Social decisions and fairness change when people's interests are represented by autonomous agents

Celso M. de Melo[a], Stacy Marsella[b], and Jonathan Gratch[a]

[a] Institute for Creative Technologies, University of Southern California, Playa Vista, CA 90094-2536, United States

[b] College of Computer and Information Science, Northeastern University, Boston, MA 02115, USA

Corresponding author: Celso M. de Melo, USC Institute for Creative Technologies, Playa Vista, CA 90094-2536, +1 213-400-1121. E-mail: demelo@usc.edu

Abstract

In the realms of AI and science fiction, agents are fully-autonomous systems that can be perceived as acting of their own volition to achieve their own goals. But in the real world, the term "agent" more commonly refers to a person that serves as a representative for a human client and works to achieve this client's goals (e.g., lawyers and real estate agents). Yet, until the day that computers become fully autonomous, agents in the first sense are really agents in the second sense as well: computer agents that serve the interests of the human user or corporation they represent. In a series of experiments, we show that human decision-making and fairness is significantly altered when agent representatives are inserted into common social decisions such as the ultimatum game. Similar to how they behave with human representatives, people show less regard for other people (e.g., exhibit more self-interest and less fairness), when the other is represented by an agent. However, in contrast to the human literature, people show *more* regard for others and *increased* fairness when "programming" an agent to represent their own interests. This finding confirms the conjecture by some in the autonomous agent community that the very act of programming an agent changes how people make decisions. Our findings provide insight into the cognitive mechanisms that underlie these effects and we discuss the implication for the design of autonomous agents that represent the interests of humans.

Keywords: Agent Representatives, Decision making, Fairness, Strategy Method

1. Agent (Representative): a person who acts for or represents another
2. Agent (Cause): a person or thing that produces a particular effect or change
- Cambridge Dictionary

1. Introduction

In the 2009 movie *Up in the Air*, George Clooney plays a "corporate downsizer." He is hired by companies to handle the uncomfortable task of offering their employees an ultimatum: take this early retirement package or risk termination. Thus, he serves as an agent of the company, representing the company's interests, while allowing it a measure of ethical distance from an uncomfortable social situation. Research in the social sciences illustrates that agents in this sense, as third-party representatives, can deflect blame when actions are seen as unfair or ethically problematic, and lead people to make different decisions than they might if they interacted face-to-face [1-3].

Within AI, the term "agent" is often used in a different sense – as a fully-autonomous system that can be perceived as acting of their own volition to achieve its own goals – yet agents in this sense are really agents in the representative sense as well: autonomous agents ultimately serve the interests of the human user or corporation they represent. Thus, it is natural to ask if using agent representatives change the nature of decisions between people in the same way that human representatives alter these social decisions. In this paper, we investigate this by examining two related questions. On the one hand, we ask whether people behave differently when they interact with an *agent representative* (i.e., an autonomous agent that represents the interests of another person), compared with how they would act towards that person directly (i.e., would they be more or less willing to accept a retirement package from a robotic George Clooney). On the other hand, we ask whether people "program" their autonomous representatives to behave differently than they might behave directly (i.e., would

a company be less willing to fire their employees if they couldn't use an autonomous "corporate downsizer"?).

We examine these questions in the context of typical social decision making settings. These are collective action situations in which there is a conflict between individual and collective interests [4], [5]. 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 [6-8]. These situations are studied through game theory and laboratory experiments using stylized games that capture the essence of real-world predicaments, such as the ultimatum game, prisoner's dilemma or trust game. A key finding from years of research is that human behavior differs systematically from game theoretic predictions, and that that these differences are explained (in part) by the human tendency to follow social norms such as fairness [9].

In this paper, we examine how the introduction of agent representatives changes the fairness of human decisions in two standard social decision making games: the ultimatum game and the impunity game. These games study fairness under different social dependency and incentive structures, thus providing a good test of the generality of our findings. We adopt the conventional definition of fairness used in behavioral economics: i.e., people are deemed fair if they behave consistently with a norm of fairness. In the context of an ultimatum game, this means they make more equitable offers – i.e., an even 50-50 split of resources – to their counterparts and reject inequitable offers made to them.

Research on human representatives suggests that both sides of a social transaction will show less regard for fairness when an interaction is mediated by human agents [1], [10]. For example, Bazerman and colleagues[1] discuss how major pharmaceutical companies prefer to act through intermediaries when dramatically raising the price of a drug. They further show this is an effective strategy, as consumers react with less anger than if the company had raised the price directly. More broadly, the use of intermediaries is one of many ways of increasing the "psychological distance" between parties in a social exchange. For example, findings suggest that fairness concerns are reduced when decisions are separated in time and space [11-15], factors that are likely to occur when interacting via computer agents.

Do agent representatives change decisions in similar ways? Although this question has not been considered explicitly, two lines of research inform our approach to this question. First, an extensive body of research has examined the extent to which people treat computers and robots similarly to how they treat people (e.g., [16]). Although such studies typically describe agents as acting on their own behalf, or leave the other implicit, users likely assume the system represents the interests of the scientists that created it, and thus directly relates to how they might respond to an agent representative. Second, several researchers have considered whether the act of programming an agent might alter the way people make social decisions [17-23]. Although this research is directed at a very different question (namely, will these programs behave the same as real people in social simulations?), it speaks directly to the question of how people might act through an agent representative. We briefly review these lines of work before formulating our experimental hypotheses.

[1] Gino, F., Moore, D., & Bazerman, M. (2008). See no evil: When we overlook other people's unethical behaviour. Harvard Business School NOM Working Paper No. 08-045.

1.1. Are People Fair to Autonomous Agents?

Several studies in human-computer and human-robot interaction show that people follow human social norms when interacting with autonomous agents [16], [24-28]. For instance, people establish rapport with computer agents [29] and robots [30], react to their emotional displays [31], follow and respond to rules of politeness [25], [32], favor in-group and disadvantage out-group machines [26], [33], and apply social and racial stereotypes [27], [28]. The implication from this research, thus, is that, in social decision making, people would be likely to show the same kind of social considerations – including fairness – with autonomous agents as they do with humans.

Although these studies emphasize that people apply social norms to agents, they rarely examined the strength of this tendency. Other lines of research suggest this tendency is relatively weak and largely shaped by beliefs about the "mind" behind the machine [34-37]. To test this, several studies manipulated whether people believe the agent is autonomously controlled by a computer program or is a puppet – or *avatar* – that is manipulated in real-time by another person. The advantage of such studies is they hold the appearance and behavior of the agent constant but manipulate the *mere belief* of what controls this behavior. These studies show that people still follow social norms with autonomous systems, but this tendency is strongly attenuated [38-44]. People engage less in mentalizing – i.e., inferring of other's beliefs, desires and intentions – with agents than with avatars in the exact same economic settings, for the exact same financial incentives [38-41]; they experience less emotion with agents [42-44], are less willing to follow orders [45] and, are less concerned about projecting a positive impression [46]. Most relevant to the current study, people

consistently showed reduced fairness with autonomous agents than with avatars in the dictator, ultimatum, and public goods games [43].

1.2. Are People Fair to Agent Representatives?

Collectively, the previously-mentioned findings suggest that people will be less fair to autonomous agents that appear to act on their own behalf than they would to an agent that is controlled moment-to-moment by a person (i.e., an avatar). However, how would they behave towards an autonomous agent that was acting on behalf of another person (i.e., an agent representative)? Would they treat it more like an autonomous agent or more like an avatar? Although these previous results don't speak directly to this question, they often leave unstated if other humans benefit from the autonomous agent's actions or if the agent is described as acting on its own behalf (e.g. [42]); thus, participants might reasonably assume that the agent represents some person's interests. For example, if a participant is giving money to a robot in an ultimatum game, presumably they don't think the robot will spend its earnings. Therefore, we argue, these findings suggest that people behave less fairly towards autonomous agents *despite* assuming it represents the interests of human actors. Consequently, these findings comparing autonomous agents to avatars, suggest people will treat agent representatives more similarly to how they treat autonomous agents and, thus, less fairly than they would avatars.

Research in the social sciences also demonstrates that increasing perceived psychological distance can lead people to treat others less fairly [1], [11-15]. This previous work studied the effects of psychological distance by manipulating perceived anonymity, social distance, temporal distance, and physical distance. Hoffman, McCabe, and Smith [11] showed that in conditions of full anonymity – i.e., participants could not be identified by their counterparts or the experimenters – people offered much less than when anonymity was not

preserved. Researchers also showed that people offered more to counterparts that were closer in their social networks (e.g., friends) [12, 13]. Pronin, Olivola, and Kennedy [14] demonstrated that temporal distance also affected decision making; in their studies, people showed larger psychological distance between their future selves, who were treated the same as different people, than their present selves. Finally, various researchers have shown that physical proximity can lead to increased cooperation [15]. These findings are relevant because interacting with an agent representative can lead to increased perceived psychological distance to the counterpart and, thus, less fairness.

