaganda), but equally it is the case that such agents may improve human welfare. Prior work has shown the capacity for online bots to facilitate coordination (9) or decrease racist speech online (43). Here, we show that physical robots can improve conversational dynamics among humans. We believe that research investigating the consequences of social agents on human behavior, like the work presented here, will help agent designers optimally take such social spillovers into account and may help inform the general public on how to use artificial intelligence (AI) technology in beneficial ways.

Machines endowed with particular kinds of programming might have effects not only by modifying the behavior of humans with whom they directly interact, but also by modifying the behavior between the humans themselves when they interact with other humans in a group, thus creating diverse sorts of behavioral cascades. This can happen even when people actually know they are interacting with machines, as in the present case [and as shown in other work (9)]. In this sense, even simple AI agents might be able to serve a beneficial function, shaping the actions of their human counterparts and modifying human–human interactions and not just human–robot interactions. Our work illustrates the idea that, in hybrid systems of humans and machines, robots could help people work better together.

Materials and Methods

Recruitment Procedures. After approval by the Yale University Institutional Review Board (IRB), we recruited 195 participants, but due to experimental malfunctions (e.g., video/audio failing to record, system glitches that prevented game play), 153 participants (51 groups) are included in this analysis. Most of our participants (73%) had very low familiarity with the other participants in their group. Further details, along with the regression procedures used to control for variation in measured covariates across treatment groups, are provided in *SI Appendix*.

The Collaborative Game. To set up a collaborative environment where the robot's vulnerability might help to ease group tension and facilitate positive group dynamics, we designed a collaborative game to be played on individual tablets where each human participant and an autonomous humanoid robot would be equal contributors (see *SI Appendix* for more details). This collaborative game for three humans and one robot (a Softbank Robotics NAO robot) was built to be played on individual Android tablets running our bespoke Railroad Route Construction game.

The game we developed was created to give players the sense that they are playing collaboratively—by making success contingent on everyone in the group completing their part of the game successfully for the whole group to succeed. In other words, if one person fails, everyone fails. To successfully complete a round, each player needs to place eight pieces along the most efficient route for the railroad. This same game procedure was completed for all 30 rounds for each group. The group's score was the total number of successfully completed rounds for that group.

Experimental Procedure. When participants arrived, they were given a consent form (or adolescent assent form after receiving parental consent), which was signed by the participant (and guardian, if applicable) before participating in any aspect of the study. Then, the participants were given a preexperiment survey. After completion of the survey, the participants entered the experiment room to play the game described above. Further details are provided in *SI Appendix*. The 3 human participants and the robot played 30 rounds of the Railroad Route Construction game. In the neutral condition, the robot made neutral utterances at the end of each round and did not acknowledge when it had made a mistake. In the vulnerable condition, the robot made vulnerable utterances when each round finished, which included acknowledging its mistakes. In the silent condition, the robot did not say anything at the end of each round. In the vulnerable and neutral conditions, the end-of-round utterances were approximately the same length across conditions. For more details, consult *SI Appendix*.

During the game, each player would take a railroad track piece from the options on the right-hand side of the screen and drag it to the active play area to complete the track (see *SI Appendix* for more details). Each player was instructed to build the optimal route during each round by creating the shortest contiguous track on their tablet. The group succeeded in a round if each player completed their path. The group failed a round when at least one participant failed to complete their track.

Over the course of the game, by design, each participant, including the robot, failed twice (see *SI Appendix* for details). In the case of the humans, we engineered this outcome by making no suitable track parts available to be placed on the game-play surface. Participants did not perceive this as accidental and felt responsibility for their "failure." After the game, participants completed a postexperiment survey. When this survey was complete, they were given a debrief form about the true nature of the game and a payment for their time.

