People Show Envy, Not Guilt, when Making Decisions with Machines

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

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

***Abstract*— Research shows that people consistently reach more efficient solutions than those predicted by standard economic models, which assume people are selfish. Artificial intelligence, in turn, seeks to create machines that can achieve these levels of efficiency in human-machine interaction. However, as reinforced in this paper, people's decisions are systematically less efficient – i.e., less fair and favorable – with machines than with humans. To understand the cause of this bias, we resort to a well-known experimental economics model: Fehr and Schmidt's inequity aversion model. This model accounts for people's aversion to disadvantageous outcome inequality (envy) and aversion to advantageous outcome inequality (guilt). We present an experiment where participants engaged in the ultimatum and dictator games with human or machine counterparts. By fitting this data to Fehr and Schmidt's model, we show that people acted as if they were just as envious of humans as of machines; but, in contrast, people showed less guilt when making unfavorable decisions to machines. This result, thus, provides critical insight into this bias people show, in economic settings, in favor of humans. We discuss implications for the design of machines that engage in social decision making with humans.**

Keywords*— Guilt; Decision Making; Bias; Human vs. Machines

I. Introduction

The last few decades have seen growing interest in the development of artificially intelligent (software and robotic) agents that can engage efficiently in decision making with humans [1]-[3]. Here, "efficiency" stands for economic outcomes that are more favorable to all parties, when compared to the rational predictions of standard economic models. However, despite evidence that machines can be treated similarly to humans [4], [5], recent research suggests that people show systematic differences in patterns of brain activation and behavior when engaging in decision making with machines, when compared to humans [6]-[10]. Building on these findings, we present further evidence that, in economic settings, people favor humans to machines. This is an important challenge for artificial intelligence because it shows a fundamental bias that needs to be overcome if we hope to achieve in human-machine interaction the kind of efficiency we see in human-human interaction.

To gather insight on this bias we resort to a well-known experimental economics theory: Fehr and Schmidt's inequity aversion theory [11]. The advantage of this theory is that, in contrast to standard economic theory, it takes into consideration people's other-regarding preferences, such as altruism, fairness, and reciprocity. These are the kinds of social considerations that often lead to higher efficiency than what is predicted by standard models [12], [13]. Specifically, the model accomplishes this by accounting for people's aversion to disadvantageous outcome inequality, or *envy*, and aversion to advantageous outcome inequality, or *guilt*. We present an experiment where participants engaged in standard economic games with human or machine counterparts. By fitting this experimental data to Fehr and Schmidt's model we show that, when the outcome was unequal in favor of others, people were not more envious of humans than machines; in contrast, when the outcome was unequal in their favor, people showed less guilt with machines than humans. To the best of our knowledge, this is the first time clear evidence is shown that increased guilt is important in explaining people's bias in favor of humans, when compared to machines.

A. Fairness, Altruism, and Reciprocity

Evidence for these other-regarding preferences has been shown in simple games where the selfish option is obvious, yet people still choose differently. Two games that are particularly relevant for this paper are the ultimatum and dictator games. In the ultimatum game [14], 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 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 [15]. A concern for fairness is usually argued to explain this behavior.

This research was supported in part by grants NSF IIS-1211064, SES-0836004, and AFOSR FA9550-09-1-0507. The content does not necessarily reflect the position or the policy of any Government, and no official endorsement should be inferred..

The dictator game [16] is similar to the ultimatum game, except that the responder doesn't get to make a decision and must accept whatever is offered by the proposer. This game, therefore, removes strategic considerations from the equation, i.e., the proposer doesn't have to fear that the offer will be rejected. Accordingly, the standard prediction is for the proposer to keep all the money and offer nothing. Nevertheless, even in this case, people offer an average of 10 to 25 percent of the initial endowment and the modal offer is 50 percent. Decisions in this game, thus, have been argued to reflect altruism.

Experimental economists have advanced theories that attempt to account both for selfish behavior (e.g., in competitive markets [17]) and the kind of deviation from selfish behavior discussed above. Three kinds of models have been proposed: (a) models of social preferences, which assume that the players' utility function not only depend on one's own payoff, but the others' payoffs as well [11], [18]; (b) models that account for the others' "type", i.e., people will adjust their behavior according to whether the other is a selfish type or a (conditionally) altruistic type [19]; finally, (c) models of intention-based reciprocity, in which case people will consider the intention behind others' actions. Thus, for instance, one is more likely to accept an unfair offer if the other did not have a choice (or did not intend to make the offer), than in the case where the other intended to act unkindly [20], [21]. In this paper, we focus on one model of social preferences that has received considerable attention and can capture behavior in many different economic settings: Fehr and Schmidt's inequity aversion model [11]. We discuss alternative models in the last section of the paper.