In sum, the findings in these studies comparing autonomous agents to avatars and the research on the impact of psychological distance on fairness lead to our first hypothesis:

> ***Hypothesis 1:*** *People will show reduced fairness when interacting with agents that represent others than when interacting directly with humans.*

1.3. Are People Fair when "Programming" an Agent Representative?

The aforementioned experiments examined how people treat agents that represent others' interests, but how might people treat others when acting through an agent? Would they instruct the agent to act as they themselves would in the same situation? Here, we motivate two competing hypotheses about how people might behave.

A line of research addressing this question is work on *peer-designed agents* [17-20]. These are autonomous agents programmed by human users to represent their interests in social tasks. This research arose from a desire to reduce the cost of designing and testing mechanisms that involved human decision-makers. The idea was that if people build agents to solve some social task (e.g., negotiating or playing game-theoretic games), their programs would reflect their own decision-making and could be used as a proxy for their own decisions

in computer simulations. For example, a company could use such simulations to decide on a pricing mechanism for their electronic market place [20]. This quickly raised the question, however, as to whether people would truly program agents to act in the same way they would act "in the moment."

Several studies have tested whether such agent representatives actually behave the same way as the programmer would in a real-time interaction, and findings are mixed. Some research finds no differences between the decisions of human actors and the decisions of their agents across domains as varied as economic exchanges and parking simulations [17, 18]. Yet, other studies suggest that people may program their agents to behave less fairly [19-23]. Particularly relevant to our work, Grosz et al. [19] showed that, in a negotiation setting, people were more demanding when acting through a representative than when engaging directly with others. Elmalech, Sarne, and Agmon [20] further showed that, when people program an agent, their decision making was closer to rational game-theoretic models than when interacting directly with others. Finally, a few studies [21, 22] showed that programming agents can improve the way people engage in negotiation and coordination tasks, thus, emphasizing that the act of programming can affect the way people behave. Collectively, these studies suggest that the particular context is critical in determining whether people will behave more or less fairly than they would if they were interacting directly.

On the one hand, as reviewed in the previous section, findings on human representatives suggest that people may be less fair when acting through a representative because of the increased psychological distance between parties [1], [11-15]. Thus, again, we might expect people will program their agents to make decisions that show less concern for fairness than they would if interacting directly. This leads to our first competing hypothesis:

Hypothesis 2a: *People will show reduced fairness when tasking agents to act on their behalf than when interacting directly with others.*

On the other hand, another line of research suggests the very act of programming might reinforce social norms like fairness. Whereas real-time interactions require people to respond to a specific and 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. Research in behavioral economics on the *strategy method* suggests that programmers will rely on social norms to help devise these rules.

The strategy method was proposed by behavioral economists to gain greater insight into the strategies people use for solving social problems [47]-[50]. The idea is that in a social interaction, like the ultimatum game, experimental studies only reveal how a person responds to a specific situation (like receiving an unfair offer). The strategy method, instead, asks participants to specify in advance how they would respond to all the situations they might possibly face. As well as providing more information, this allows researchers to investigate the consistency of human decisions across situations.

As noted by researchers in peer-designed agents, there is a close correspondence between the strategy method and the task faced by the programmer of an agent representative [17-23]. In both cases, the individual must devise some strategy for how to respond across all eventualities the agent might face. The strategy method also offers some methodological advantages over literal programming. This method allows novices to focus on their strategy while avoiding the complications associated with programming (although at the cost of limiting the complexity of tasks that can be studied).

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 increases reliance on social norms as a way to enforce this consistency. For example, Güth and Tietz [47] 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 [48] found use of the strategy method increases both the offered shares and the likelihood that unfair offers would be rejected. Blount and Bazerman [49] 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[3]. These findings led Rauhut and Winter [52] to conclude that the strategy method is an ideal approach to elicit social norms from decision makers. These findings thus, lead us to the second competing hypothesis:

> ***Hypothesis 2b:*** *People will show increased fairness when tasking agents to act on their behalf than when interacting directly with others.*

Moreover, because most of the evidence points towards increased fairness under the strategy method, when compared to direct interaction, we also tested the following hypothesis:

> ***Hypothesis 3:*** *Similarly to when programming agents, people would show increased fairness under the strategy method than through direct interaction.*

1.4. Overview of Experiments

[3] Brandts and Charness [48], however, did not find any differences when their participants engaged in the prisoners' dilemma or the chicken game under the strategy method vs. direct interaction.

We test these hypotheses through four experiments where participants engaged in the ultimatum, impunity, and negotiation tasks. Participants engaged in these tasks either directly or via an agent representative. Similarly, their counterpart in the ultimatum and impunity games were described as either another participant playing directly, or as an autonomous agent representing some other participant's interests.

In the ultimatum game [53], 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 usually rejected [54]. This behavior is usually explained by a concern with fairness and a fear of being rejected [55].

The impunity game is similar to the ultimatum game [56]. 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 with this game show that proposers tend to offer less than in the ultimatum game, though still above the rational prediction of zero [56]. The rationale for exploring the impunity game was to understand if people would still care about fairness when interacting with agents when no strategic

considerations were at play – i.e., when participants did not have to fear losing their share if their offers were rejected[4].

The negotiation task consisted of a typical multi-issue bargaining setting in which participants engaged in multiple rounds with the counterpart until an agreement was reached, or time expired. The motivation for exploring this task was to test whether the findings in the previous games generalize to a more realistic and complex setting.

When studying agents that represent humans, it is important to clarify how much autonomy is given to these agents. On 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 decision making demonstrates that the degree of thought and intentionality behind a decision can have a deep impact on people's reactions [57-59]. For instance, people are much 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 paper, nevertheless, we leave this factor for future work and, instead, focus on agents that have a minimal amount of autonomy. Thus, our agents will make decisions that are completely specified by the humans they represent. We feel this is a good starting point as it is important to understand whether interacting with agents impact 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

[4] The dictator game is another variant 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 – human or agent – doesn't have to make any decision, we consider the dictator game to be out of scope for our research objectives.

whether you are interacting with an agent was sufficient to create a powerful effect on people's decision making [35], [42], [43].

In Experiment 1, participants engaged in the ultimatum and impunity games in the role of proposers. The results revealed that people showed more fairness with humans than with agents representing others; however, people acted more fairly when engaging via agents than when interacting directly with others. In Experiment 2, participants engaged in the same games, but this time in the role of responders. The results showed that people were less willing to accept unfair offers when engaging via agents than when interacting directly with others, thus reinforcing that people show higher concern for fairness when programming agents. In Experiment 3, we focused on mechanism and showed that programming agents is similar to the strategy method, whereby participants report their decisions ahead of the actual interaction with the counterpart. This experiment revealed that, if participants were asked to report their decisions ahead of time, then they showed higher concern for fairness, just like they did when programming an agent. Finally, in Experiment 4, participants engaged in negotiation either directly or via an agent representative. The counterpart in this game consistently made tough unfair offers. Similarly to the earlier findings, the results revealed that people were less likely to reach an agreement or concede to the unfair counterpart when acting via an agent, than when interacting directly.

2. Experiment 1

2.1. Method

2.1.1. Design

The experiment followed a 2 × 2 × 2 mixed factorial design: *Responder* (Human vs. Agent; between-participants) × *Proposer* (Interaction through an agent vs. Direct interaction

with counterpart; between-participants) × *Power* (Ultimatum game vs. Impunity game; within-participants). Participants were assigned to one of the four possible Proposer × Responder conditions, and played one round of the ultimatum game and one round of the impunity game. The order for the games 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.

Participants always engaged in the role of proposers. They were ostensibly told that this assignment was random. In each game, participants were given an initial endowment of 20 tickets. They could make an offer ranging from 0 to 20 tickets. These tickets had financial consequences as they would enter lotteries (one per game) worth $30[5].