Vulnerable utterances made by the robot after each round of the game in the vulnerable condition were constructed based on the principles of psychological safety (44). The utterances by the robot were of three types: self-disclosure, personal story, and humor. Both self-disclosure and telling a personal story show vulnerability through the sharing of personal information with another member of the group (34). The robot used self-disclosure in utterances such as: "Sorry guys, I made the mistake this round. I know it may be hard to believe, but robots make mistakes too." Personal stories from the robot consisted of sharing information about prior experiences and passions.

These utterances were statements such as: "Awesome! I bet we can get the highest score on the scoreboard, just like my soccer team went undefeated in the 2014 season!" Lastly, the robot expressed vulnerability by using humor, as humorous expressions constitute a social gamble that has the potential to reduce discomfort within the group (35, 36). An example of a humorous phrase the robot used was "Sometimes failure makes me angry, which reminds me of a joke: Why is the railroad angry? Because people are always crossing it!" (We quite realize that this joke is corny, but that was the point.) In contrast, neutral statements were fact-based and did not contain any personal information or humor. For example, the robot would say, "We have completed 17 rounds thus far and have successfully built 76 percent of them." All end-of-round utterances can be found in *SI Appendix*.

To verify that the comments made by the robot at the end of each round were perceived to be vulnerable in the vulnerable condition and fact-based in the neutral condition, we asked 210 human judges (via Amazon Mechanical Turk) to assess pairs of utterances. Judges were provided with random pairs of utterances (one utterance from each condition in a pair) and were asked which of the two indicated more vulnerability. By a ratio of nearly 3:1, the judges selected the utterances that we designed to be vulnerable as vulnerable (details can be found in *SI Appendix*).

Measures of Conversational Dynamics. In order to analyze the conversational dynamics within groups, we transcribed and categorized each utterance made by the participants using ELAN software (45). All utterances made throughout the game were included. These utterances fell broadly into two categories: comments and responses. We define comments as utterances that are addressed to others within the group, but that are not contingent on what has been said previously in the conversation. In other words, comments are new thoughts. In contrast, we define a response as an utterance that is dependent on what has just been said in the conversation. Often, responses are to comments, but they can also be a response to a response. Both comments and responses could be directed speech to certain individuals in particular or to the group as a whole (see *SI Appendix* for details on the utterance categorizations). For example, a comment would be an utterance such as, "Alright, we need to beat the top team" followed by a response of "We can do it!" The coder agreement (Cohen's κ) across the four coders on these categorizations was $k = 0.92$.

Control Variables. Because characteristics of the participants could shape their willingness to engage socially with their group members, we asked participants to report their age, gender, familiarity with others in their group, and degree of extraversion.

Before entering the experiment room, we provided participants with a survey in which we asked them to select their familiarity with the two other human participants in their group on a 5-point scale. In addition, we asked participants if they were "Facebook friends" with, or had the telephone numbers of, the other members of their group. For more details on how familiarity was calculated, consult *SI Appendix*. In the postexperiment survey, we asked participants about the psychological safety of the group, what the participant thought of the robot, how extraverted the participant was, and a series of general questions about the game and the group.

We believed that whether or not a participant was extraverted would have an impact on their willingness to engage socially with their other group members. To avoid priming participants, we asked participants 6 yes or no questions from the abbreviated, revised Eysenck personality questionnaire (EPQR-A) (46) in the postexperiment survey. Participants were given a score from 0 (lowest extraversion) to 6 (highest extraversion) by adding the number of affirmative answers to the 6 questions in our survey. More details can be found in *SI Appendix*.

Data Availability. Data and reproducible R scripts are available on the Human Nature Lab website at http://humannaturelab.net/publications/vulnerable-robots (raw participant videos are not available due to IRB stipulations).

ACKNOWLEDGMENTS. We thank Adam Erickson for experimental design assistance as well as Evy Roberts, Ely Sibarium, Victoria Dombrowski, Taylor Triplett, and Anmolpreet Kaur for assisting with video coding. We are grateful for comments on the manuscript provided by Marcus Alexander, Laura Forastiere, and Giorgios Iosifidis. Support for this work was provided by the Robert Wood Johnson Foundation and the National Science Foundation Awards 1139078 and GR104893.

