PLOS ONE

RESEARCH ARTICLE

Cooperation with autonomous machines through culture and emotion

Celso M. de Melo[1]*, **Kazunori Terada**[2]

1 CCDC U.S. Army Research Laboratory, Playa Vista, CA, United States of America, **2** Gifu University, Gifu, Yanagido, Japan

* celso.miguel.de.melo@gmail.com

Citation: de Melo CM, Terada K (2019) Cooperation with autonomous machines through culture and emotion. PLoS ONE 14(11): e0224758. https://doi.org/10.1371/journal.pone.0224758

Editor: Francisco C. Santos, Instituto Superior Técnico, Univesidade de Lisboa, PORTUGAL

Received: July 19, 2019

Accepted: October 20, 2019

Published: November 11, 2019

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0224758

Data Availability Statement: All data collected and analyzed during this study is available in the SI.

Funding: This research was supported by JSPS KAKENHI Grant Number JP16KK0004 [KT], and the US Army. The funder had no role in study

Abstract

As machines that act autonomously on behalf of others–e.g., robots–become integral to society, it is critical we understand the impact on human decision-making. Here we show that people readily engage in social categorization distinguishing humans ("us") from machines ("them"), which leads to reduced cooperation with machines. However, we show that a simple cultural cue–the ethnicity of the machine's virtual face–mitigated this bias for participants from two distinct cultures (Japan and United States). We further show that situational cues of affiliative intent–namely, expressions of emotion–overrode expectations of coalition alliances from social categories: When machines were from a different culture, participants showed the usual bias when competitive emotion was shown (e.g., joy following exploitation); in contrast, participants cooperated just as much with humans as machines that expressed cooperative emotion (e.g., joy following cooperation). These findings reveal a path for increasing cooperation in society through autonomous machines.

Introduction

Humans often categorize others as belonging to distinct social groups, distinguishing "us" versus "them", and this categorization influences cooperation, with decisions tending to favor in-group members and, at times, discriminating against out-group members [1–4]. As autonomous machines–such as self-driving cars, drones, and robots–become pervasive in society [5–7], it is important we understand whether humans also apply social categories when engaging with these machines, if decision making is shaped by these categories and, if so, how to overcome unfavorable biases to promote cooperation between humans and machines. Here we show that, when deciding whether to cooperate with a machine, people engage, by default, in social categorization that is unfavorable to machines but, it is possible to override this bias by having machines communicate cues of affiliative intent. In our experiment, participants from two distinct cultures (Japan and United States), engaged in a social dilemma with humans or machines that had a virtual face from the same culture or not and, additionally, expressed emotion conveying cooperative or competitive intent. The results confirmed that people cooperated less with machines than with humans perceived to be from a different culture, except when the emotion indicated an intention to cooperate. Our findings strengthen earlier research indicating that humans rely on social categories–such as culture–to detect coalitional

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

alliances [8–10] and show that this mechanism applies to autonomous machines. The results further confirm that it is possible to override these default encodings with situational cues of affiliative intent [10, 11]–in our case, communicated through emotion. The results also have important practical implications for the design of autonomous machines, indicating that it is critical to understand the social context these machines will be immersed in and, adopt mechanisms to convey affiliative intent in order to minimize unfavorable biases and promote cooperation with humans.

In social interaction, people categorize others into groups while associating, or self-identifying, more with some–the in-groups–than others–the out-groups [1–4]. This distinction between "us" and "them" can lead to a bias that favors cooperation with in-group members [2, 4]. An evolutionary justification for such a bias is to promote prosperity of the in-group which, in turn, leads to increased chance of survival and longer-term benefits for the individual [12]. In fact, perceptions of group membership have consistently been found to be effective in promoting cooperation in social dilemmas [13]. But, do humans engage in social categorization when engaging with autonomous machines?

Experimental evidence suggests that people categorize machines similarly to how they do with other people: in one experiment, in line with gender stereotypes, people assigned more competence to computers with a female voice than a male voice on the topic of "love and relationships" [14]; in another experiment, people perceived computers with a virtual face of the same race as being more trustworthy and giving better advice than a computer with a face of a different race [15]; in a third experiment, machines with voices that had an accent of the same culture or not as the participants, impacted perceptions of the appropriateness of the machine's decisions in social dilemmas [16]. Findings such as these led Reeves and Nass [17] to propose a general theory arguing that to the extent that machines display human-like cues (e.g., human appearance, verbal and nonverbal behavior), people will treat them in a fundamentally social manner and automatically apply the same rules they use when interacting with other people. A strict interpretation of this theory would, thus, suggest that not only can people apply categories to machines, but machines are in-group members.