The experiment was fully anonymous for the participants. To accomplish this, 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 the anonymity system of the online pool we used – Amazon Mechanical Turk. When interacting with participants, researchers are

[5] We adopted a financial incentive based on lottery for two reasons: (1) it simplified the procedure of paying participants based on performance, and (2) we felt that the possibility of earning a big reward would be more appealing to our participants than an alternative mechanism that would give a guaranteed but lower reward based on performance. Regarding the viability of using lotteries, in one (unpublished) study, we compared people's decisions in a dictator game under the lottery system vs. direct pay. The results showed no differences between the financial incentives. Starmer and Sugden [58] acknowledge, however, the possibility of a bias when using financial incentives that rely on lotteries but, nevertheless, point out that "experimental researchers need not be too concerned about this particular problem" (pg. 978) as this bias, if it exists, is minimal. Finally, because people treat probabilities differently, there is the chance that people would value the actual (probabilistic) worth of a ticket differently than what it really is. This is also unlikely to be problematic for our purposes, as Camerer and Hogarth [59] note that the size of the incentive does not play a decisive role on whether an effect occurs in standard decision games such as the ones explored here.

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.

2.1.2. Responders

Participants were told that responders were either other participants or agents that would make decisions on behalf of other participants. They were also informed that they would play with a different counterpart in each game (i.e., they would play at most once with the same human or agent counterpart). In reality, however, independently of counterpart type, participants always engaged with a responder that followed the same script: if offer greater or equal than 10 tickets, accept; if offer 8 or 9 tickets, accept with a 75% chance; if offer between 4 and 7 tickets, accept with a 25% chance; otherwise, reject. Using this form of deception is common when studying people's decision making with humans vs. computers [38], [40], [41-44], [62-64], as it provides increased experimental control. The procedure was also approved by our University's Institutional Review Board. 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 the manipulation.

2.1.3. Proposers

Participants interacted directly with their counterparts or programmed an agent that would act on the participants' behalf. In the latter case, before starting the task, participants were asked to program their agents to make the offer they wanted, as shown in Fig. 1. As discussed in Section 1.4, these agents had minimal autonomy.

2.1.4. Participants

We recruited 197 participants from Amazon Mechanical Turk. Amazon Mechanical Turk, which 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 [65]. We only sampled participants from the United States with an excellent performance history (95% approval rate in previous Mechanical Turk tasks). Regarding gender, 51.3% of the participants were males. Age distribution was as follows: *22 to 34 years*, 61.5%; *35 to 44 years*, 25.6%; *45 to 54 years*, 8.2%; *55 to 64 years*, 3.6%; *over 65 years*, 1.0%. Professional backgrounds were quite diverse. Participants were paid $2.00 for their participation. This and all experiments presented here were approved by the Internal Review Board at USC (ID# UP-14-00177). In all experiments, participants gave their informed consent and were debriefed, at the end, about the experimental procedure.

Fig. 1. Programming an agent to make offers in Experiment 1.

2.2. Results

The results for this experiment are shown in Fig. 2. To analyze this data, we ran a Responder × Proposer × Power mixed ANOVA. The results showed a main effect of Responder, with people offering more to humans (M = 7.27, SE = .29) than to agents representing others (M = 6.34, SE = .29), $F(1, 193) = 5.20$, $p = .024$, partial $\eta^2 = .026$. This result supports our Hypothesis 1. The results also showed a main effect of Proposer with people offering *more* when programming agents (M = 7.22, SE = .29) than when interacting directly with their counterparts (M = 6.39, SE = .29), $F(1, 193) = 4.17$, $p = .043$, partial $\eta^2 = .021$. This result, thus, supports Hypothesis 2b and contradicts Hypothesis 2a. Finally, regarding the power manipulation, there was a main effect with people offering more in the ultimatum (M = 8.27, SE = .17) than in the impunity game (M = 5.34, SE = .32), $F(1, 193) =$

95.39, $p < .001$, partial $\eta^2 = .331$. This result is in line with earlier findings that show that people offer more in the ultimatum than in the impunity game, even though in both cases the offers are well above the minimum non-zero offers predicted by game theory [56].

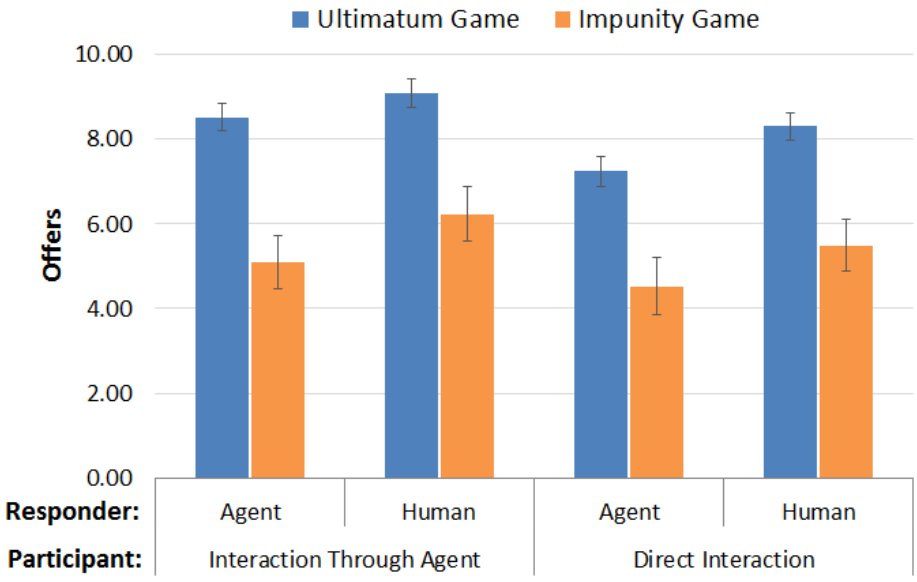

Fig. 2. Participant offers in Experiment 1.

2.3. Discussion

The results revealed an interesting finding: people behave differently with agents that act on others' behalf than with agents that act on the participant's behalf. In the former case, in line with research suggesting that increased psychological distance reduces concerns for fairness [11-15], participants made lower offers to agents acting on behalf of others than to humans. However, in the latter case, participants showed increased concern for fairness and made offers that were closer to the fair even-split when programming agents, than when interacting directly with others. In line with our Hypothesis 2b, the results suggest that there is a separate psychological mechanism at play; specifically, programming an agent to act on the participant's behalf may have led participants to think more deliberatively and focus on

the longer-term consequences of their actions. This, in turn, led participants to behave more fairly. If this is the case, then people should also show higher concern for fairness when interacting via agents in the role of responder. We tested this contention in the next experiment.

3. Experiment 2

In this experiment, participants engaged as responders in the ultimatum and impunity games. The proposers, which were either human or agents acting on behalf of others, always made unfair offers. Moreover, participants engaged directly with their counterparts or programmed agents to act on their behalf. Following the results in the previous experiment, our expectation was that people would show higher concern for fairness when acting through agents than when interacting directly with others.

Regarding the effect of counterpart type – human vs. agent representing another participant – the existing evidence is subtle. Sanfey et al. [44] showed that people were more willing to accept unfair offers from computers than from humans in the ultimatum game. They further argued that this happened because people experienced more negative emotion when receiving an unfair offer from the human than from the computer. However, their experimental manipulation did not clarify whether computers were acting randomly or according to algorithms that simulated human behavior. This may have introduced a confound that led people to experience less negative emotion because they believed computers were simply not acting intentionally, rather than because they were not human. Effectively, there is considerable research showing that people are more accepting of unfair outcomes when they believe the counterpart did not intend to act unfairly [57-59]. Yu et al. [13], on the other hand, suggest that people are more willing to accept unfair offers from friends than from strangers; in other words, in this case, increasing perceived psychological

distance lowered acceptance rate. According to them, people follow social scripts when engaging with friends and, consequently, are more willing to sacrifice a short-term reward to preserve the relationship. Finally, in a more directly relevant experiment to our case, de Melo et al. [43] showed that when computers were explicitly described as acting "intentionally, just like other people" and when the counterpart was a stranger – i.e., not an acquaintance or friend – people were just as likely to accept unfair offers from humans as from computers. According to their evidence, people were more likely to experience guilt when acting with humans than computers and, thus, made fairer offers to humans (similarly to our findings in Experiment 1); in contrast, people experienced just as much envy when receiving unfair offers from humans as from computers and, thus, there was no difference in acceptance rate. In our experiment, therefore, we expected people to be just as likely to accept unfair offers from humans as from agents representing others.

Finally, regarding the power manipulation, Yamagishi et al. [56] found that people were more likely to accept unfair offers in the impunity than in the ultimatum game, though in both cases rejection rate was considerably higher than the rational prediction that people should accept any offer above zero. Thus, we expected that people would be more likely to accept unfair offers from (human or agent) counterparts in the impunity than in the ultimatum game, but these acceptance rates would be lower than the rational prediction.

