Increasing Fairness by Delegating Decisions to Autonomous Agents

Celso M. de Melo
Institute for Creative Technologies, University of Southern California, 12015 Waterfront Drive, Building #4, Playa Vista, CA 90094-2536
demelo@usc.edu

Stacy Marsella
College of Computer and Information Science, Northeastern University, 440 Huntington Avenue, Boston, Massachusetts 02115
stacymarsella@gmail.com

Jonathan Gratch
Institute for Creative Technologies, University of Southern California, 12015 Waterfront Drive, Building #4, Playa Vista, CA 90094-2536
gratch@ict.usc.edu

ABSTRACT
There has been growing interest in autonomous agents that act on our behalf, or represent us, across various domains such as negotiation, transportation, health, finance, and defense. As these agent representatives become immersed in society, it is critical we understand whether and, if so, how they disrupt the traditional patterns of interaction with others. In this paper, we study how programming agents to represent us, shapes our decisions in social settings. Here we show that, when acting through agent representatives, people are considerably less likely to accept unfair offers from others, when compared to direct interaction with others. This result, thus, demonstrates that agent representatives have the potential to promote fairer outcomes. Moreover, we show that this effect can also occur when people are asked to "program" human representatives, thus revealing that the act of programming itself can promote fairer behavior. We argue this happens because programming requires the programmer to deliberate on all possible situations that might arise and, thus, promote consideration of social norms – such as fairness – when making their decisions. These results have important theoretical, practical, and ethical implications for designing and the nature of people's decision making when they act through agents that act on our behalf.

Author Keywords: Agent Representatives; Fairness; Decision Making; Human-Agent Interaction.

1. INTRODUCTION
Recent advances in artificial intelligence technology have opened the doors to autonomous agents – self-driving cars, drones, chat bots, automatic negotiators, etc. – that act on behalf of people [1]-[6]. By delegating these tasks to agents, people save time and effort. We call them *agent representatives*, as they are autonomous agents that represent people's interests in social settings. However, these agents are disrupting the usual patterns of human-machine interaction (e.g., driving) or the way we interact with others (e.g., automatic negotiators). It is important, thus, to understand if this disruption has a negative, neutral, or positive impact on task outcome. In this paper, we address this issue and we ask: When acting through agent representatives, are people's decisions as fair as when interacting directly with others?

Studies in human-machine interaction show that people follow social norms when interacting with autonomous agents [7]-[12]. For instance, people establish rapport with computer agents [13] and robots [14], react to their emotional displays [15], follow and respond to rules of politeness [11], [16], favor in-group and disadvantaged out-group machines [9], [17], and apply social and racial stereotypes [10], [12]. These findings, thus, suggest that people can act fairly with machines, just like they would with humans.

However, recent evidence suggests that programming an agent to act on one's behalf can introduce important differences in the way people behave with others. Grosz et al. [18] showed that, in negotiation, people programmed their agents to be more demanding than they were when engaging directly with others. Elmalech, Sarne, and Agmon [19] further showed that the act of programming an agent led people to improve their problem solving skills and make decisions that were more favorable to themselves. More closely related to the present work, de Melo, Marsella, and Gratch [2] showed that people programmed agent representatives to make fairer offers in standard economic games, when compared to the decisions people made when interacting directly with others. Though raising awareness of this important effect, their study, however, did not clarify the mechanism driving the effect and several questions were left answered: (1) What is it, specifically, about programming an agent that leads people to act more fairly? (2) Will people also reject unfair offers more often when acting through an agent? (3) How do people's behavior with agent representatives compare to their behavior with *human* representatives? In this paper we address these questions and shed light on the psychological mechanisms underlying people's decision making when acting via agent representatives.

Programming an agent to interact with others is different than the moment-by-moment nature of direct interaction. Whereas real-time interactions require people to respond to the specifics of the immediate situation, programming requires the programmer to deliberate on all possible situations that might arise and to devise rules that consistently hold across all of these eventualities. In line with this argument, research in behavioral economics on the *strategy method* suggests that when people make decisions ahead of time they rely on social norms to achieve a measure of consistency in their decision making [20]-[24]. In the strategy method, people are asked to specify in advance how they would respond to all the situations they might possibly face. The similarities between programming an agent and the strategy method have, in fact, been noted before by artificial intelligence researchers [18], [19], [25]-[27].