However, studies show that, despite being able to treat machines as social actors, people make different decisions and show different patterns of brain activation when engaging with machines, when compared to humans. As detailed in our recent review of this research [18]: "Gallagher et al. [19] showed that when people played the rock-paper-scissors game with a human there was activation of the medial prefrontal cortex, a region of the brain that had previously been implicated in mentalizing (i.e., inferring of other's beliefs, desires and intentions); however, no such activation occurred when people engaged with a machine that followed a known predefined algorithm. McCabe et al. [20] found a similar pattern when people played the trust game with humans vs. computers, and Kircher et al. [21], Krach et al. [22], and Rilling et al. [23] replicated this finding in the prisoner's dilemma. (. . .) Sanfey et al. [24] further showed that, when receiving unfair offers in the ultimatum game, people showed stronger activation of the bilateral anterior insula–a region associated with the experience of negative emotions–when engaging with humans, when compared to machines." The evidence, thus, suggests that people experienced less emotion and spent less effort inferring mental states with machines than with humans. These findings align with research that shows that people perceive less mind in machines than in humans [25]. Denying mind to others or perceiving inferior mental ability in others, in turn, is known to lead to discrimination [26]. Overall, these findings suggest that machines are treated, at least by default, as members of an out-group. Effectively, recent studies showed that participants favored humans to computers in several economic games, including the ultimatum, dictator, and public goods social dilemma [18, 27]. As autonomous machines become pervasive in society, it is important we find solutions to

promote cooperation between humans and machines, including overcoming these types of unfavorable biases.

To accomplish this, we first look at cross categorization, i.e., the idea of associating a positive category with an entity to mitigate the impact of a negative category [3]. Research indicates that humans have the cognitive capacity to process multiple categories simultaneously and crossing categories can reduce intergroup bias [3]. Here we look at culture–pertaining to the shared institutions, social norms, and values of a group of people [9]–as a possible moderator to this bias with machines. Culture is an appropriate first choice in the study of interaction with machines as research indicates that people respond to cultural cues in machines, such as language style [28], accent [16], social norms [29], and race [15]. Research also shows that individuals from different cultures can have different initial expectations about whether the interaction is cooperative or competitive, follow different standards of fairness, and resort to different schemas when engaging in social decision making [30, 31]. Finally, culture has been argued to be important in explaining cooperation among non-kin [8, 32]. Our first hypothesis, therefore, was that associating positive cues of cultural membership could mitigate the default unfavorable bias people have towards machines.

However, it may not always be possible to control the social categories people associate with machines and, thus, it is important to consider a more reliable solution to overcoming negative biases with autonomous machines. Research indicates that, even though social categorization is pervasive, it is possible to override initial expectations of coalitional alliances based on social categories by resorting to more situationally-relevant cues of affiliate intent [10, 11]. Kurzban, Tooby and Cosmides [10] confirmed that people form expectations about coalitions from race but, these were easily overridden by counterparts' verbal statements about intentions to cooperate. They argue social categories are useful only insofar as they are relevant in identifying coalitions. In that sense, cues specific to the social situation should supersede the influence of social categories in perceptions of coalitional alliances. Here we consider emotion expressions for this important social function.

Emotion expressions influence human decision making [33, 34]. One of the important social functions of emotions is to communicate one's beliefs, desires, and intentions to others [35] and, in that sense, emotion displays can be important in identifying cooperators [36]. de Melo, Carnevale, Read and Gratch [37] showed that people were able to retrieve information about how counterparts were appraising the ongoing interaction in the prisoner's dilemma and, from this information, make inferences about the counterparts' likelihood of cooperating in the future. Moreover, emotion displays simulated by machines have also been shown to influence human behavior in other social settings [38]. Our second hypothesis, thus, was that emotion expressions could override expectations of cooperation based on cultural membership.

We present an experiment where participants engaged in the iterated prisoner's dilemma. In this dilemma, two players make a simultaneous decision to either defect or cooperate. Standard decision theory argues that individuals should always defect because defection is the best response to any decision the counterpart may make: if you believe your counterpart will defect, you should defect as well; if you believe your counterpart is going to cooperate, then you still maximize your payoff by defecting. However, if both players follow this reasoning, then they will both be worse off than if they had cooperated. Participants engaged in 20 rounds of this dilemma. Repeating the dilemma a finite number of rounds does not change this prediction since the last round is effectively a one-shot prisoner's dilemma and, by induction, so is every previous round. However, in practice, people often cooperate in such social dilemmas [13]. The payoff matrix we used is shown in Fig 1A. The points earned in the task had real financial consequences as they would be converted to tickets for a $30 lottery. Finally, to prevent any