Fehr and Schmidt assume that people are inequity averse, i.e., people dislike outcomes where others are better off than them (envy) or outcomes where they are better off than others (guilt). To model this, they propose the following utility function for dyadic interactions:

$$U_k(x_k, x_l) = x_k - \alpha_k \max\{x_l - x_k, 0\} - \beta_k \max\{x_k - x_l, 0\} \qquad (1)$$

with $0 \leq \beta_i \leq \alpha_i$ and $\beta_i \leq 1$ and where x_i, i □ $\{k, l\}$, are the payoffs for the players. Notice that the parameter α captures aversion to disadvantageous inequality (envy), whereas the parameter β captures aversion to advantageous inequality (guilt). This simple utility function has successfully modeled people's behavior in many economic settings, including the ultimatum and dictator games [12]. Thus, people's decisions with other humans can be succinctly described by their α and β parameters. However, what is less clear, and is tested in this paper, is whether people have the same α and β parameters with machines as they do with humans.

B. *Humans vs. Machines*

Early work by Nass and colleagues showed that people are capable of treating machines in a social manner, just like they would treat humans in similar situations [5]. For instance, Nass, Fogg, and Moon [22] showed that machines that were perceived to be teammates were rated more positively than machines that were not. However, despite having the ability to treat machines like social actors, recent research suggests that people still make important distinctions between machines and humans. Specifically, this evidence reveals that people can reach different decisions and show different patterns of brain activation with machines in the exact same decision making tasks, for the exact same financial incentives, when compared to humans [6]-[10]. For instance, Gallagher, Anthony, Roepstorff, and Frith [6] showed that when people played the rock-paper-scissors game with a human there was activation of the medial prefrontal cortex, a region of the brain that had previously been implicated in mentalizing (i.e., inferring of other's beliefs, desires and intentions); however, no such activation occurred when people engaged with a machine that followed a known predefined algorithm. In another study, Sanfey, Rilling, Aronson, Nystrom, and Cohen [10] showed that, when receiving unfair offers in the ultimatum game, people showed stronger activation of the bilateral anterior insula – a region associated with the experience of negative emotions – when engaging with humans, when compared to machines. This evidence, thus, suggests that people experienced less emotion and spent less effort inferring mental states with machines than with humans.

These findings are compatible with research that shows that people perceive less mind in machines than in humans [23]. Denying mind to others or perceiving inferior mental ability in others, in turn, is known to lead to discrimination [24]. In the context of human-machine interaction, Blascovich et al. [4] also propose that machines are less likely to influence humans, the lower the perceived mental ability in machines.

C. *Overview of Approach*

We ran an experiment where participants engaged in the ultimatum game and a modified version of the dictator game with human vs. computer counterparts. This experiment is described in Section II. The results allowed us to understand whether people were showing a bias in favor of humans, which we expected given the findings presented above. In Section III, we fit this experimental data to Fehr and Schmidt's inequity aversion model. We calculate separate parameters for envy and guilt according to whether the counterpart is a human or a computer. We then compare these parameters to understand the nature of this bias. This section also presents the model's predictions, for human and computer counterparts, and compares them to the experimental data. Finally, Section IV presents our conclusions.

II. Experiment

In this section we present an experiment where participants engaged in decision making tasks with humans and computers. The first aim of the experiment was to test whether people would show systematic differences in behavior with computers, when compared to humans. The second aim was to collect data to estimate the parameters in Fehr and Schmidt's inequity aversion model. Following the methodology of Blanco, Engelmann, and Normann [25], participants engaged, in a repeated measures design, in the ultimatum game and a modified version of the dictator game. As discussed in Section III, these games are sufficient to determine, respectively, the α and β parameters.