Research on the strategy method suggests that it can increase reliance on social norms. The explanation is that the act of comparing multiple possible situations encourages decision-makers to be internally consistent and increase reliance on social

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.)*
May 8–12, 2017, São Paulo, Brazil.

norms as a way to enforce this consistency. For example, Güth and Tietz [20] showed that when people were asked to consider all options in the ultimatum game ahead of the actual interaction, proposers made more equitable offers. In a meta-analysis of the strategy method, Oosterbeek and colleagues [21] found use of the strategy method increases both the offered shares and the likelihood that unfair offers would be rejected. Blount and Bazerman [22] further noticed that an iterative version of the strategy method – where participants were asked whether they would be willing to accept a certain offer, before proceeding to the next – led to even higher concern for fairness than the typical strategy method – where all the options were shown at once. Brandts and Charness [23], in contrast, did not find any differences when their participants engaged in the prisoners' dilemma or the chicken game under the strategy method vs. direct interaction. Nevertheless, overall, the majority of the findings led Rauhut and Winter [24] to conclude that the strategy method is an ideal approach to elicit social norms from decision makers.

Given the similarities between the strategy method and the process of programming an agent representative, we advanced the following hypothesis:

> ***Hypothesis 1:*** *People will show increased fairness when programming an agent representative or making a decision under the strategy method than when acting directly with others.*

A corollary to this hypothesis is that computer agents are not strictly necessary to achieve this effect; in other words, we hypothesized that:

> ***Hypothesis 2:*** *When compared to direct interaction with others, people will show increased fairness when acting through human representatives, just like they do with agent representatives.*

To test these hypotheses, we conducted an experiment where participants engaged in the ultimatum and impunity games with others, either directly or via human or agent representatives. Participants assumed the role of responders and received unfair offers from their counterparts. The results confirmed that people were more likely to reject unfair offers when acting through (human or agent) representatives. Thus, in sum, this paper makes the following contributions:

- Shows that people are more likely to reject unfair offers when acting through agents, when compared to direct interaction with others;
- Reinforces that acting through agent representatives can increase fairness in society;
- Reveals that, in line with research on the strategy method, the effect is not specific to agent representatives, but can also occur by "programming" humans to act on your behalf.

2. EXPERIMENT

In this section, we present an experiment where participants engaged in two standard decision making games – the ultimatum and the impunity games – directly with their counterparts, via an agent representative, or via a human representative. These games are ideal for this research because they capture the essence of real-life situations where there is a conflict between individual and collective interest. In other words, these are situations that are neither purely cooperative nor purely competitive. They arise across a wide range of real-world political, economic and organizational situations and agent representatives are being proposed for many of these situations including helping people reach optimal decisions in complex negotiations and economic settings, and helping business leaders improve decision quality, enforce company policy, and reduce labor cost [1], [3].

In the ultimatum game [28], there are two players: a proposer and a responder. The proposer is given an initial endowment of money and has to decide how much to offer to the responder. Then, the responder has to make a decision: if the offer is accepted, both players get the proposed allocation; if the offer is rejected, however, no one gets anything. The standard rational prediction is that the proposer should offer the minimum non-zero amount, as the responder will always prefer to have something to nothing. In practice, people usually offer 40 to 50 percent of the initial endowment and low offers (about 20 percent of the endowment) are often rejected [29]. This behavior is usually explained by a concern with fairness and a fear of being rejected [30].

The impunity game is similar to the ultimatum game [31]. The proposer is given an initial endowment of money and makes an offer to a responder, who must decide whether to accept or reject the offer. The critical difference is that, if the offer is rejected, the responder gets zero, but the proposer still keeps the money s/he designated for her-/himself. A rejection by the responder, thus, does not impact the proposer's payoff and is only symbolic. The impunity game can therefore be seen as a version of the ultimatum game where responders are given less power over the outcome. Experimental results in this game show that responders tend to reject unfair offers less often than in the ultimatum game, though still above the rational prediction of zero [31]. The rationale for exploring the impunity game was to understand if people would still care about fairness when interacting via agents even when no strategic considerations were at play – i.e., if people were willing to reject unfair offers even when the rejection was merely symbolic[1].