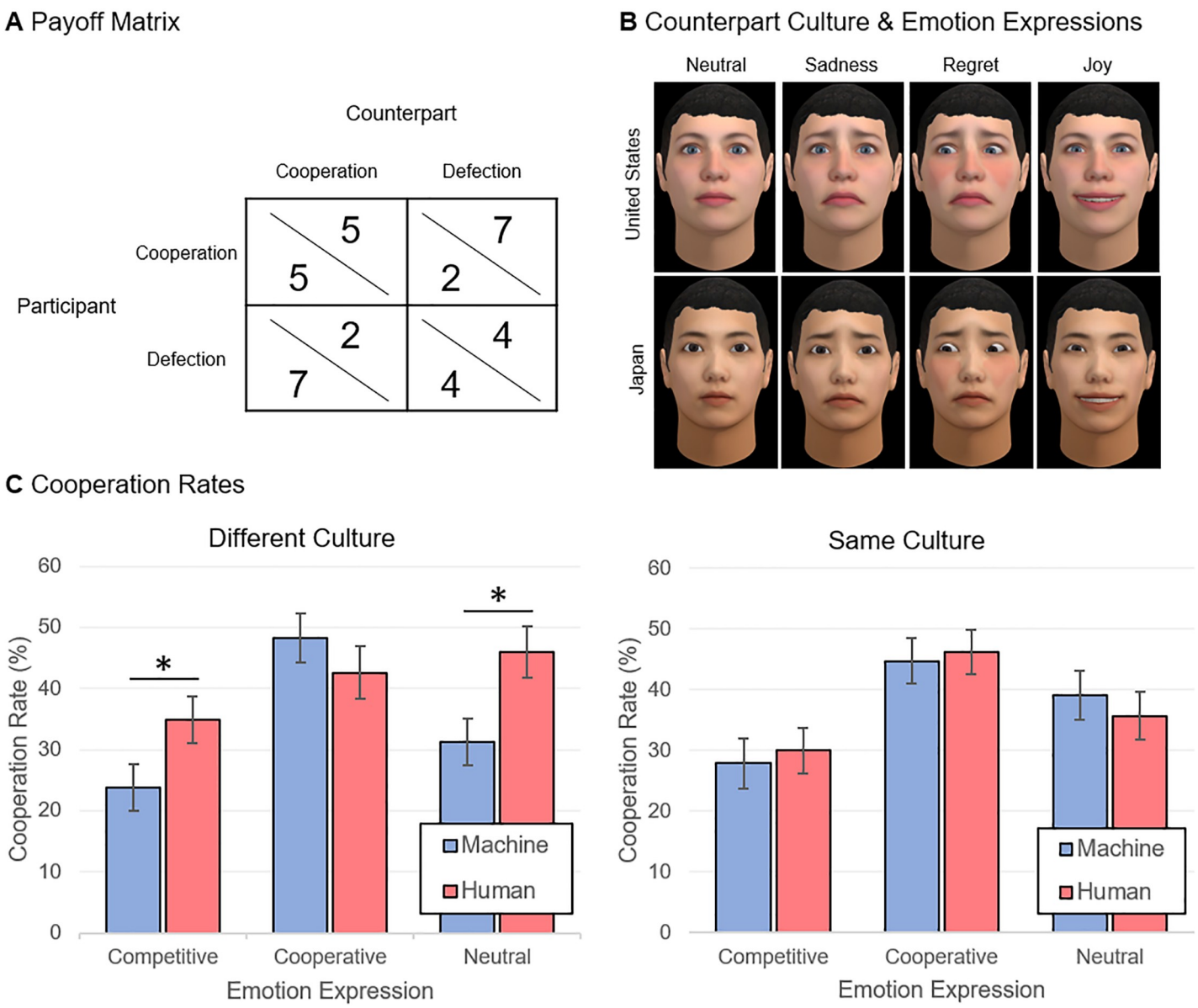

Fig 1. Experimental manipulations and cooperation rates. (**A**) The payoff matrix for the prisoner's dilemma, (**B**) Counterparts' virtual faces typical in the United States (top) and Japan (bottom) and corresponding emotion expressions, (**C**) Cooperation rates when the counterpart was from a different culture as the participant (left) or the same culture (right). The error bars correspond to standard errors. * $p < .05$.

https://doi.org/10.1371/journal.pone.0224758.g001

reputation concerns, the experiment was fully anonymous–i.e., the participants were anonymous to each other and to the experimenters (see the Materials and methods section for details on how this was accomplished).

Participants were told they would engage in the prisoner's dilemma with either another participant or with an autonomous machine. In reality, to maximize experimental control, they always engaged with a computer script. Similar methods have been followed in previous studies of human behavior with machines [18, 19, 21, 23, 24] and the experimental procedures were fully approved by the Gifu University IRB. This script followed a tit-for-tat strategy (starting with a defection) and showed a pre-defined pattern for emotion expression (see below). The focus of the experiment was in studying whether participants would cooperate distinctly

with humans vs. machines and, if so, whether culture or emotion expressions could moderate this effect.