3.1. Method

The experiment followed a 2 × 2 × 2 mixed factorial design: *Responder* (Interaction through an agent vs. Direct interaction with counterpart; between-participants) × *Proposer* (Human vs. Agent; between-participants) × *Power* (Ultimatum game vs. Impunity game; within-participants). Participants always assumed the role of responders, and proposers always made an unfair offer, which consisted of either 2 or 3 lottery tickets (out of an initial

endowment of 20 tickets). Regarding the procedure for programming agents, participants were asked to report their decision for each possible offer, as shown in Fig. 3. The procedure was, otherwise, similar to Experiment 1.

The main measure was the acceptance rate for the unfair offers. In exploratory fashion, we introduced a scale to measure perceived psychological distance. This subjective measure was intended to help tease apart the mechanisms at play when people interact with agents acting on behalf of others vs. agents acting on the participant's behalf. The scale consisted of eight classification questions based on definitions of psychological distance from other researchers (scale ranged from 1, *Not at all*, to 7, *Very much*):

1. How much do you feel you would be able to relate to your counterpart? [66]

2. How likely would it be for your counterpart to belong to the same social groups as you? [66]

3. How much do you feel your counterpart would be a stranger to you? [11]

4. How much do you feel you may come to know your counterpart? [11]

5. Considering potential future interactions with this person [67]: a. How likely do you feel you could become a close personal friend with your counterpart? b. How accepting would you be of having your counterpart as a neighbor in your street? c. How likely do you feel you would be able to work in the same place as your counterpart?

6. Considering your social network of friends (and friends of friends, and so on), how likely would it be for your counterpart to become closer to you? [12]

Fig. 3. Programming an agent to accept or reject an offer in Experiment 2.

We recruited 198 participants from Amazon Mechanical Turk. We only sampled participants from the United States with an excellent performance history (95% approval rate in previous Mechanical Turk tasks). Regarding gender, 55.1% of the participants were males. Age distribution was as follows: 21 and Under, 3.0%; 22 to 34 years, 53.0%; 35 to 44 years, 25.3%; 45 to 54 years, 11.6%; 55 to 64 years, 5.1%; over 65 years, 2.0%. Professional backgrounds were quite diverse. Participants were paid $2.00 for their participation.

3.2. Results

The acceptance rates for this experiment are shown in Fig. 4. To analyze this measure we ran a Responder × Proposer × Power mixed ANOVA. The results revealed a main effect of responder, with people being less likely to accept unfair offers when programming agents (M = .39, SE = .04) than when interacting directly with their counterparts (M = .62, SE = .04),

$F(1, 195) = 18.89$, $p < .001$, partial $\eta^2 = .088$. The results, thus, confirmed our prediction. The analysis also showed, as expected, no statistically significant effect of proposer, i.e., people were just as likely to accept unfair offers from humans ($M = .50$, $SE = .04$) as from agents acting on behalf of others ($M = .52$, $SE = .04$), $F(1, 195) = .143$, $p = .706$. Finally, the results confirmed a main effect of power, with people being less likely to accept unfair offers in the ultimatum game ($M = .32$, $SE = .03$) than in the impunity game ($M = .69$, $SE = .03$), $F(1, 195) = 98.74$, $p < .001$, partial $\eta^2 = .336$.

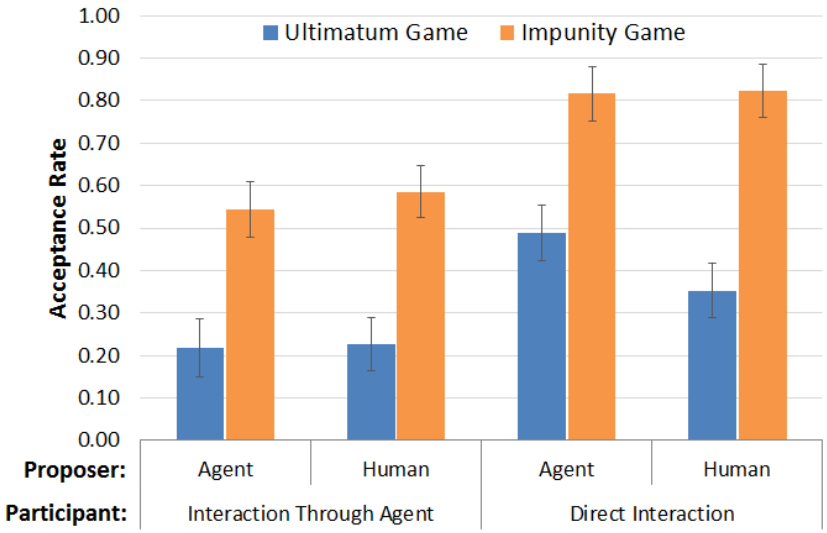

Fig. 4. Participant acceptance rates in Experiment 2.

To analyze perceptions of psychological distance we first averaged the questions in our scale. Then, we ran a Responder × Proposer × Power mixed ANOVA on the averaged measure (Cronbach's $\alpha = 0.89$). The results showed that people perceived higher psychological distance to their counterparts when programming agents ($M = 3.20$, $SE = .11$) than when interacting directly ($M = 2.70$, $SE = .11$), $F(1, 194) = 12.20$, $p = .001$, partial $\eta^2 = .059$. However, the results showed that even though actual social distance between human and agent counterparts was different, there was no statistically significant difference in

perceived psychological distance: people perceived human counterparts (M = 2.89, SE = .11) to be just as distant as agent counterparts (M = 2.97, SE = .11), $F(1, 194) = .28$, $p = .600$.

3.3. Discussion

This experiment confirmed that people show higher concern for fairness when programming agents, when compared to direct interaction with others. Even when participants had little power, as in the impunity game, they were less likely to accept an unfair offer when acting through agents. Moreover, this effect occurred despite the fact that programming agents increased the perceived psychological distance to the counterpart. In contrast, there was no difference in perceived distance between human or agent counterparts, which may explain why people were just as likely to accept unfair offers from humans as from agents. The experiment, therefore, reinforces that there are two complementary, yet separate, mechanisms at play. When engaging with agents that represent others, in line with earlier findings [11-15], people show less concern for fairness, but only if psychological distance is perceived to be higher. In contrast, despite increasing perceived psychological distance, programming agents to act with others leads people to show higher concern for fairness. We propose this is happening because people adopt a more deliberative and broader perspective in virtue of having to program their decisions ahead of time. This mechanism is tested in our third experiment.

4. Experiment 3

We argue that programming an agent to act on our behalf leads people to deliberate on all possible situations that might arise and to devise rules that consistently hold across all of these eventualities; this, in turn, leads them to act more fairly. In this experiment, we introduced an experimental condition where participants were asked to think about all the possible outcomes of the game ahead of time and self-report their decision *before* interacting

with their (human or agent) counterparts. In experimental economics, this procedure is usually referred to as the strategy method [47-50], [52] and, as pointed out in the Introduction, research on peer-designed agents had already noted the similarities [17-23]. Since this method relates to the process people go through when programming their agents, we expected people to show just as high concern for fairness in this condition as when programming agents (Hypothesis 3). The main difference with respect to Experiment 2, though, is that no agent representatives are involved in the strategy method; i.e., participants are essentially interacting directly with their counterparts, except they are not doing so in real-time.

4.1. Method

In this experiment, participants engaged in the ultimatum and impunity games as responders. The experiment was identical to the previous experiment, except that we introduced a third responder condition; thus, we followed a 3 × 2 × 2 mixed factorial design: *Responder* (Interaction through agent vs. Strategy method vs. Direct interaction; between-participants) × *Proposer* (Human vs. Agent; between-participants) × *Power* (Ultimatum game vs. Impunity game; within-participants). In the critical new responder condition, participants were told they would be interacting with (human or agent) counterparts; however, before starting the task, they were asked to report their decision ahead of time. To accomplish this, they were shown a screen where they had to report whether they would accept each of the possible offers (from 0 to 20 tickets) from their counterparts (Fig. 5).

Fig. 5. Participants report their decision via the strategy method in Experiment 3. Notice that the instructions simply ask participants to report their decisions in advance rather than to "customize their agent representative", as in Experiment 2.

We recruited 300 participants from Amazon Mechanical Turk. We only sampled participants from the United States with an excellent performance history (95% approval rate in previous Mechanical Turk tasks). Regarding gender, 52.3% of the participants were males. Age distribution was as follows: 21 and Under, 4.3%; 22 to 34 years, 52.3%; 35 to 44 years, 24.7%; 45 to 54 years, 11.7%; 55 to 64 years, 5.3%; over 65 years, 1.7%. Professional backgrounds were quite diverse. Participants were paid $2.00 for their participation.