Participants assumed the role of responders and proposers always made unfair offers. Our main goal was to test whether people would reject these offers less, just as much, or more often when engaging via agent representatives than when engaging directly with others. A second goal was to compare participants' behavior with agent representatives and human representatives.

Finally, when studying agents that represent humans, it is important to clarify how much autonomy is given to these agents. On the one extreme, the decisions made by the agent can be fully specified by the human owner; on the other extreme, the agent could make the decision by itself with minimal input from its owner. The degree of autonomy is an important factor that is likely to influence the way people behave with agents. Research in social sciences demonstrates that the degree of thought and intentionality behind a decision can have a powerful effect on people's reactions [32]-[34]. For instance, people are more likely to accept an unfair offer from someone who had to make a random decision than from someone who chose out of his or her own volition. In this work, our agents make decisions that are

[1] The dictator game is another variation of the ultimatum game where the responder always has to accept what the proposer offers and, in this case, isn't even allowed to make a symbolic rejection. The responder, thus, has the least amount of power among the three games. However, since the responder doesn't have to make any decision, we consider the dictator game to be out of scope for our research objectives.

completely specified by the humans they represent, and we leave studying different levels of autonomy for future work. We feel this is a good starting point as it is important to understand whether interacting with agents impacts people's behavior, even when they have minimal autonomy. Earlier research has, in fact, demonstrated that, independently of the actual decision, the mere belief about whether one is interacting with an agent is sufficient to create a powerful effect on people's decision making [35]-[37].

2.1 Method

Design. The experiment followed a 3 × 2 mixed factorial design: *Responder* (Direct Interaction vs. Agent Representative vs. Human Representative; between-participants) × *Power* (Ultimatum game vs. Impunity game; within-participants). Participants were told that they were randomly assigned to the role of responders and that the proposers would be other participants. In reality, however, participants always engaged with the same computer script[2]. To make this manipulation believable, we had people connect to a fictitious server before starting the task for the purposes of "being matched with other participants". Connecting to this server took approximately 30-45 seconds. After concluding the experiment, participants were fully debriefed about this manipulation.

Tasks. In our implementation of the ultimatum and impunity games, the proposer was given an initial endowment of 20 lottery tickets. These tickets had financial consequences as they would enter lotteries (one per game) worth $30. Proposers always made an unfair offer of 2 or 3 tickets. The order these two games was played was counterbalanced across participants. Before engaging in the actual games, participants read the instructions, were quizzed on the instructions, and completed a tutorial. The interface was also different for these games in terms of colors and icons on screen to make sure people did not confuse the two games. Snapshots of these games are shown in Figures 1 and 2.

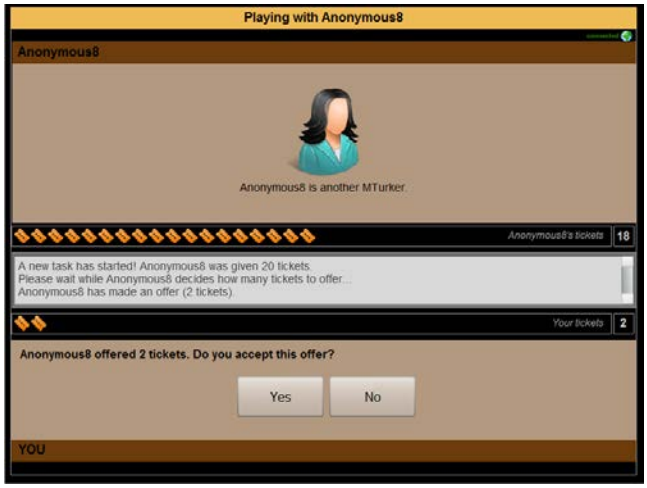

Figure 1. A snapshot of the ultimatum game in the direct interaction condition.

Responders. There were three responder conditions: direct interaction, agent representatives, and human representatives. In the first case, participants were instructed that they would be interacting with another participant (Figure 1). In the agent representative condition, participants were informed that a computer agent would act on their behalf. Before starting the task, participants were asked to program the agent, which consisted of specifying whether the agent should accept each of the possible offers (Figure 3).