To manipulate perceptions of cultural membership, counterparts were given virtual faces that were either typical in the United States or in Japan–see Fig 1B. See the Supporting Information (S1 File) appendix for a validation study, with a separate sample of participants, for perceptions of the corresponding ethnicities in these faces (S1 File). We recruited 945 participants from the United States ($n = 468$) and Japan ($n = 477$) using online pools (see Materials and methods for more details about recruitment and sample demographics). Participants were either matched with a counterpart of the same or different culture, counterbalanced across participants. This manipulation allowed us to study, in two distinct cultures, how participants behaved with (human or machine) counterparts that were either in- or out-group members to the culture.

Counterparts expressed emotion through their virtual faces corresponding to either a competitive, neutral, or cooperative orientation. Building on earlier work that shows that emotion expressions can shape cooperation in the prisoner's dilemma [37], we chose the following patterns: competitive emotions–regret following mutual cooperation (given that it missed the opportunity to exploit the participant), joy following exploitation (participant cooperates, counterpart defects), sadness in mutual defection and, neutral otherwise; cooperative emotions–joy following mutual cooperation, regret following exploitation, sadness in mutual defection, and neutral otherwise; neutral emotions–neutral expression for all outcomes. For a validation study for perception of the intended emotions, please see the SI appendix (S1 File). In sum, we ran a $2 \times 2 \times 3$ between-participants factorial design: *counterpart type* (human vs. machine) × *counterpart culture* (United States vs. Japan) × *emotion* (competitive vs. neutral vs. cooperative). Our main measure was cooperation rate, averaged across all rounds.

Results

The focus of our analysis was two-fold: (1) understand the cooperation rate when participants engaged with humans vs. machines that had the same vs. different culture; and, (2) understand the moderating role of emotion expressions. To accomplish this, we first split the data into two sets: the first corresponding to pairings of participants with counterparts of a different culture, and the second corresponding to pairings with counterparts of the same culture. For each set, we ran a participant sample (United States vs. Japan) × counterpart type (human vs. machine) × emotion (competitive vs. neutral vs. cooperative) between-participants factorial ANOVA. Fig 1C shows the cooperation rates for this analysis.

When participants were paired with counterparts of a different culture, we found the expected main effect of counterpart type, $F(1, 436) = 4.17$, $P = 0.042$, partial $\eta^2 = 0.01$: participants cooperated more with humans ($M = 41.15$, $SE = 2.39$) than machines ($M = 34.45$, $SE = 2.25$). There was also a main effect of emotion–$F(2, 436) = 8.17$, $P < 0.001$, partial $\eta^2 = 0.04$ –and Bonferroni post-hoc tests showed that: participants cooperated more with cooperative ($M = 45.42$, $SE = 2.97$) than competitive counterparts ($M = 29.34$, $SE = 2.70$), $P < .001$; and, participants tended to cooperate more with neutral ($M = 38.63$, $SE = 2.85$) than competitive counterparts, $P = 0.055$. Interestingly, however, there was a counterpart type × emotion interaction, $F(2, 436) = 3.48$, $P = 0.032$, partial $\eta^2 = 0.16$. To get insight into this interaction, we split the data by emotion condition and ran a participant sample × counterpart type between-participants factorial ANOVA. This analysis revealed that: when counterparts showed competitive or neutral emotions, there was a main effect of counterpart type–respectively, $F(3, 159) = 5.39$, $P = 0.022$, partial $\eta^2 = 0.03$, and $F(3, 146) = 5.83$, $P = 0.017$, partial $\eta^2 = 0.04$ –with participants cooperating more with humans than machines; but, when the counterparts showed

cooperative emotion, there was no statistically significant difference in cooperation between humans and machines, $F(3, 131) = 0.83$, $P = 0.365$. Finally, there was no main effect of participant sample–$F(1, 436) = 1.28$, $P = 0.259$ –and no statistically significant interactions with the other factors, suggesting that the effects apply both for participants in Japan and the United States.

When participants engaged with counterparts of the same culture, in contrast to the previous case, there was no statistically significant main effect of counterpart type–$F(1, 485) = 0.01$, $P = 0.972$ –or counterpart type × emotion interaction–$F(2, 485) = 0.29$, $P = 0.749$. The main effect of emotion, on the other hand, was still statistically significant–$F(2, 485) = 558.71$, $P < 0.001$, partial $\eta^2 = 0.54$ –and Bonferroni post-hoc tests revealed that participants: cooperated more with cooperative ($M = 45.55$, $SE = 2.61$) than competitive counterparts ($M = 28.99$, $SE = 2.78$), $P < .001$; tended to cooperate more with cooperative than neutral counterparts ($M = 37.42$, $SE = 2.81$), $P = 0.104$; and, tended to cooperate more with neutral than competitive counterparts, $P = 0.100$. Once again, there was no main effect of participant sample–$F(1, 485) = 0.97$, $P = 0.813$ –and no statistically significant interactions.