1. A. Pentland, The new science of building great teams. *Harv. Bus. Rev.* **90**, 60–69 (2012).
2. N. Katz, D. Lazer, H. Arrow, N. Contractor, Network theory and small groups. *Small Group Res.* **35**, 307–332 (2004).
3. M. T. Rivera, S. B. Soderstrom, B. Uzzi, Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annual Rev. Sociol.* **36**, 91–115 (2010).
4. A. L. Hill, D. G. Rand, M. A. Nowak, N. A. Christakis, Emotions as infectious diseases in a large social network: The SISa model. *Proc. Biol. Sci.* **277**, 3827–3835 (2010).
5. S. G. Barsade, The ripple effect: Emotional contagion and its influence on group behavior. *Adm. Sci. Q.* **47**, 644–675 (2002).
6. R. M. Bond *et al*., A 61-million-person experiment in social influence and political mobilization. *Nature* **489**, 295–298 (2012).
7. J. H. Fowler, N. A. Christakis, Cooperative behavior cascades in human social networks. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5334–5338 (2010).
8. N. Riether, F. Hegel, B. Wrede, G. Horstmann, "Social facilitation with social robots?" in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, New York, NY, 2012), pp. 41–48.
9. H. Shirado, N. A. Christakis, Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* **545**, 370–374 (2017).
10. A. Ramachandran, S. S. Sebo, B. Scassellati, "Personalized robot tutoring using the assistive tutor pOMDP (AT-POMDP)" in *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI, Palo Alto, California, 2019), pp. 8050–8057.
11. A. Ramachandran, A. Litoiu, B. Scassellati, "Shaping productive help-seeking behavior during robot-child tutoring interactions" in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, Piscataway, NJ, 2016), pp. 247–254.
12. I. Leite *et al*., Narratives with robots: The impact of interaction context and individual differences on story recall and emotional understanding. *Frontiers in Robotics and AI* **4**, 29 (2017).
13. G. Briggs, M. Scheutz, How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *Int. J. Soc. Robot.* **6**, 343–355 (2014).
14. M. Salem, G. Lakatos, F. Amirabdollahian, K. Dautenhahn, "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust" in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, New York, NY, 2015), pp. 141–148.
15. W. A. Bainbridge, J. Hart, E. S. Kim, B. Scassellati, "The effect of presence on human-robot interaction" in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication* (IEEE Press, Piscataway, NJ, 2008), pp. 701–706.
16. J. W. Crandall *et al*., Cooperating with machines. *Nat. Commun.* **9**, 233 (2018).
17. B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, N. Hagita, "Footing in human-robot conversations: how robots might shape participant roles using gaze cues" in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (ACM, New York, NY, 2009), pp. 61–68.
18. E. Short, M. J. Mataric, "Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions" in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (IEEE Press, Piscataway, NJ, 2017), pp. 385–390.
19. S. Strohkorb *et al*., "Improving human-human collaboration between children with a social robot" in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (IEEE Press, Piscataway, NJ, 2016), pp. 551–556.
20. M. F. Jung, N. Martelaro, 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* (ACM, New York, NY, 2015), pp. 229-236.
21. H. Tennent, S. Shen, M. Jung, "Micbot: A peripheral robotic object to shape conversational dynamics and team performance" in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction* (IEEE Press, Piscataway, NJ, 2019), pp. 133–142.
22. G. Skantze, M. Johansson, J. Beskow, "Exploring turn-taking cues in multi-party human-robot discussions about objects" in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (ACM, New York, NY, 2015), pp. 67–74.
23. M. Vázquez *et al*., "Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze" in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, New York, NY, 2017), pp. 42–52.
24. B. Scassellati *et al*., Improving social skills in children with ASD using a long-term, in-home social robot. *Science Robotics* **3**, eaat7544 (2018).
25. H. Admoni, T. Weng, B. Hayes, B. Scassellati, "Robot nonverbal behavior improves task performance in difficult collaborations" in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, Piscataway, NJ, 2016), pp. 51–58.
26. F. Correia, S. Mascarenhas, R. Prada, F. S. Melo, A. Paiva, "Group-based emotions in teams of humans and robots" in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, New York, NY, 2018), pp. 261–269.
27. J. L. Levere, Artificial intelligence, like a robot, enhances museum experiences. *The New York Times*, 25 October 2018. https://www.nytimes.com/2018/10/25/arts/artificial-intelligence-museums.html. Accessed 22 April 2019.
28. K. Tingley, Learning to love our robot co-workers. *New York Times Magazine*, 23 February 2017. https://www.nytimes.com/2017/02/23/magazine/learning-to-love-our-robot-co-workers.html. Accessed 11 March 2019.
29. N. Salomons, M. van der Linden, S. Strohkorb Sebo, B. Scassellati, "Humans conform to robots: Disambiguating trust, truth, and conformity" in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, New York, NY, 2018), pp. 187–195.