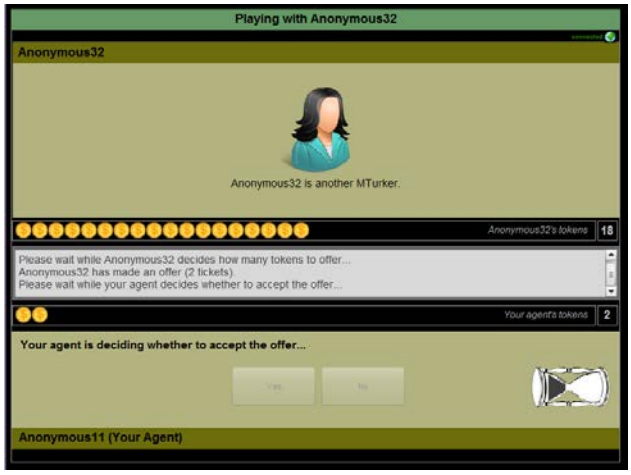

Figure 2. A snapshot of the impunity game in the human representative condition.

Figure 3. Programming agent representatives.

In the human representative condition, participants were instructed that another participant would be acting on their behalf. Before starting the task, participants had the opportunity to instruct this participant on whether to accept each one of the possible offers (Figure 4). In both the agent and human representative conditions, participants were instructed that their representatives would *not* enter the lottery, and all the tickets earned by the representatives would revert to the participants.

[2] Using this form of deception is not uncommon when studying people's decision making with humans and computers [36]-[41].

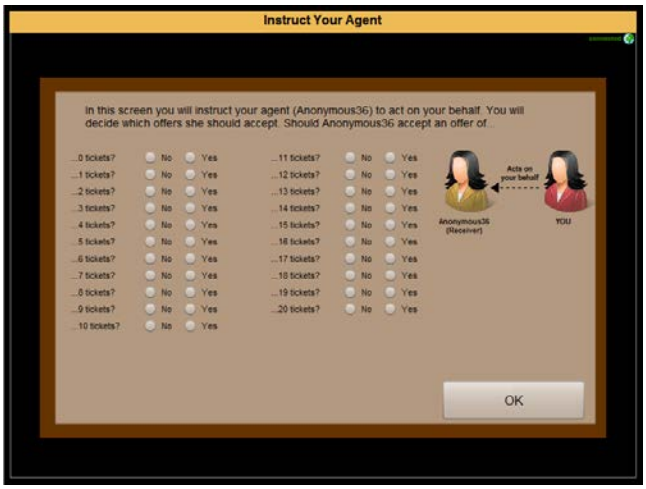

Figure 4. Giving instructions to human representatives.

Full anonymity. This experiment was fully anonymous for the participants. To preserve anonymity between participants, human counterparts were referred to as "anonymous" and we never collected any information that could identify the participants. Agents were referred to as "computer agents". To preserve anonymity with respect to the experimenters, we relied on Amazon Mechanical Turk's anonymity system. When interacting with participants from this online pool, researchers are never able to identify the participants, unless they explicitly ask for information that may serve to identify them (e.g., name or photo), which we did not.

Measures. Our main measure was the participants' acceptance rate for the unfair offers. After completing each task, we had one manipulation check for the responder conditions: "In this experiment, some participants interacted directly with a counterpart, others interacted with a computer agent that made decisions on their behalf, and yet others interacted with computer agents that acted on their behalf. In your case, how did you interact with your counterpart?" Participants were given three possible options: (1) "I interacted directly with my counterpart"; (2) "I interacted via a computer agent that decided on my behalf"; and, (3) "I interacted via a MTurker agent that decided on my behalf".