For further insight and analyses, please refer to the SI for a table with all descriptive statistics (S1 Table) and the raw data (S2 File).

Discussion

Despite the changes autonomous machines promise to bring to society, here we show that humans will resort to familiar psychological mechanisms to identify alliances and collaborate with machines. Whereas autonomous machines may be perceived by default as out-group members [18–27], our experimental results with participants from two distinct cultures (Japan and United States) indicate that simple cues of cultural in-group membership–based on the ethnicity of the machine's virtual face–can mitigate this unfavorable bias in the decisions people make with machines, when compared to humans. More fundamentally, the results indicate that situational cues of affiliative intent–in our experiment, through expressions of emotion–can override default expectations created from social categorization and promote cooperation between humans and machines.

Our results confirm that, in the context of interaction with autonomous machines, social categorization occurs naturally and influences human decision making. Participants cooperated more with machines that were perceived to belong to the same culture than those that were not. Culture has been argued to be central to cooperation with non-kin [8, 32] and our results indicate that this can extend to interactions with machines. This is also in line with earlier research showing that humans readily apply social rules, including social categorization, when interacting with machines in social settings [14–17]. However, the results further show that, when counterparts were perceived to belong to a different culture, participants cooperated less with machines than with humans. This reinforces that, despite being able to treat machines in a social manner, by default, people still show an unfavorable bias with machines, when compared to humans [18–27]; in other words, machines are perceived, by default, as belonging to an out-group. By crossing this default negative category with a positive cue for culture membership, as we do in our experiment, we were able to mitigate this bias, suggesting that multiple social categorization [3] can be a solution for reducing intergroup bias with machines.

A more reliable solution, however, may be to communicate affiliative intent through emotion expression. Emotion had the strongest effect in our experiment, showing that even a machine from a different culture group could be treated like an in-group member through judicious expression of emotion–in our case, joy following cooperation and regret after

exploitation. Emotion expressions have been argued to serve important social functions [35] and help regulate decision making [34, 36, 37], and here we strengthen research indicating that emotion in autonomous machines is a powerful influencer of human behavior [38], including social decision making [27, 37]. More generally, in line with earlier research [10, 11], this research indicates that default coalition expectations from social categories can be overridden if situational coalition information is available. This is encouraging as it may not always be possible to control the social categories that will be perceived in machines.

The work presented here has some limitations that introduce opportunities for future work. First, the effects of emotion expressions, when compared to the neutral (control) condition, could be strengthened. For instance, whereas people cooperated more with cooperative than competitive counterparts, there was only a trend for higher cooperation with cooperative than neutral counterparts. To strengthen these effects, people may need to be exposed longer to the expressions (e.g., by increasing the number of rounds), or an even clearer emotional signal may need to be communicated (e.g., verbal or multimodal expression of emotion [37]). Second, social discrimination among humans is complex and here we have only begun studying this topic. Dovidio and Gartner [39] point out that, rather than blatant and open, modern racism tends to be subtle. In fact, in some cases, to avoid being perceived as racist, people can over-compensate when interacting with out-group members [40]. This may explain why, in our case, participants tended to cooperate more, when no emotion was shown, with humans of a different than the same culture. This compensation mechanism, however, did not occur with machine counterparts, where we see the expected in-group bias–thus, exemplifying that machines are often treated less favorably than humans. It is important, therefore, to understand whether interaction with machines will also evolve to reflect subtler forms of discrimination, especially as they become more pervasive in society. Finally, here we looked at race as a signal for cultural membership, but several other indicators have been studied [15, 16, 28, 29]. Follow-up work should study the relative effects of these different indicators–e.g., speech accent or country of origin–on cooperation with machines. As noted in the SI, perception of ethnicity from race can still lead to some ambiguity, with Japanese participants more easily identifying the Caucasian face as being from the United States than American participants. Multiple complementary signals could, thus, potentially be used to strengthen perceptions of in-group cultural membership and, consequently, help further reduce bias.

The results presented in this paper have practical importance for the design of autonomous machines. The existence of a default unfavorable bias towards machines means designers should take action if they hope to achieve the levels of cooperation seen among humans. It wouldn't be satisfactory to conceal that a machine is autonomous–i.e., is not being directly controlled by a human–as there is increased expectation in society of transparency and interpretability from algorithms [41]. Instead, designers should consider the broader social context and the cognitive mechanisms driving humans to promote cooperation with machines. Here we show that social categorization is pervasive and can be leveraged–through simple visual [15], verbal [16, 28], or behavioral [29] cues–to increase perceptions of group membership with machines and, subsequently, encourage more favorable decisions. However, on the one hand, it may not always be possible to provide those cues and, on the other, social categories can be activated by unexpected cues. In this sense, designers should consider specific situational cues as a more explicit signal of the machine's affiliative intent. Here we exemplify how this signaling can be achieved, effectively and naturally, through expressions of emotion.