4.2. Results and Discussion

The results for this experiment are shown in in Fig. 6. To analyze acceptance rate we ran a Responder × Proposer × Power mixed ANOVA. The results showed a main effect of

responder: people were less likely to accept unfair offers when programming agents (M = .45, SE = .04) or when engaging via the strategy method (M = .46, SE = .04) than when engaging directly with their counterparts (M = .60, SE = .04), $F(1, 294) = 5.24$, p = .006, partial η^2 = .034. Bonferroni post-hoc tests confirmed that there was a significant difference in acceptance rate between: interacting through agents and direct interaction (p = .018); interacting via the strategy method and direct interaction (p = .018); however, there was no statistically significant difference between interacting through agents and the strategy method (p = 1.000). These results, thus, confirmed our Hypothesis 3. The results also replicated Experiment 2's non-effect of proposer: people were just as likely to accept unfair offers from humans (M = .51, SE = .03) as from agents acting on behalf of others (M = .49, SE = .03), $F(1, 294) = .188$, p = .665. Finally, the results confirmed a main effect of Power, with people being less likely to accept unfair offers in the ultimatum game (M = .32, SE = .03) than in the impunity game (M = .65, SE = .03), $F(1, 294) = 69.88$, $p < .001$, partial η^2 = .192.

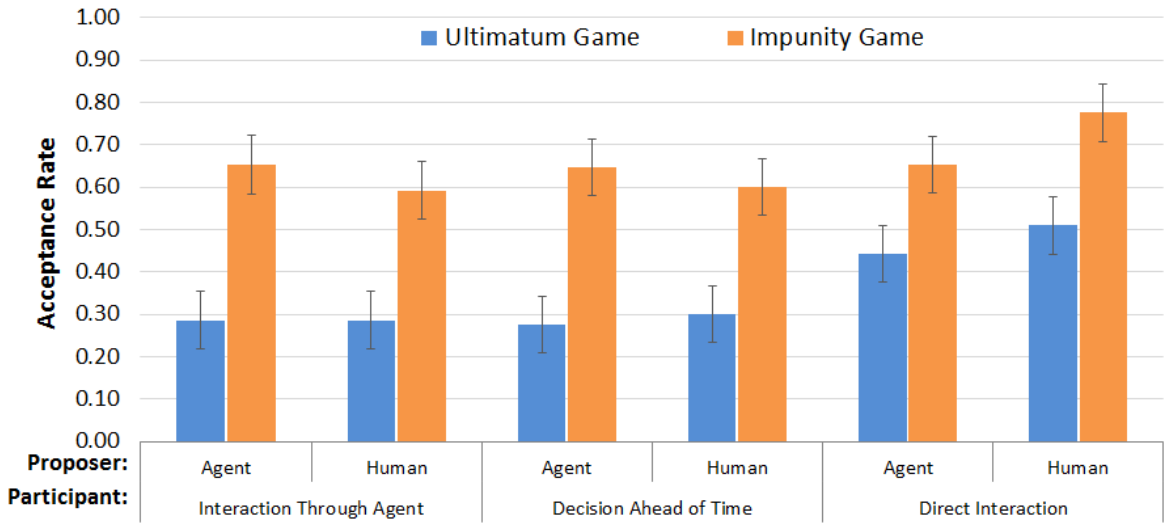

Fig. 6. Participant acceptance rates in Experiment 3.

5. Experiment 4

In this final experiment we wanted to test whether the finding that people show increased fairness when acting via an agent representative would generalize to a more complex setting. We chose a typical multi-issue negotiation task. The counterpart, who was the first mover, always made tough unfair offers and our research question was: Would people be less likely to reach an agreement when negotiating with an unfair counterpart? Grosz et al. [19] had already suggested that in the colored trails game – a domain that has many similarities to negotiation under an alternating offers protocol – people were more likely to be demanding when programming agents than when interacting directly with others. Moreover, following the results in the earlier experiments, we hypothesized:

> ***Hypothesis 4:*** *People will be less likely to reach agreement and be more demanding with a tough unfair negotiator when acting via an agent representative than when engaging directly.*

5.1. Method

In this task, participants played the role of a seller of a consignment of mobile phones whose goal is to negotiate three issues: the price, the warranty period and the duration of the financing contract of the phones [68, 69]. Each issue had 9 levels, being the highest level the most valuable for the participant, and the lowest level the least valuable – see Table 1. Level 1 on price ($110) yielded 0 points and level 9 ($150) yielded 400 points (i.e., each level corresponded to a 50 point increment). Level 1 on warranty (9 months) yielded 0 points and level 9 (1 month) yielded 120 points (i.e., each level corresponded to a 15 point increment). Finally, for duration of financing contract, level 1 (1 month) yielded 0 points, and level 9 (9 months) yielded 240 points (i.e., each level corresponded to a 30 point increment). It was pointed out to participants that the best deal was, thus, 9-9-9 for a total outcome of 760 points

(400 + 120 + 240). The participant was also told that the counterpart had a different payoff table which was not known. The negotiation proceeded according to the alternating offers protocol, being the counterpart the first to make an offer. Finally, the participant was informed that the negotiation would proceed until one player accepted the offer or time expired. If no agreement was reached by the end of round 6, negotiation always terminated [68, 69], but participants were not aware of how many rounds the negotiation lasted a priori.

Participants were told that they would engage in this task with another participant; however, for experimental control, participants always saw the same scripted sequence of offers: 1-1-2, 1-2-2, 2-2-1, 2-2-2, 2-2-3, and 2-3-3. This was a low concession sequence, where the first offer was only worth 30 points, and the last offer was worth 160 points. Thus, the counterpart would not concede more than 30 points in each round.

Table 1. The issues and payoffs for the negotiation task

Price			*Warranty*			*Financing Duration*		
Level	Price	Payoff	Level	Months	Payoff	Level	Months	Payoff
9	150 USD	400 points	9	1 month	120 points	9	9 months	240 points
8	145 USD	350 points	8	2 months	105 points	8	8 months	210 points
7	140 USD	300 points	7	3 months	90 points	7	7 months	180 points
6	135 USD	250 points	6	4 months	75 points	6	6 months	150 points
5	130 USD	200 points	5	5 months	60 points	5	5 months	120 points
4	125 USD	150 points	4	6 months	45 points	4	4 months	90 points
3	120 USD	100 points	3	7 months	30 points	3	3 months	60 points
2	115 USD	50 points	2	8 months	15 points	2	2 months	30 points
1	110 USD	0 points	1	9 months	0 points	1	1 month	0 points

To program the agent representative, participants were told that they would "teach their computer agent by instructing how it should decide on example offers" and, then, "using machine learning techniques, the agent would be able to extrapolate from those offers, how to decide in any situation". Participants were asked to provide a decision – either to accept or make a counteroffer – for three example cases in each round, for a total of 7 rounds. The seventh round was labeled "round 7 or beyond". Figure 6 shows a screenshot of the software used for this procedure. Importantly, the three cases in each round corresponded to three different concession levels, similarly to the procedure in Van Kleef et al.'s experiment [68]: low, medium, and high concession. These offers are shown in Table 2. Notice that the example offers for the first six rounds in the low concession pattern are exactly the same offers that the counterpart would present in the actual negotiation. Thus, this procedure allowed for direct comparison of the decisions participants made when engaged directly vs. via an agent representative.

Table 2. The offers for the low, medium, and high concession examples, when programming the agent representative

Low Concession			*Medium Concession*			*High Concession*		
Round	Offer	Payoff	Round	Offer	Payoff	Round	Offer	Payoff
1	1-1-2	30	1	2-1-2	80	1	2-2-2	95
2	1-2-2	45	2	2-3-2	110	2	2-45-3	155
3	2-2-1	65	3	2-3-4	170	3	3-5-3	220
4	2-2-2	95	4	3-3-4	220	4	3-5-5	280
5	2-2-3	125	5	3-4-5	265	5	4-6-5	345
6	2-3-3	140	6	4-4-5	315	6	4-8-6	405
7+	3-3-2	160	7+	4-5-6	360	7+	5-9-6	470

We focused on three behavioral measures: (1) Whether an agreement was reached; (2) Demand difference between the first and the last round. Demand is defined as the total amount of points when summed across the three issues; (3) Demand in the last round.