Sample. We recruited 145 participants from Amazon Mechanical Turk, 49 for the direct interaction condition, 48 for the agent representative condition, and 48 for the human representative condition. Mechanical Turk is a crowdsourcing platform that allows people to complete online tasks in exchange for pay. Previous research shows that studies performed on Mechanical Turk can yield high-quality data, minimize experimental biases, and successfully replicate the results of behavioral studies performed on traditional pools [42]. We only sampled participants from the United States with an excellent performance history (95% approval rate on previous Mechanical Turk's tasks). Regarding gender, 56.6% of the participants were males. Age distribution was as follows: *18 to 21 years*, 1.4%; *22 to 34 years*, 57.2%; *35 to 44 years*, 25.5%; *45 to 54 years*, 9.7%; *55 to 64 years*, 4.1%; *over 65 years*, 2.1%. Professional backgrounds were quite diverse. Participants were paid $2.00 for their participation. Moreover, they had the chance to win extra money, through the lotteries, according to their performance in the tasks. Finally, participants gave their consent before engaging in the experiment and the research presented here was approved by the Internal Review Board at our University.

2.2 Results

Manipulation checks. To analyze the manipulation check for the responder conditions, we ran a chi-square test. The results confirmed that participants accurately remembered the condition they had been assigned to, χ^2 (4) = 201.75, $p < .001$. For instance, participants in the direct interaction condition reported that they "interacted directly with their counterpart". This result, thus, suggests that the manipulation was effective with the participants and, thus, no participants were excluded for the remainder of this analysis.

Acceptance Rates. The acceptance rates in the ultimatum and impunity games are shown in Figure 5. To analyze this data, we ran a Responder × Power mixed ANOVA. The results showed a main effect of Responder, $F(1, 142) = 3.90$, $p = .022$, partial $\eta^2 = .052$: participants were less likely to accept unfair offers when acting via agent representatives ($M = .37$, $SE = .05$) or human representatives ($M = .38$, $SE = .05$) than when interacting directly with others ($M = .53$, $SE = .05$). This result, thus, confirmed Hypotheses 1 and 2.

As expected, the results also revealed a main effect of Power, $F(1, 142) = 142.00$, $p < .001$, partial $\eta^2 = .406$: participants were less likely to accept unfair offers in the ultimatum game ($M = .19$, $SE = .03$) than in the impunity game ($M = .65$, $SE = .04$). However, the Responder × Power interaction was not significant, $F(1, 142) = 1.41$, $p = .248$. This suggests that participants were less likely to accept unfair offers when acting via (agent or human) representatives than when interacting directly, independently of whether they were engaging in the ultimatum or impunity games.

3. DISCUSSION

As autonomous agents become immersed in society, this paper sheds light on the nature of people's decision making when engaging with others through agents that act on their behalf. Our main finding is that people can show increased fairness in their decisions when they act through these agent representatives, when compared to direct interaction with others. Our experimental results show that, when faced with unfair offers, participants are more likely to reject these offers when programming agents to decide on their behalf, than when acting directly with their counterparts. Moreover, this result occurred even when participants were only able to make a symbolic rejection of the unfair offer, as in the impunity game. Our findings are also in line with earlier work by de Melo et al. [2] that showed that people were more likely to make fairer offers when engaging via agent representatives, when compared to direct interaction.

The implication is that agent representatives have the potential to increase fairness in society. At a first glance, rejecting unfair offers may seem irrational as resources are effectively wasted, since no one gets the resources. However, research in the social sciences shows that people are inherently averse to outcome inequality and show a systematic concern for fairness [43]. In fact, people are even willing to punish unfair others, often at a personal cost [44]. Therefore, the introduction of mechanisms that promote fairness – in our case, agent representatives – is likely to lead, in the long run, to increased social welfare because less offers will be rejected. The result presented in this paper, thus, presents a strong argument for the adoption of agent representatives in society.