At a time of increasing divisiveness in society, it may seem unsurprising that autonomous machines are perceived as being outsiders and, consequently, being less likely to benefit from the advantages afforded to in-group members. However, autonomous machines presumably act on behalf of (one or several) humans who, logically, are the ultimate targets of any decision

made with these machines. Nevertheless, it is encouraging to learn that behavior with these machines appears to be driven by the same psychological mechanisms in human-human interaction. This introduces familiar opportunities for reducing intergroup bias, such as cultural membership priming and cross categorization. It is also comforting to learn that if a machine intends to communicate with humans, and is able to effectively communicate that affiliative intent, then any default expectations derived from social categorization can be overridden. Since autonomous machines can be designed to take advantage of these cognitive-psychological mechanisms driving human behavior, they introduce a unique opportunity to promote a more cooperative society.

Materials and methods

This section describes details for the experimental methods that are not described in the main body of the text.

Experimental task

Building on previous work [37], the prisoner's dilemma game was recast as an investment game and described as follows to the participants: "You are going to play a two-player investment game. You can invest in one of two projects: project green and project blue. However, how many points you get is contingent on which project the other player invests in. So, if you both invest in project green, then each gets 5 points. If you choose project green but the other player chooses project blue, then you get 2 and the other player gets 7 points. If, on the other hand, you choose project blue and the other player chooses project green, then you get 7 and the other player gets 2 points. A fourth possibility is that you both choose project blue, in which case both get 4 points". Thus, choosing project green corresponded to the cooperative choice, and project blue to defection. Screenshots of the software are shown in the Supporting Information (S3 and S4 Figs). The software was presented in English to US participants and translated to Japanese for participants in Japan.

Participant samples

All participants were recruited from online pools: the US sample was collected from Amazon Mechanical Turk, and the Japanese sample from Yahoo! Japan Crowdsourcing. Previous research shows that studies performed in online platforms can yield high-quality data and successfully replicate the results of behavioral studies performed on traditional pools [42]. To estimate sample size per country, we used G*Power 3. Based on earlier work [18, 37], we predicted a small to medium effect size (Cohen's $f = 0.20$). Thus, for $\alpha = .05$ and statistical power of .85, the recommended total sample size was 462 participants. In practice, we recruited 468 participants in the US and 477 in Japan. The demographics for the US sample were as follows: 63.2% were males; age distribution– 18 to 21 years, 0.9%; 22 to 34 years, 58.5%; 35 to 44 years, 21.6%; 45 to 54 years, 10.7%; 55 to 64 years, 6.2%; over 64 years, 2.1%; ethnicity distribution–Caucasian, 77.3%; African American, 10.3%; East Indian, 1.3%; Hispanic or Latino, 9.2%; Southeast Asian, 6.0%. The demographics for the Japanese sample were as follows: 67.4% were males; age distribution– 18 to 21 years, 0.6%; 22 to 34 years, 15.7%; 35 to 44 years, 36.4%; 45 to 54 years, 34.7%; 55 to 64 years, 10.4%; over 64 years, 2.1%; ethnicity distribution–East Indian, 0.6%; Southeast Asian, 99.4%.

Financial incentives

Participants in the US were paid $2.00 for participating in the experiment, whereas participants in Japan were paid 220 JPY (~$2.00). Moreover, they had the opportunity to earn more

money according to their performance in the task. Each point earned in the task was converted to a ticket for a lottery worth $30.00 for the US sample and 3,000 JPY (~$27.00) for the Japanese sample.

Full anonymity

All experiments were fully anonymous for participants. To accomplish this, counterparts had anonymous names and we never collected any information that could identify participants. To preserve anonymity with respect to experimenters, we relied on the anonymity system of the online pools we used. When interacting with participants, researchers are never able to identify the participants, unless we explicitly ask for information that may serve to identify them (e.g., name, email, or photo), which we did not. This experimental procedure is meant to minimize any possible reputation effects, such as a concern for future retaliation for the decisions made in the task.

Data analyses

The main analysis consisted of participant sample (United States vs. Japan) × counterpart type (human vs. machine) × emotion (competitive vs. neutral vs. cooperative) between-participants factorial ANOVAs on average cooperation rate, for the case where participants were matched with counterparts of the same culture and the case where counterparts had a different culture. To get insight into statistically significant interactions, we ran Bonferroni post-hoc tests and follow-up participant sample × counterpart type between-participants factorial ANOVAs, where the data was split on emotion condition.