A. Method

a) Design: The experiment followed a 2 × 2 repeated-measures factorial design: *Counterpart* (Human vs. Computer) × *Game* (Ultimatum Game vs. Modified Dictator Game). The counterpart and game order was counterbalanced across participants. Regarding the ultimatum game, participants played in both the role of proposer and responder (in a counterbalanced order). The initial endowment was 20 lottery tickets. The lottery tickets had financial consequences as they would enter a lottery for $50; thus, the more tickets participants got, the higher were their chances of winning the lottery. Moreover, participants were informed that computers would enter the lottery. Participants were quizzed on these instructions before playing the game. Participants also engaged in an interactive tutorial, prior to starting the game.

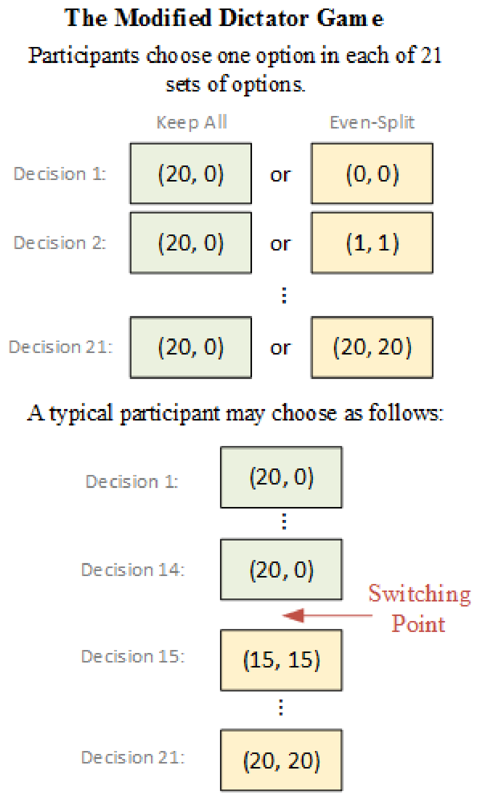
In the modified dictator game [25] participants have to decide between keeping all the money or accepting an even-split option. In our case, participants had to make twenty-one decisions; each decision consisted of a choice between the option where they keep all 20 lottery tickets – i.e., (20, 0) – or an even-split option – (0, 0), (1, 1), …, (20, 20). Fig.1 shows a visual description of the game. Notice that subjects with monotone preferences should have a single switching point (if any) where they change from choosing the left (20, 0) option to choosing the right even-split option. We are interested in what that switching point is, and whether it is different with humans vs. computers. Participants were informed that the lottery for this game was independent from the lottery for the ultimatum game, i.e., the outcome in this game would not affect the chances in the other lottery (and vice-versa).

b) Strategy-elicitation method: Following the methodology of Blanco et al. [25], we used the strategy-elicitation method. In this method, which has been extensively used in the past [26], participants report a choice for each possible decision node in the game. After reporting all choices, participants are randomly matched with a counterpart and assigned a role. The game is then simulated according to the reported decisions. This method contrasts with the direct-elicitation method, where participants are matched with other participants and engage in the game in real time. The advantage of the strategy-elicitation method is that it gathers more information per subject. This was also critical in our case, as decisions were necessary for all choice nodes in order to calculate the parameters in Fehr and Schmidt's model (see Section III).

The strategy-elicitation method also affected the logic of the lottery. Specifically, participants were instructed that, once all decisions were reported, the lottery would proceed as follows: (a) Participants are assigned a random role; (b) The counterpart is determined randomly to be a human or a computer; (c) The game is simulated according to the reported decisions and, if the counterpart is a computer, an algorithm that makes decisions "just like real people" is used; (d) Tickets are assigned to the players according to the outcome of the simulated game; and (e) A lottery winner is selected, according to the odds set by the lottery tickets.

c) Measures: We focused on the following dependent variables: (a) The proposer's offer in the ultimatum game; (b) For responders in the ultimatum game, the point at which participants switched from rejecting to accepting the proposer's offers; (c) In the modified dictator game, the point at which participants switched from choosing (20, 0) to the even-split.

Fig. 1. The modified dictator game.



d) Participants: Participants were recruited at the Psychology student pool of the University of Southern California. We recruited a total of 165 participants. We ran 6 participants at a time, throughout several sessions. Regarding gender, 53.0% were males. Age distribution was as follows: 21 years and under, 83.0%; 22 to 34 years, 17.0%. All participants were undergraduate or graduate students majoring in Psychology-related courses and mostly with citizenship from the United States (78.8%). Regarding incentive, first, participants were given school credit for their participation; second, with respect to their goal in the task, participants were instructed to earn as many tickets as possible, as the total amount of tickets would increase their chances of winning the lottery. All participants were asked to give informed consent before starting the task.