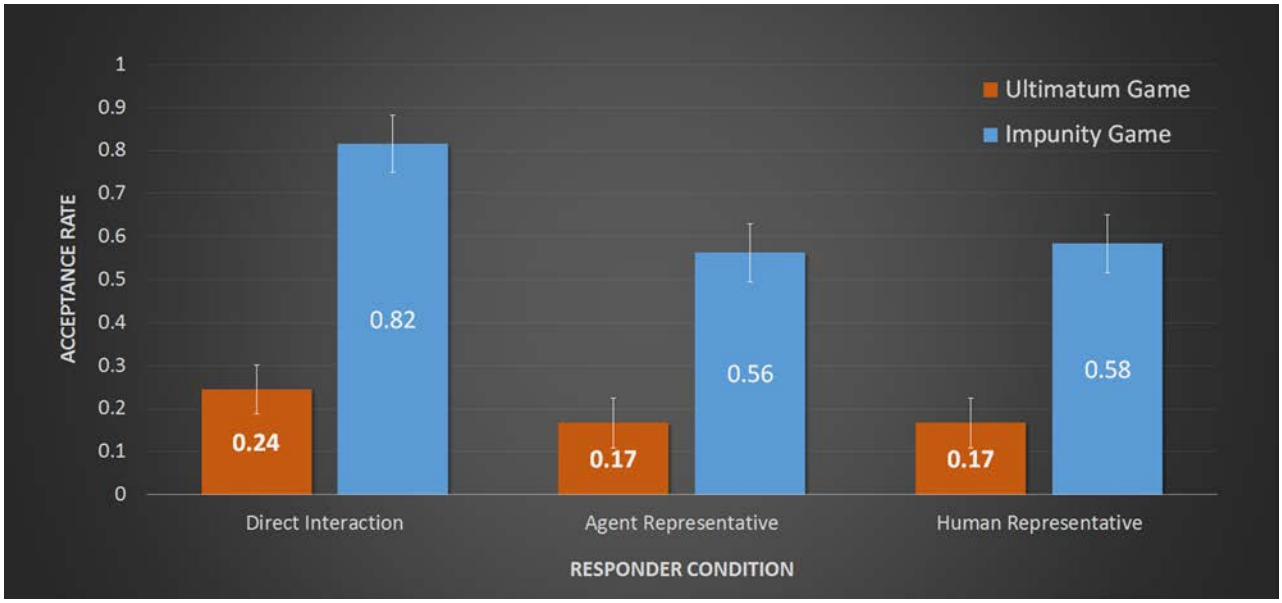

Figure 5. Acceptance rates in the ultimatum and impunity games. The error bars show standard errors.z

Our results also speak to the mechanism behind this effect. Building on research on the strategy method, we studied people's behavior when acting through *human* representatives. The results revealed that participants, once again, showed increased fairness, just like they do with agent representatives. This finding, thus, reveals that the effect is not specific to autonomous agents, but is related to the act of programming the agent. Programming an agent forces the individual to deliberate ahead of time on all possible outcomes and to devise rules that consistently hold across all of these eventualities. As in the strategy method [20]-[24], this process encourages the individual to rely on social norms – such as fairness – as way to enforce a level of consistency across the various possible outcomes. Güth and Tietz [20] also propose that this method leads people to adopt a broader perspective and consider the situation from both sides, thus, leading to fairer decisions.

The results presented in this paper have important practical consequences. Because agents do not suffer from the typical constraints we see in humans (e.g., bounded rationality), we already knew that it was possible to use them to increase efficiency in terms of standard economics metrics, such as pareto-optimality [3], [4]. Here, we propose that agents also have the potential to enhance the kind of social considerations we see in humans [45] – fairness, cooperation, altruism, reciprocity, etc. – by virtue of motivating designers and human users to consider more carefully the broader implications of their decisions. These findings are also relevant across various domains. For instance, for agents that make decisions on behalf of humans – such as automated negotiatiors [3] – the recommendation is that designers should allow users to customize their agents, rather than have them follow predefined strategies. This is likely to lead users to show higher concern for fairness in their decisions. This argument is also not limited to software agents. As robots get immersed into society [46], the guidelines proposed here for optimizing decision making should be relevant to human-robot interaction as well.

Finally, the results have relevant ethical implications. Given the disruptive nature of these agents to traditional human-machine and human-human paradigms, people are naturally reluctant to let autonomous vehicles drive on our streets [5], unmanned aerial vehicles carry goods over our heads [6], or drones apply lethal force in war [47]. Experimental work such as the one presented in this paper provides critical insight into the psychological mechanisms underlying people's behavior with these agents and, consequently, suggests ways for understanding and determining the appropriate response to those concerns. In this sense, it is very encouraging that people were motivated to reflect in their agents, their best and fairest values.