Ethics

All experimental methods were approved by the Medical Review Board of Gifu University Graduate School of Medicine (IRB ID#2018–159). As recommended by the IRB, written informed consent was provided by choosing one of two options in the online form: 1) "I am indicating that I have read the information in the instructions for participating in this research and have had a chance to ask any questions I have about the study. I consent to participate in this research.", or 2) "I do not consent to participate in this research." All participants gave informed consent and, at the end, were debriefed about the experimental procedures.

Supporting information

S1 Fig. Perception of emotion in Japanese and Caucasian faces.
(TIF)

S2 Fig. Perception of ethnicity in Japanese and Caucasian faces by US and Japanese participant samples.
(TIF)

S3 Fig. The prisoner's dilemma software for the United States participants. The counterpart in this case has the same culture and is showing cooperative emotions.
(TIF)

S4 Fig. The prisoner's dilemma software for the Japanese participants. The counterpart in this case has the same culture and is showing competitive emotions.
(TIF)

S1 File. Appendix with validation experiment for emotion and ethnicity perception in virtual faces.
(DOCX)

S2 File. CSV file with raw data.
(CSV)

S1 Table. Descriptive statistics for main experiment.
(DOCX)

Author Contributions

Conceptualization: Celso M. de Melo, Kazunori Terada.

Data curation: Celso M. de Melo.

Formal analysis: Celso M. de Melo, Kazunori Terada.

Funding acquisition: Celso M. de Melo, Kazunori Terada.

Investigation: Celso M. de Melo, Kazunori Terada.

Methodology: Kazunori Terada.

Resources: Celso M. de Melo, Kazunori Terada.

Software: Celso M. de Melo.

Validation: Celso M. de Melo, Kazunori Terada.

Writing – original draft: Celso M. de Melo, Kazunori Terada.

Writing – review & editing: Celso M. de Melo, Kazunori Terada.

References

1. Tajfel H, Turner J (1986) The social identity theory of intergroup behavior. In: Worchel S, Austin W, editors. Psychology of intergroup relations. Nelson-Hall; pp. 7–24.
2. Brewer M (1979) In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. Psychol Bull 86: 307–324.
3. Crisp R, Hewstone M (2007) Multiple social categorization. Adv Exp Soc Psychol 39: 163–254.
4. Baillet D, Wu J, De Dreu C (2014) Ingroup favoritism in cooperation: A meta-analysis. Psychol Bull 140: 1556–1581. https://doi.org/10.1037/a0037737 PMID: 25222635
5. de Melo C, Marsella S, Gratch J (2019) Human cooperation when acting through autonomous machines. Proc Nat Acad Sci U.S.A. 116: 3482–3487.
6. Bonnefon J-F, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352: 1573–1576. https://doi.org/10.1126/science.aaf2654 PMID: 27339987
7. Stone R, Lavine M (2014) The social life of robots. Science 346: 178–179. https://doi.org/10.1126/science.346.6206.178 PMID: 25301612
8. Richerson P, Newton E, Baldini R, Naar N, Bell A, Newson L et al. (2016) Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. Behav Brain Sci 39, 1–68.
9. Betancourt H, Lopez S (1993) The study of culture, ethnicity, and race in American Psychology. Am Psychol 48: 629–637.
10. Kurzban R, Tooby J, Cosmides L (2001) Can race be erased? Coalitional computation and social categorization. Proc Nat Acad Sci U.S.A. 98: 15387–15392.
11. Pietraszewski D, Cosmides L, Tooby J (2014) The content of our cooperation, not the color of our skin: An alliance detection system regulates categorization by coalition and race, but not sex. PLOS ONE 9, e88534. https://doi.org/10.1371/journal.pone.0088534 PMID: 24520394
12. Bernhard H, Fischbacher U, Fehr E (2006) Parochial altruism in humans. Nature 442: 912–915. https://doi.org/10.1038/nature04981 PMID: 16929297