B. Results

The means and standard deviations for our measures in the ultimatum and dictator games are shown in Table I.

a) Ultimatum game (proposer): We compared participants' offers to humans and computers in the ultimatum game. The results showed that people offered more to humans than computers, $t(164) = 5.538$, $p = .000$, $r = .397$, mean difference = 1.20, 95% CI [.77, 1.63].

b) Ultimatum game (responder): We compared the point at which participants switched, with computers and humans, from rejecting to accepting offers in the ultimatum game. Participants that had more than one switching point – i.e., had inconsistent preferences – were excluded. We excluded 8 participants using this criterion. The results showed that participants did not show a statistically significant difference in switching point between humans and computers, $t(156) = 1.184$, $p = .238$.

c) Modified dictator game: We compared the point at which participants switched, with humans and computers, from choosing (20, 0) to the even-split in the modified dictator game. Participants that had more than one switching point – i.e., had inconsistent preferences – were excluded. We excluded 22 participants using this criterion. Participants that never switched to the even split – i.e., always chose (20, 0) – were assigned a switch point of 21 for this analysis. The results showed that participants switched to the even-split later with computers than with humans, $t(142) = 5.590$, $p = .000$, $r = .425$, mean difference = 2.52, 95% CI [1.63, 3.42].

C. Discussion

The results show that people can treat machines in a social manner [5]. Effectively, people's decisions with machines were, in general, more favorable than what selfish models of rational behavior predict. Nevertheless, as we expected, the results confirm that people still make important distinctions in their decisions with machines, when compared to humans. In the ultimatum game, people offered more tickets to humans than computers. In the modified dictator game, people required more tickets in the even split with computers than with humans, before they were willing to forfeit keeping all tickets. The only case where participants did not seem to distinguish between humans and computers was when they assumed the role of responders in the ultimatum game, in which case they accepted offers at the same level.

III. Model

In this section we use the data collected in our experiment to model people's decision making with machines and humans. As discussed in the Introduction, we focus on Fehr and Schmidt's inequity aversion model [11] and discuss alternative models in the General Discussion. We, first, estimated and compared the α and β parameters for human and machine counterparts. Then, we used these parameters to run the model and compare its predictions in the ultimatum and modified dictator games.

A. Parameter Estimation

We use the procedure from Blanco et al. [25] to derive the α (envy) and β (guilt) parameters. As responder in the ultimatum game, suppose r'_k is the lowest offer responder k is willing to accept and, thus, $r'_k - 1$ is the highest offer rejected. A responder will hence be indifferent between accepting some offer $r_k \in [r'_k - 1, r'_k]$ and getting a zero payoff for rejection. This means that $U_i(r_k, 20 - r_k) = r_k - \alpha_i(20 - r_k - r_k) = 0$. Thus, the estimate for the envy parameter is:

$$\alpha_k = \frac{r_k}{2(10 - r_k)} \qquad (2)$$

where we set $r_k = r'_k - 0.5$ as the point at which responders switched from rejecting to accepting offers. Participants that had more than one switch point – i.e., their prefences were inconsistent – were excluded from analysis. We excluded 8 participants using this criterion.

The guilt parameter is calculated based on the decisions in the modified dictator game. We get β_k by finding the egalitarian allocation, (x_k, x_k), such that the dictator is indifferent between keeping the entire endowment, the (20, 20) outcome, and (x_k, x_k). Suppose that an individual switches to the egalitarian outcome at (x'_k, x'_k); i.e., s/he prefers (20, 0) over $(x'_k - 1, x'_k - 1)$ but (x'_k, x'_k) over (20, 0). Therefore, s/he is indifferent between the (20, 0) distribution and $(\tilde{x}_k, \tilde{x}_k)$, where $\tilde{x}_k \in [x'_k - 1, x'_k]$ and $x'_k \in \{1, \ldots, 20\}$. Thus, $U_i(20,0) = U_i(\tilde{x}_k, \tilde{x}_k)$ if, and only if, $20 - 20\beta_k = \tilde{x}_k$. This leads to:

$$\beta_k = 1 - \frac{\tilde{x}_k}{20} \qquad (3)$$

where we set $\tilde{x}_k = x'_k - 0.5$ as the point at which people switched from choosing (20, 0) to choosing the even-split option. Participants that had more than one switch point – i.e., their preferences were inconsistent – were excluded from analysis. Using this criterion, we excluded 22 participants. Finally, for participants that always chose (20, 0) – i.e., never split – we set $\tilde{x}_k$ to 21. Notice that this means that $\beta < 0$, which is an extension of Fehr and Schmidt's assumptions. People that fall on this category might enjoy being better off than others.