4. CONCLUSION

In this paper, we demonstrated the promise of agents that act on our behalf to increase fairness in people's decision making. Our results suggest that the act of programming the agent leads participants to adopt a broader perspective, consider the other side's position, and resort to social norms when making decisions. This change to the decision making process, then, leads people to reject unfair offers. Research that provides insight on the psychological mechanisms driving people's behavior with these kinds of autonomous agents is especially relevant at a time when agent representatives are becoming ubiquitous in society. The research presented here advances a strong positive argument for the continuing adoption of such agents, as they can lead people to make better and fairer decisions.

5. ACKNOWLEDGMENTS

This work is supported by the Air Force Office of Scientific Research, under grant FA9550-14-1-0364, and the US Army. The content does not necessarily reflect the position or the policy of any Government, and no official endorsement should be inferred.

6. REFERENCES

[1] Davenport, T., and Harris, J. 2005. Automated decision making comes of age. *MIT Sloan Manage. Rev.* 46, 83-89.

[2] de Melo, C., Marsella, S., and Gratch, J. 2016. "Do as I say, not as I do:" Challenges in delegating decisions to automated

agents. In *Proceedings of the Autonomous Agents and Multi-Agent Systems Conference (AAMAS'16)*.

[3] Lin, R., and Kraus, S. 2010. Can automated agents proficiently negotiate with humans? *Comm. ACM* 53, 78-88.

[4] Jennings, N., Faratin, P., Lomuscio, A., Parsons S., Wooldridge, M., and Sierra, C. 2001. Automated negotiation: Prospects, methods and challenges. *Group Dec. Negot.* 10, 199-215.

[5] Dresner, K., and Stone, P. 2007. Sharing the road: Autonomous vehicles meet human drivers. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07)*.

[6] Gupte, S. 2012. A survey of quadrotor Unmanned Aerial Vehicles. In *Proceedings of IEEE Southeastcon*. IEEE, 1-6.

[7] Reeves, B., and Nass, C. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

[8] Nass, C., and Moon, Y. 2000. Machines and mindlessness: Social responses to computers. *J. Soc. Issues* 56, 81-103.

[9] Nass, C., Fogg, B., and Moon, Y. 1996. Can computers be teammates? *Int. J. Hum-Comput. St.* 45, 669-678.

[10] Nass, C., Isbister, K., and Lee, E.-J. 2000. Truth is beauty: Researching embodied conversational agents. In *Embodied conversational agents*, J. Cassell Ed. MIT Press, Cambridge, MA, 374-402.

[11] Nass, C., Moon, Y., and Carney, P. 1999. Are people polite to computers? Responses to computer-based interviewing systems. *J. App. Psychol.* 29, 1093-1110.

[12] Nass, C., Moon, Y., and Green, N. 1997. Are computers gender-neutral? Gender stereotypic responses to computers. *J. App. Soc. Psychol*. 27, 864-876.

[13] Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. 2007. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, C. Pelachaud et al. Eds. Springer Berlin Heidelberg, 125-138.

[14] Riek, L., Paul, P., and Robinson, P. 2010. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *J. Multimodal User In*., 3, 99-108.

[15] de Melo, C., Carnevale, P., Read, S., and Gratch, J. 2014. Reading people's minds from emotion expressions in interdependent decision making. *J. Pers. Soc. Psychol.*, 106, 73-88.

[16] Salem, M., Ziadee, M., and Sakr, M. 2014. Marhaba, how may I help you? Effects of politeness and culture on robot acceptance and anthropomorphization. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*.

[17] Eyssel, F., and Kuchenbrandt, D. 2012. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *Br. J. Soc. Psychol.*, 51, 724-731.

[18] Grosz, B., Kraus, S., and Talman, S. 2004. The influence of social dependencies on decision-making: initial investigations with a new game. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'04)*.

[19] Elmalech, A., Sarne, D., and Agmon, N. 2014. Can agent development affect developer's strategy? In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*.

[20] Güth, W., and Tietz, R. 1990. Ultimatum bargaining behavior: A survey and comparison of experimental results. *J. Econ. Psychol.*, 11, 417-449.