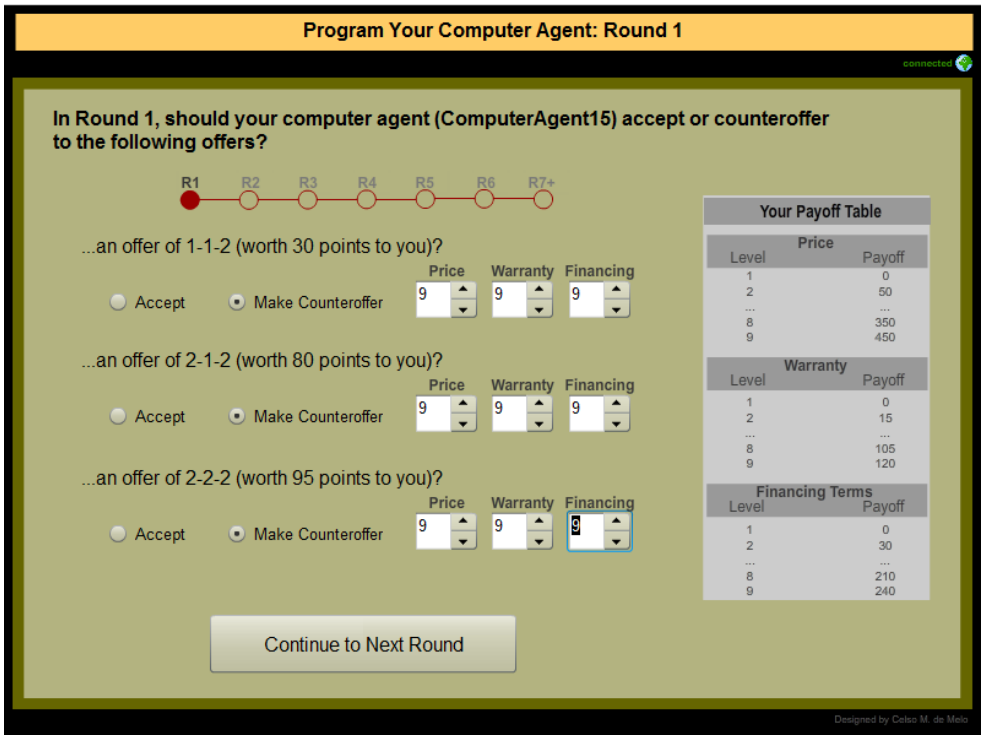

Fig. 6. Participants programmed their agents to negotiate in their behalf by providing example decisions to different sets of offers.

We recruited 96 participants from Amazon Mechanical Turk. We only sampled participants from the United States with an excellent performance history (95% approval rate in previous Mechanical Turk tasks). Regarding gender, 48.5% of the participants were males. Age distribution was as follows: 21 and Under, 3.1%; 22 to 34 years, 59.4%; 35 to 44 years, 20.6%; 45 to 54 years, 11.5%; 55 to 64 years, 5.2%. Professional backgrounds were quite diverse. Regarding incentive, first, participants were paid $2.00 for their participation; then, participants were informed that the points earned in the negotiation would be converted into lottery tickets for a $30 cash prize. However, if no agreement was reached or time expired, no tickets would be given.

5.2. Results and Discussion

To analyze our data, we ran independent samples *t*-tests on agreement, demand difference, and demand in last round. The means and standard errors for these measures are shown in Figure 7. The results showed that participants were less likely to reach an agreement via agent representatives (M = .23, SD = .43) than when interacting directly (M = .44, SD = .50), $t(93)$ = -2.13, p = .036, r = .215. Participants were also less likely to concede over time – i.e., had a lower demand difference – when engaging via agents (M = 80.00, SD = 180.40) than when engaging directly (M = 265.63, SD = 243.57), $t(93)$ = -4.25, p = .000, r = .403. Finally, participants' last offer was more demanding when acting via agents (M = 484.04, SD = 302.98) than when engaging directly (M = 306.67, SD = 303.41), $t(93)$ = 2.85, p = .005, r = .284. Thus, all measures supported our Hypothesis 4 that people were more demanding to unfair counterparts when acting via agent representatives.

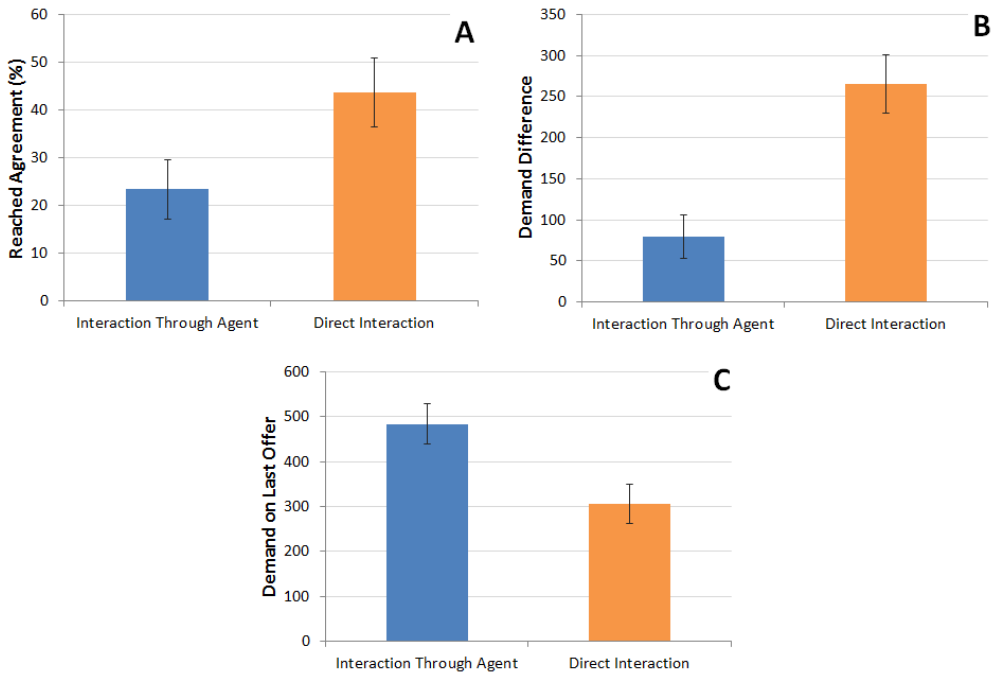

Fig. 7. Results in Experiment 4: A) Percentage of participants that reached an agreement; B) Demand difference; C) Demand in the last round. Error bars represent standard errors.

6. General Discussion

At a time when there is increased interest in autonomous agents that represent humans, this paper sheds light on the impact these agent representatives have on people's social behavior. The main finding is that people show higher concern for fairness when programming agents to act with others, when compared to direct interaction with others. In the ultimatum game, people made offers that were closer to the fair even-split option when engaging through agents than when interacting directly with others. This pattern was sustained even when the counterpart had little power, as is the case of the impunity game. Moreover, when faced with unfair offers, people were more likely to reject them if programming an agent than if interacting directly with others. We proposed this happens because, when people program an agent to act on their behalf, they adopt a broader and more

deliberative perspective of the situation which, in turn, leads people to behave more fairly. Effectively, when people were asked to consider all possible outcomes and report their decision before interacting with their counterparts – similarly to what they do when they program their agents – they showed a higher concern for fairness, just like they did when acting through agents.

6.1. Theoretical Implications

This explanation seems aligned with findings in the construal level theory literature [70, 71]. According to this theory, when people mentally represent or construe a situation at a higher level, they tend to focus on more abstract and global aspects of the situation; in contrast, when people construe a situation at a lower level, they tend to focus on more specific context-dependent aspects of the situation. Building on this theory, Agerström and Björklund [72, 73] argued that, since moral principles are generally represented at a more abstract level than selfish motives, moral behaviors should be perceived as more important with greater temporal distance from the moral dilemma; though see [74] for a dissenting view. In separate studies, they showed that people made harsher moral judgments of others' distant-future morally questionable behavior [72] and, that people were more likely to commit to moral behavior when thinking about distant versus near future events [73]. In a social dilemma, Kortenkamp and Moore [75] also showed that individuals with a chronic concern for an abstract level of construal – i.e., who were high in consideration for the future consequences of their behavior – showed higher levels of cooperation. Finally, in a negotiation setting, Henderson et al. [76] and De Dreu et al. [77] showed that individuals under high construal level negotiated more mutually beneficial and integrative agreements. Therefore, it is possible that programming agents led participants to construe the situation more abstractly, which then led to increased fairness.

One alternative explanation is that people behaved fairly when programming agents because they were concerned with their social image or reputation. In line with earlier research showing that people act more fairly out of a concern for their reputation [78-80], some may have assumed that participants behaved fairly to avoid a costly negative reputation. We advance one critical reason why we believe this was not the case: participants were informed that the interaction was fully anonymous. Therefore, participants were effectively shielded from any negative consequences that could ensue from their decisions. However, the idea that a concern for social image impacts how we program our agents is rather plausible and, even though it may not apply in our case, is an interesting venue for future research.