B. Parameter Comparison

Using these formulas, we calculated parameters, for every participant, for human and computer counterparts. Then, we

TABLE I. Means and standard deviations for decisions in the ultimatum and dictator games

	n	Human		Computer	
		Mean	*SD*	*Mean*	*SD*
Ultimatum Game					
Offers (as proposer)	165	7.91	3.050	6.71	3.434
Minimum accepted offer (as responder)	157[a]	7.36	3.183	7.62	3.755
Modified Dictator Game					
Switching point from (20, 0) to even-split	143[b]	15.60	6.868	18.13	5.862

[a] Participants that switched from rejecting to accepting at more than one point (i.e., were inconsistent) were excluded.

[b] Participants that switched from keeping all to the even-split at more than one point (i.e., were inconsistent) were excluded.

ran dependent *t* tests to compare these parameters. The results showed that people did not show more envy (α parameter) with humans (*M* = 1.77, *SD* = 4.85) than with machines (*M* = 1.98, *SD* = 4.83), *t*(139) = .517, *p* = .606; in contrast, people showed more guilt (β parameter) with humans (*M* = .24, *SD* = .34) than with machines (*M* = .11, *SD* = .29), *t*(139) = 5.603, *p* = .000, *r* = .429, mean difference = .13, 95% CI [.08, .17].

C. Predictions

Having estimated the parameters, we used them to determine Fehr and Schmidt's predictions in the ultimatum and dictator games. Our focus was comparing the predictions with human vs. machine counterparts. Nevertheless, as a reference, we also compared these predictions to those made by a standard model that assumed players were selfish.

The predicted expected utilities for responders in the ultimatum game are shown in Table II, for each possible offer. Since the expected utility of rejecting the offer is 0, whenever the utitlity for accepting is greater than zero, the offer is accepted. Notice that to calculate these values we did not need to resort to the distribution of α and β parameters in the dataset. In this case, the responder makes the final decision and, thus, the decision is only contingent on the responder's α and β parameters. As shown in Table II, Fehr and Schmidt's model predicts that people will switch, both with humans and machines, from rejecting to accepting, when the offer is at least 8 tickets. This is a replication of our empirical finding reported above. Moreover, this prediction is much closer to the data than the selfish model's prediction of switching when the offer is at least 1 ticket.

TABLE II. EXPECTED PAYOFF IF RESPONDERS ACCEPT THE OFFER IN THE ULTIMATUM GAME. THE POINT AT WHICH RESPONDERS SWITCH FROM REJECTING TO ACCEPTING THE OFFER IS IN BOLD

Offer	Selfish	Fehr & Schmidt	Fehr & Schmidt
	Model	(Humans)	(Machines)
0	0	-35.328	-39.629
1	**1**	-30.795	34.666
2	2	-26.262	-29.703
3	3	-21.729	-24.740
4	4	-17.197	-19.777
5	5	-12.664	-14.814
6	6	-8.131	-9.852
7	7	-3.598	-4.889
8	8	**.935**	**0.0743**
9	9	5.467	5.037
10	10	10	10
11	11	10.517	10.773
12	12	11.034	11.546
13	13	11.551	12.319
14	14	12.068	13.092
15	15	12.585	13.865
16	16	13.102	14.638
17	17	13.619	15.411
18	18	14.136	16.184
19	19	14.653	16.957
20	20	15.170	17.731

Regarding the proposers in the ultimatum game, the predicted expected utilities, for each possible offer, are shown in Table III. Since these utilities are contingent on the counterparts' expected decision (to accept or reject the offer), we use the distribution of α and β parameters defined by our dataset to make this calculation. As shown in the table, Fehr and Schmidt's model predicts that people will offer 10 tickets whether they are engaging with human or machine counterparts. These predictions are much better than what is predicted by the selfish model; however, they fail to replicate the finding in our data that people favor humans to machines. This might have happened because, even though there is difference in the estimated parameters, this difference may have not been large enough[1].