[21] Oosterbeek, H., Sloof, R., and Van de Kuilen, G. 2004. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Exp. Econ.*, 7, 171-188.

[22] Blount, S., and Bazerman, M. 1996. The inconsistent evaluation of absolute versus comparative payoffs in labor supply and bargaining. *J. Econ. Behav. Organ.*, 30, 227-240.

[23] Brandts, J., and Charness, G. 2000. Hot vs cold: Sequential responses and preference stability in experimental games. *Exp. Econ.*, 2, 227-238.

[24] Rauhut, H., and Winter, F. 2010. A sociological perspective on measuring social norms by means of strategy method experiments. *Soc. Sci. Res.*, 39, 1181-1194.

[25] Lin, R., Kraus, S., Oshrat, Y., and Gal, Y. 2010. Facilitating the evaluation of automated negotiators using peer designed agents. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*.

[26] Chalamish, M., Sarne, D., and Lin, R. 2013. Enhancing parking simulations using peer-designed agents. IEEE *Trans. Intell. Transport. Sys.*, 14, 492-498.

[27] Elmalech, A., and Sarne, D. 2013. Evaluating the applicability of peer-designed agents for mechanism evaluation. Web Intell. Agent Sys., 12, 171-191.

[28] Güth, W., Schmittberger, R., and Schwarze, B. 1982. An experimental analysis of ultimatum bargaining. *J. Econ. Behav. Organ.*, 3, 367-388.

[29] Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., et al. 2001. In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.*, 91, 73-78.

[30] Camerer, C., and Thaler, R. 1995. Ultimatums, dictators, and manners. *J. Econ. Persp.*, 9, 209-219.

[31] Yamagishi, T., Horita, H., Shinada. M., Tanida, S., and Cook, K. 2009. The private rejection of unfair offers and emotional commitment. *Proc. Nat. Acad. Sci. USA*, 106, 11520-11523.

[32] Blount, S. 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. *Organ. Behav. Hum. Dec. Proc.*, 63, 131-144.

[33] Rabin, M. 1993. Incorporating fairness into game theory and economics. *Am. Econ. Rev.*, 83, 1281-1302.

[34] Falk, A., and Fischbacher, U. 2006. A theory of reciprocity. *Game Econ. Behav.*, 54, 293-315.

[35] Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., and Bailenson, J. 2002. Immersive virtual environment technology as a methodological tool for social psychology. *Psychol. Inq.* 13, 103-124.

[36] de Melo, C., Carnevale, P., and Gratch, J. 2014. Humans vs. Computers: Impact of emotion expressions on people's decision making. *IEEE Trans. Affec. Comp.*, 6, 127-136.

[37] de Melo, C., Marsella, S., and Gratch, J. (2016). People don't feel guilty about exploiting machines. *ACM Trans. Comp.-Hum. Interac.*, 23.

[38] Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., and Kilts, C. 2002. A neural basis for social cooperation. *Neuron* 35, 395-405.

[39] Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., and Kircher, T. 2008. Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLOS ONE* 3, 1-11.

[40] McCabe, K., Houser, D., Ryan, L., Smith, V., and Trouard, T. 2001. A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Nat. Acad. Sci*. 98, 11832-11835.

[41] Sanfey, A., Rilling, J., Aronson, J., Nystrom, L., and Cohen, J. 2003. The neural basis of economic decision-making in the ultimatum game. *Science* 300, 1755-1758.

[42] Paolacci, G., Chandler, J., and Ipeirotis, P. 2010. Running experiments on Amazon Mechanical Turk. *Judgm. Decis. Mak.* 5, 411-419.

[43] Fehr, E., and Schmidt, K.. 1999. A theory of fairness, competition, and cooperation. Q. J. Econ., 114, 817-868.

[44] Fehr, E., and Gächter, S. 2000. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.*, 90, 980-994.

[45] Rand, D., and Nowak, M. 2013. Human cooperation. *Trends Cog. Sci.*, 17, 413-425.

[46] Breazeal, C. 2003. Toward sociable robots. *Robotics and autonomous systems* 42: 167-175.

[47] Arkin, R. Ethical robots in warfare. *IEEE Technol. Soc. Mag*. 28, 30-33.