30. G. R. Jones, J. M. George, The experience and evolution of trust: Implications for cooperation and teamwork. *Acad. Manage. Rev.* **23**, 531–546 (1998).
31. R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust. *Acad. Manage. Rev.* **20**, 709–734 (1995).
32. G. A. Van Kleef, C. K. De Dreu, A. S. Manstead, "An interpersonal approach to emotion in social decision making: The emotions as social information model" in *Advances in Experimental Social Psychology* (Elsevier, 2010), vol. 42, pp. 45–96.
33. L. R. Wheeless, A follow-up study of the relationships among trust, disclosure, and interpersonal solidarity. *Hum. Commun. Res.* **4**, 143–157 (1978).
34. P. C. Cozby, Self-disclosure: A literature review. *Psychol. Bull.* **79**, 73–91 (1973).
35. O. H. Lynch, Humorous communication: Finding a place for humor in communication research. *Commun. Theory* **12**, 423–445 (2002).
36. C. M. Smith, L. Powell, The use of disparaging humor by group leaders. *South. Speech Commun. J.* **53**, 279–292 (1988).
37. N. Martelaro, V. C. Nneji, W. Ju, P. Hinds, "Tell me more: Designing HRI to encourage more trust, disclosure, and companionship" in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, Piscataway, NJ, 2016), pp. 181–188.
38. C. Breazeal, B. Scassellati, "How to build robots that make friends and influence people" in *Proceedings 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No. 99CH36289)* (IEEE Press, Piscataway, NJ, 1999), pp. 858–863.
39. S. Strohkorb Sebo, M. Traeger, M. Jung, B. Scassellati, "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams" in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, New York, NY, 2018), pp. 178–186.
40. R. M. Siino, J. Chung, P. J. Hinds, "Colleague vs. tool: Effects of disclosure in human-robot collaboration" in *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication* (IEEE Press, Piscataway, NJ, 2008), pp. 558–562.
41. B. Corgnet, R. Hernán-Gonzalez, R. Mateo, Rac(g)e Against the Machine? Social Incentives When Humans Meet Robots (2019). https://halshs.archives-ouvertes.fr/halshs-01994021/document. Accessed 26 April 2019.
42. I. Rahwan *et al.*, Machine behaviour. *Nature* **568**, 477–486 (2019).
43. K. Munger, Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Polit. Behav.* **39**, 629–649 (2017).
44. A. Edmondson, Psychological safety and learning behavior in work teams. *Adm. Sci. Q.* **44**, 350–383 (1999).
45. ELAN, Version 5.8. https://tla.mpi.nl/tools/tla-tools/elan/. Accessed 12 October 2019.
46. L. J. Francis, L. B. Brown, R. Philipchalk, The development of an abbreviated form of the revised Eysenck personality questionnaire (EPQR-A): Its use among students in England, Canada, the USA and Australia. *Pers. Individ. Dif.* **13**, 443–449 (1992).

COMPUTER SCIENCES

SOCIAL SCIENCES