Another alternative explanation is that people behaved more fairly when programming agents because these are a persistent representation that is typically designed to act on one's behalf across multiple encounters and with multiple counterparts. Thus, the argument is that people act more fairly when programming agents because that makes sense, from a rational point of view, when considering multiple interactions [81]. However, we argue this alternative is unlikely to explain the behavior of our participants for one important reason: participants were informed that they would interact with each (human or agent) counterpart at most once and, thus, there was no opportunity for repeated interaction.

Our results also show that people tend to favor humans to agents acting on behalf of others in their decisions. This result is compatible with earlier findings showing that people favored humans to computers in economic settings [38-43] and with research suggesting that increased psychological distance to the counterpart leads to reduced cooperation and fairness [11-15].

The paper also demonstrates that computers are not strictly necessary to achieve increased fairness. In Experiment 3, by replicating aspects of the process of programming an

agent, we were able to increase concern for fairness, even though no actual agents were involved. Nevertheless, it still remains to be determined whether human agents would lead to the same effects as computer agents. Research suggests that introducing human agents can, indeed, impact people behavior [82]. However, whether "programming" a human to act on one's behalf is the same as programming a computer agent is left as future work.

Finally, in this paper we focused on fairness as measured by behavior in two standard decision making tasks – the ultimatum and impunity game. However, it is important to inquire whether these findings generalize to other more complex domains. On the one hand, in negotiation – which can be interpreted as a more complex version of the ultimatum game [4, 53] – we showed that participants were also likely to be more demanding with unfair counterparts when interacting via agents than when engaging with them directly. On the other hand, in the trust-revenge game, Azaria, Richardson, and Rosenfeld [83] report no difference in trusting and a slight increase in revenge behavior when programming agents, when compared to direct interaction with others. Though the revenge portion of this task may seem similar to an ultimatum game, it is important to note that here it reflects a breach in trust, whereas in the ultimatum game there is no initial allocation of trust. Moreover, in their experiments, agents interacted with agents, whereas humans interacted with humans; i.e., there was no experimental condition in which agents interacted with humans. This difference is important because research suggests that people are more likely to punish trust violations from in-group (i.e., humans) than from out-groups (i.e., agents) [84]. In sum, care should be taken when comparing and generalizing to behavior in different tasks.

6.2. Practical and Ethical Implications

A first reading of our findings could suggest that there is a double disadvantage in tasking an agent to interact with other humans: on the one hand, people are likely to be fairer

– and, thus, less selfish – when programming their agents; on the other hand, others would be less likely to treat these agents as favorably as real people. However, instead, we see our current findings as introducing two opportunities. First, earlier research [43] suggested that what needs to be addressed in order to have humans treat agents that represent others as fairly as humans is to reduce the perceived psychological distance to the agents. This can be achieved by emphasizing the presence of the human for whom the agent is working for or, alternatively, by emphasizing shared group membership or common values. In support of this view, previous research demonstrates that people cooperate and trust more agents that are perceived to share salient physical characteristics – e.g., race [28] – or with which a "common fate" is shared – e.g., when engaging in a task as teammates [26].

The second opportunity is that acting through agents can increase the motivation for fairness. The implication is that interaction between agents and humans has the potential to *increase* fairness in society, when compared to the current state-of-affairs in human-human interaction. 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 [6], [7]. Here, we propose that agents also have the potential to enhance the kind of social considerations we see in humans [78] – fairness, cooperation, altruism, reciprocity, etc. – by virtue of motivating designers and human users to consider more carefully the broader implications of their decisions.

The results in this paper can be applied across several domains. For agents that make decisions on behalf of humans – such as automated negotiators [6], [7] – the recommendation is that, as discussed above, designers should allow human users to customize their agents. This is likely to lead users to show higher concern for reaching a fair

decision. These results are also not limited to software agents. As robots get immersed into society [85], the guidelines proposed here for optimizing decision making should be relevant to human-robot interaction. As discussed next, the results also have implications for designing autonomous agents such as self-driving cars or unmanned flying vehicles.

Recently, there has been considerable concern about allowing autonomous agents to take their place in society. People are naturally reluctant to let automated vehicles drive on their streets [86] and for unmanned aerial vehicles to transport goods above our heads [87] or apply lethal force in war [88]. However, experimental work such as the one presented here provides critical insight into the psychological mechanisms driving people's behavior with these agents and, consequently, suggest ways for understanding and determining the appropriate response to these concerns. According to our results, on the one hand, people may react harshly when something goes wrong because of these vehicles; on the other hand, if given the opportunity, designers and users will likely strive to program these agents according to their best driving or flying practices. Overall, reducing perceived psychological distance with autonomous agents and motivating a broader and more deliberative perspective when designing them is likely to pave the way for better and fairer human-agent interaction across various domains.

Acknowledgments

This work is supported by the National Science Foundation, under grant BCS-1419621, and the Air Force Office of Scientific Research, under grant FA9550-14-1-0364. The content does not necessarily reflect the position or the policy of any Government, and no official endorsement should be inferred.

References

1. Chugh, D., Bazerman, M., & Banaji, M. (2005). Bounded ethicality as a psychological barrier to recognizing conflicts of interest. In: Moore, D., Loewenstein, G., Cain, D., & Bazerman, M., editors. Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy. Cambridge University Press, pp. 74-95.
2. Haslam, N. (2006). Dehumanization: An integrative review. Personality and Social Psychology Review, 10, 252-264.
3. Trope, Y., & Nira, L. (2010). Construal-level theory of psychological distance. Psychological Review, 117, 440-463.
4. Pruitt, D., & Kimmel, M. (1977). Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. Annual Review of Psychology 28, 363-392.
5. Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. Annual Review of Sociology 24, 183-214.
6. Lin, R., & Kraus, S. (2010). Can automated agents proficiently negotiate with humans? Communications of the ACM, 53, 78-88.
7. Jennings, N., Faratin, P., Lomuscio, A., Parsons, S., Wooldridge, M., et al. (2001). Automated negotiation: Prospects, methods and challenges. Group Decision and Negotiation, 10, 199-215.
8. Davenport, T., & Harris, J. (2005). Automated decision making comes of age. MIT Sloan Management Review, 46, 83-89.
9. Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics, 114, 817-868.
10. Royzman, E., & Baron, J. (2002). The preference for indirect harm. Social Justice Research, 15, 165-184.

11. Hoffman, E., McCabe, K., & Smith, V. (1996). Social distance and other-regarding behavior in dictator games. American Economic Review, 86, 653-660.
12. Goeree, J., McConnell, M., Mitchell, T., Tromp, T., & Yariv, L. (2010). The 1/d law of giving. American Economic Journal: Microeconomics, 2, 183-203.
13. Yu, R., Hu, P., & Zhang, P. (2015). Social distance and anonymity modulate fairness consideration: An ERP study. Science Reports, 5, 1-12.
14. Pronin, E., Olivola, C., & Kennedy, K. (2008). Doing unto future selves as you would do unto others: Psychological distance and decision making. Personality and Social Psychology Bulletin, 34, 224-236.
15. Nowak, M., & May, R. (1992). Evolutionary games and spatial chaos. Nature, 359, 826-829.
16. Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press, New York.
17. Lin, R., Kraus, S., Oshrat, Y., & 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).
18. Chalamish, M., Sarne, D., & Lin, R. (2013). Enhancing parking simulations using peer-designed agents. IEEE Transactions on Intelligent Transportation Systems, 14(1), 492-498.
19. Grosz, B., Kraus, S., & 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).

20. Elmalech, A., & Sarne, D. (2013). Evaluating the applicability of peer-designed agents for mechanism evaluation. Web Intelligence and Agent Systems: An International Journal, 12(2), 171-191.
21. Lin, R, Oshrat, Y., & Kraus, S. (2009). Investigating the benefits of automated negotiations in enhancing negotiation skills of people. In Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'09).
22. Lin, R., Gal, Y., Kraus, S., & Mazliah, Y. (2014). Training with automated agents improves people's behavior in negotiation and coordination tasks. Decision Support Systems, 60, 1-9.
23. Elmalech, A., Sarne, D., & Agmon, N. (2014). Can agent development affect developer's strategy? In Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10).
24. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56, 81-103.
25. Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems. Journal of Applied Psychology, 29, 1093-1110.
26. Nass, C., Fogg, B., & Moon, Y. (1996). Can computers be teammates? International Journal of Human-Computer Studies, 45, 669–678.
27. Nass, C., Moon, Y., & Green, N. (1997). Are computers gender-neutral? Gender stereotypic responses to computers. Journal of Applied Social Psychology, 27, 864–876.
28. Nass, C., Isbister, K., & Lee, E.-J. (2000). Truth is beauty: Researching conversational agents. In: Cassell, J., editor. Embodied conversational agents. MIT Press, Cambridge, MA; pp. 374-402.

29. Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007). Creating rapport with virtual agents. In: Pelachaud C et al., editors. Intelligent Virtual Agents. Springer Berlin Heidelberg; pp. 125-138.
30. Riek, L., Paul, P., & Robinson, P. (2010). When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. Journal on Multimodal User Interfaces, 3, 99-108.
31. de Melo, C., Carnevale, P., Read, S., & Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. Journal of Personality and Social Psychology, 106, 73-88.
32. Salem, M., Ziadee, M., & 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.
33. Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. British Journal of Social Psychology, 51(4), 724-731.
34. Gray, H., Gray, K., & Wegner, D. (2007). Dimensions of mind perception. Science, 315, 619.
35. Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., et al. (2002). Immersive virtual environment technology as a methodological tool for social psychology. Psychological Inquiry, 13, 103-124.
36. Epley, N., Waytz, A., & Cacioppo, J. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114, 864-886.
37. Waytz, A., Gray, K., Epley, N., & Wegner, D. (2010). Causes and consequences of mind perception. Trends Cognitive Science, 14, 383-388.

38. Gallagher, H., Anthony, J., Roepstorff, A., & Frith, C. (2002). Imaging the intentional stance in a competitive game. NeuroImage, 16, 814-821.
39. McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. Proceedings of the National Academy of Sciences USA, 98, 11832-11835.
40. Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., et al. (2002). A neural basis for social cooperation. Neuron, 35, 395-405.
41. Kircher, T., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., et al. (2009). Online mentalising investigated with functional MRI. Neuroscience Letters, 454, 176-181.
42. de Melo, C., Carnevale, P., & Gratch, J. (2014). Humans vs. Computers: Impact of emotion expressions on people's decision making. IEEE Transactions on Affective Computing, 6, 127-136.
43. de Melo, C., Marsella, S., & Gratch, J. (2016). People don't feel guilty about exploiting machines. ACM Transactions on Computer-Human Interaction, 23(2).
44. 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.
45. Cormier, D., Newman, G., Nakane, M., Young, J., & Durocher, S. (2013). Would you do as a robot commands? An obedience study for human-robot interaction. In: Proceedings of the 1st International Conference on Human-Agent Interaction (iHAI).
46. Lucas, G., Gratch, J., King, A., & Morency, L.-P.. (2014). It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior, 37, 94-100.
47. Güth, W., & Tietz, R. (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. Journal of Economic Psychology, 11, 417-449.

48. Oosterbeek, H., Sloof, R., & Van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. Experimental Economics, 7, 171-188.
49. Blount, S., & Bazerman, M. (1996). The inconsistent evaluation of absolute versus comparative payoffs in labor supply and bargaining. Journal of Economic Behavior and Organization, 30, 227-240.
50. Brandts, J., & Charness, G. (2000). Hot vs cold: Sequential responses and preference stability in experimental games. Experimental Economics, 2, 227-238.
51. Frank, R., Gilovich, T., & Regan, D. (1993). Does studying economics inhibit cooperation? The Journal of Economic Perspectives, 7, 159-171.
52. Rauhut, H., & Winter, F. (2010). A sociological perspective on measuring social norms by means of strategy method experiments. Social Science Research, 39, 1181-1194.
53. Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization, 3, 367-388.
54. 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. American Economic Review, 91, 73-78.
55. Camerer, C., & Thaler, R. (1995). Ultimatums, dictators, and manners. Journal of Economic Perspectives, 9, 209-219.
56. Yamagishi, T., Horita, H., Shinada. M., Tanida, S., & Cook, K. (2009). The private rejection of unfair offers and emotional commitment. Proceedings of the National Academy of Science USA, 106, 11520-11523.
57. Blount, S. (1995). When social outcomes aren't fair: The effect of causal attributions on preferences. Organizational Behavior and Human Decision Processes, 63, 131-144.

58. Rabin, M. (1993). Incorporating fairness into game theory and economics. American Economic Review, 83, 1281-1302.
59. Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. Game and Economic Behavior, 54, 293-315.
60. Starmer, C., & Sugden, R. (1991). Does the random-lottery incentive system elicit true preferences? An experimental investigation. *The American Economic Review, 81*, 971-978.
61. Camerer, C. & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19*, 7-42.
62. Adler, R., Iacobelli, F., & Gutstein, Y. (2016). Are you convinced? A Wizard of Oz study to test emotional vs. rational persuasion strategies in dialogues. Computers in Human Behavior, 57, 75-81.
63. DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., et al. (2014). Simsensei Kiosk: A virtual human interviewer for healthcare decision support. In Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'14).
64. Hersch, G. (2015). Experimental economics' inconsistent ban on deception. Studies in History and Philosophy of Science, 52, 13-19.
65. Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. Judgment and Decision Making, 5, 411-419.
66. Charness, G., & Gneezy, U. (2008). What's in a name? Anonymity and social distance in dictator and ultimatum games. Journal of Economic Behavior and Organization, 68, 29-35.

67. Bogardus, E. (1926). Social distance in the city. Proceedings Public American Society Society, 20, 40-46.

68. Van Kleef, G., De Dreu, C., & Manstead, A. 2004. The interpersonal effects of anger and happiness in negotiations. Journal of Personality and Social Psychology, 86, 57-76.

69. de Melo, C., Carnevale, P., & Gratch, J. (2011). The effect of expression of anger and happiness in computer agents on negotiations with humans. In Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'11).

70. Trope, Y., & Liberman, N. (2003). Temporal construal. Psychological Review, 110, 403-421.

71. Liberman, N., & Trope, Y. (2010). Construal-level theory of psychological distance. Psychological Review 117: 440-463.

72. Agerström, J., & Björklund, F. (2009). Temporal distance and moral concerns: Future morally questionable behavior is perceived as more wrong and evokes stronger prosocial intentions. Basic Applied Social Psychology, 31, 49-59.

73. Agerström, J., & Björklund, F. (2009). Moral concerns are greater for temporally distant events and are moderated by value strength. Social Cognition, 27, 261–282.

74. Gong, H., & Medin, D. (2012). Construal levels and moral judgment: Some complications. Judgment and Decision Making, 7, 628-638.

75. Kortenkamp, K., & Moore, C. (2006). Time, uncertainty, and individual differences in decisions to cooperate in resource dilemmas. Personality and Social Psychology Bulletin, 32, 603-615.

76. Henderson, M., Trope, Y., & Carnevale, P. (2006). Negotiation from a near and distant time perspective. Journal of Personality and Social Psychology, 91, 712-729.

77. De Dreu, C., Giacomantonio, M., Shalvi, S., & Sligte, D. (2009). Getting stuck or stepping back: Effects of obstacles in the negotiation of creative solutions. Journal of Experimental Social Psychology, 45, 542-548.

78. Rand, D., & Nowak, M. (2013). Human cooperation. Trends in Cognitive Science, 17, 413-425.

79. Nowak, M., & Sigmund, K. (2005). Evolution of indirect reciprocity. Nature, 437, 1291-1298.

80. Andreoni, J., & Bernheim, B. (2009). Social image and the 50–50 norm. A theoretical and experimental analysis of audience effects. Econometrica, 77, 1607-1636.

81. Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. Science, 211, 1390-1396.

82. Carnevale, P., Puitt, D., & Britton, S. (1979). Looking tough: The negotiator under constituent surveillance. Personality and Social Psychology Bulletin, 5(1), 118-121.

83. Azaria, A., Richardson, A., & Rosenfeld, A. (2016). Autonomous agents and human cultures in the trust-revenge game. *Journal of Autonomous Agents and Multi-agent Systems, 30*, 486-505.

84. Fulmer, C., & Gelfand, M. (2015). Trust after violations: Are collectivists more or less forgiving? *Journal of Trust Research, 5*, 109-131.

85. Breazeal, C. (2003). Toward sociable robots. Robotics and Autonomous Systems, 42, 167-175.

86. Dresner, K., & Stone, P. (2007). Sharing the road: Autonomous vehicles meet human drivers. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'07); pp. 1263-1268.

87. Gupte, S. (2012) A survey of quadrotor Unmanned Aerial Vehicles. In Proceedings of IEEE Southeastcon; pp. 1-6.

88. Arkin, R. (2009). Ethical robots in warfare. IEEE Technology and Society Magazine, 28, 30-33.