13. Kollock P (1998) Social dilemmas: The anatomy of cooperation. Annu Rev Sociol 24: 183–214.
14. Nass C, Moon Y, Green N (1997) Are computers gender-neutral? Gender stereotypic responses to computers. J App Soc Psychol 27: 864–876.
15. Nass C, Isbister K, Lee E-J (2000) Truth is beauty: Researching embodied conversational agents. In: Cassell J, editor. Embodied conversational agents. MIT Press; pp. 374–402.
16. Khooshabeh P, Dehghani M, Nazarian A, Gratch J (2017) The cultural influence model: When accented natural language spoken by virtual characters matters. AI & Soc 32: 9–16.
17. Reeves B, Nass C (1996) The media equation: How people treat computers, television, and new media like real people and places, Cambridge University Press.
18. de Melo C, Marsella S, Gratch J (2016) People do not feel guilty about exploiting machines. ACM T Comput-Hum Int 23: 1–17.
19. Gallagher H, Anthony J, Roepstorff A, Frith C (2002) Imaging the intentional stance in a competitive game. NeuroImage 16: 814–821. PMID: 12169265
20. McCabe K, Houser D, Ryan L, Smith V, Trouard T (2001) A functional imaging study of cooperation in two-person reciprocal exchange. Proc Nat Acad Sci U.S.A. 98: 11832–11835.
21. Kircher T, Blümel I, Marjoram D, Lataster T, Krabbendam L, Weber J et al. (2009) Online mentalising investigated with functional MRI. Neurosci Lett 454: 176–181. https://doi.org/10.1016/j.neulet.2009.03.026 PMID: 19429079
22. Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, Kircher T (2008) Can machines think? Interaction and perspective taking with robots investigated via fMRI. PLOS ONE 3: 1–11.
23. Rilling J, Gutman D, Zeh T, Pagnoni G, Berns G, Kilts C (2002) A neural basis for social cooperation. Neuron 35: 395–405. https://doi.org/10.1016/s0896-6273(02)00755-9 PMID: 12160756
24. Sanfey A, Rilling J, Aronson J, Nystrom L, Cohen J (2003) The neural basis of economic decision-making in the ultimatum game. Science 300: 1755–1758. https://doi.org/10.1126/science.1082976 PMID: 12805551
25. Waytz A, Gray K, Epley N, Wegner D (2010) Causes and consequences of mind perception. Trends Cogn Sci 14: 383–388. https://doi.org/10.1016/j.tics.2010.05.006 PMID: 20579932
26. Haslam N (2006) Dehumanization: An integrative review. Pers Soc Psychol Rev 10: 252–264. https://doi.org/10.1207/s15327957pspr1003_4 PMID: 16859440
27. Terada K, Takeuchi C (2017) Emotional expression in simple line drawings of a robot's face leads to higher offers in the ultimatum game. Front. Psychol. 8: https://doi.org/10.3389/fpsyg.2017.00724 PMID: 28588520
28. Rau P, Li Y, Li D (2009) Effects of communication style and culture on ability to accept recommendations from robots. Comp Hum Behav 25: 587–595.
29. Aylett R, Paiva A (2012) Computational modelling of culture and affect. Emotion Rev 4: 253–263.
30. Brett J, Okomura T (1998) Inter- and intracultural negotiations: US and Japanese negotiators. Acad Manag J 41: 495–510.
31. Henrich J, Boyd R, Bowles S, Camerer C, Fehr E et al. (2000) In search of homo economicus: behavioral experiments in 15 small-scale societies. Am Econ Rev 91: 73–78.
32. Efferson C, Lalive R, Fehr E (2008) The coevolution of cultural groups and ingroup favoritism. Science 321: 1844–1849. https://doi.org/10.1126/science.1155805 PMID: 18818361
33. Damasio A (1994) Descartes' error: Emotion, reason, and the human brain, Putnam Press.
34. Van Kleef G, De Dreu C, Manstead A (2010) An interpersonal approach to emotion in social decision making: The emotions as social information model. Adv Exp Soc Psychol 42: 45–96.
35. Morris M, Keltner D (2000) How emotions work: An analysis of the social functions of emotional expression in negotiations. Res Organ Behav 22: 1–50.
36. Frank R (2004) Introducing moral emotions into models of rational choice. In: Manstead A, Frijda N, Fischer A, editors. Feelings and emotions. Cambridge University Press; pp. 422–440.
37. de Melo C, Carnevale P, Read S, Gratch J (2014) Reading people's minds from emotion expressions in interdependent decision making. J Pers Soc Psychol 106: 73–88. https://doi.org/10.1037/a0034251 PMID: 24079297
38. Marsella S, Gratch J, Petta P (2010) Computational models of emotion. In: Scherer K, Bänziger T, Roesch E, editors. A blueprint for an affectively competent agent: Cross-fertilization between emotion psychology, affective neuroscience, and affective computing. Oxford University Press; pp. 21–45.
39. Dovidio J, Gaertner S (2000). Aversive racism and selection decisions: 1989 and 1999. Psychol Sci 11: 319–323.

40. Axt J, Ebersole C, Nosek B (2016) An unintentional, robust, and replicable pro-Black bias in social judgment. Soc Cogn 34: 1–39.

41. Goodman B, Flaxman S (2017) European Union regulations on algorithmic decision making and a "Right to Explanation". AI Magazine Fall 2017: 51–57.

42. Paolacci G, Chandler J, Ipeirotis P (2010) Running experiments on Amazon Mechanical Turk. Judg Decis Making 5: 411–419.