TABLE III. EXPECTED PAYOFFS FOR PROPOSERS IN THE ULTIMATUM GAME, FOR EVERY POSSIBLE OFFER. THE OFFER WITH THE HIGHEST EXPECTED UTILITY IS IN BOLD

Offer	Selfish	Fehr & Schmidt	Fehr & Schmidt
	Model	(Humans)	(Machines)
0	10[a]	2.167	2.533
1	**19**	3.349	3.876
2	18	3.332	3.815
3	17	3.210	3.633
4	16	3.276	3.660
5	15	3.686	4.061
6	14	4.138	4.489
7	13	5.281	5.632
8	12	7.803	8.165
9	11	8.564	8.772
10	10	**10**	**10**
11	9	5.467	5.037
12	8	.935	.074
13	7	-3.598	-4.889
14	6	-8.131	-9.852
15	5	-12.664	-14.814
16	4	-17.197	-19.777
17	3	-21.729	-24.740
18	2	-26.262	-29.703
19	1	-30.795	-34.666
20	0	-34.066	-38.214

[a] The selfish model predicts that responders are indifferent between accepting or rejecting an offer of 0; thus, in this case, the expected payoff for the proposer is 10.

Regarding the decision in the modified dictator game, Fehr and Schmidt's model predicts that the expected utility of keeping all tickets with humans is 15.17, whereas with machines is 17.73. (As for responders in the ultimatum game, this calculation is only contingent on the decision maker's α and β parameters.) Thus, the model predicts that people will switch from choosing (20, 0) to choosing the even-split, when the offer is 16 with humans and 18 with machines. This is a replication of our finding that people wait longer with machines than humans to switch to the even-split. The selfish model would predict that people never choose the even-split option.

[1] In particular, replicating this might require assuming that some people have βs << 0 – i.e., they have a high preference for getting more tickets than others – especially when engaging with machines.

D. Discussion

Fitting the data to Fehr and Schmidt's model, clarified that people experience less guilt with machines, when compared to humans. In contrast, they seem to experience just as much envy with machines as with humans. Moreover, the model is rather successful in replicating the experimental data. For instance, in the ultimatum game, the model predicts that people will make fair offers (10 tickets) as proposers, and reject low offers (7 tickets or less) as responders. The model also had moderate success in replicating the distinctions people make between human vs. machine counterparts. On the one hand, in the modified dictator game, the model replicated that people wait longer with machines than with humans before choosing the even-split. The model also replicated that people start accepting offers, as responders in the ultimatum game, at the same level with humans and machines (8 tickets). On the other hand, the model did not replicate our finding that, in the ultimatum game, people make better offers with humans than machines.

IV. General Discussion

This paper, first, shows that people can reach efficiency with machines that goes beyond what is predicted by selfish models of decision making behavior. This is compatible with earlier findings that people often treat machines in a social manner, when immersed in social settings [5]. Nevertheless, the paper reinforces that people still make important distinctions between machines and humans. Everything else being equal, people offered more money to humans than machines, and they expected more money with machines than with humans before they were willing to forfeit an option where they kept everything vs. an option that made an even split.

Some could argue that this bias occurs because people are avoiding giving lottery tickets to computers, since if they win the lottery no one else can win it. However, we can present three reasons why this explanation is unlikely: (a) if people truly wanted to minimize the chances of computers winning the lottery, then they should have behaved according to the predictions of selfish models (e.g., offered 1 ticket as proposers in the ultimatum game); (b) people should have also avoided giving tickets to humans, as they can also win the lottery; (c) studies in human-computer interaction [5], [6] suggest that people often show social considerations for computers and, in many cases, this happens automatically.

To gain insight into this bias, we fit our experimental data to Fehr and Schmidt's [11] inequity aversion model. This revealed that, on the one hand, people did not show more aversion to disadvantageous outcome inequality (envy) with machines than with humans; but, on the other hand, people showed less aversion to advantageous inequality (guilt) with machines than humans. These results are compatible with general findings that emotion plays a critical role on people's decision making [27], [28]. Sanfey et al. [10], in particular, show that people experience less negative emotion with machines than with humans, and this impacts their decisions. Our results go further and suggest that the experience of guilt, in particular, is critical in explaining the differences in people's behavior with humans, when compared to machines.

These results have important implications for the design of autonomous agents that engage in decision making with humans. Designers need to acknowledge that people behave differently, at least by default, with (software or robotic) agents when compared to humans. To overcome this bias, designers can try to de-emphasize that people are interacting with autonomous agents. However, there are ethical and legal concerns that might limit this type of approach. For instance, the UK's 1998 Data Protection Act gives employees the right to ask for human intervention in the case of any decision made solely by automated means, when personal data is involved. A better alternative, thus, might be to emphasize that autonomous agents represent the interests of real humans. Our results also emphasize that people experience less emotion with machines than with humans. However, one way to get people to experience more emotion is to simulate expression of emotion in agents [29]. Research shows, in fact, that emotion in agents can have a powerful positive effect on people's behavior [1], [2], [30], [31]. Lastly, the findings reported in this paper are not relevant exclusively to software agents but also to other kinds of machines, such as social robots [32]. Research shows that, just like for computers, people have lower expectations about robots' mental capacities, when compared to humans [23]. Thus, we expect people will show a similar bias that favors humans to robots and, thus, designers should take appropriate measures to compensate for it.

Finally, in this paper we focused on Fehr and Schmidt's model, but experimental economists have proposed other models [18], [21], [19], [20]. Fehr and Schmidt fall within the class of inequity aversion models and, as shown in this paper, these models can provide insight into people's behavior with humans vs. machines. Another class of models that is likely to be useful is the one that focuses on the *intentions* underlying (fair or unfair) outcomes. These models argue that what is critical is not the outcome itself, but the intentions that led to it. For instance, a counterpart might have chosen a rather unfair outcome, but if he or she did not have a reasonable alternative, people are much more likely to accept the outcome [20]. Falk and Fischbacher [21], in turn, attempt to integrate inequity aversion with intentions and propose a model that is sensitive both to the outcome as well as to the underlying intentions. A potentially interesting line of inquiry, thus, could explore whether people are influenced by the machines' perceived intentions in the same manner as with humans. Overall, we propose that these economic models have much to offer to artificial intelligence; in particular, these models can help us understand the mechanisms underlying people's decision making with machines, when compared to humans.

References

[1] C. de Melo, P. Carnevale, and J. Gratch, "The effect of expression of anger and happiness in computer agents on negotiations with humans." in Proceedings of Autonomous Agents and Multiagent Systems, pp. 937-944, 2011.

[2] C. de Melo, P. Carnevale, and J. Gratch, "The impact of emotion displays in embodied agents on emergence of cooperation with people," Presence vol. 20, pp. 449-465, 2012.

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

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

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

[6] H. Gallagher, J. Anthony, A. Roepstorff, and C. Frith, “Imaging the intentional stance in a competitive game,” NeuroImage, vol. 16, pp.814-821, 2002.

[7] T. Kircher, I. Blümer, D. Marjoram, T. Lataster, L. Krabbendam, J. Weber et al., “Online mentalising investigated with functional MRI,” Neurosci Lett, vol. 3, pp. 176-181, 2009.

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

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

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

[11] E. Fehr, and K. Schmidt, “A theory of fairness, competition, and cooperation,” Q J Econ, vol. 114, pp. 817-868, 1999.

[12] E. Fehr, and K. Schmidt, “The economics of fairness, reciprocity and altruism – Experimental evidence and new theories,” in Handbook of the Economics of Giving, Altruism and Reciprocity, S-C. Kolm and J. Ythier, Eds., Oxford, UK: Elsevier, 2006, pp. 615-691.

[13] C. Camerer, Behavioral Game Theory, Experiments in Strategic Interaction. Princeton: Princeton University Press, 2003.

[14] W. Güth, R. Schmittberger, and B. Schwarze, “An experimental analysis of ultimatum bargaining,” J Econ Behav Organ, vol. 3, pp. 367-388, 1982.

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

[16] R. Forsythe, J. Horowitz, N. Savin, and M. Sefton, “Fairness in simple bargaining games,” Games Econ Behav, vol. 6, pp. 347-369, 1994.

[17] V. Smith, “An experimental study of competitive market behavior,” J Pol Econ, vol. 70, pp. 111-137, 1962.

[18] G. Bolton and A. Ockenfels, “A theory of equity, reciprocity and competition,” Am Econ Rev, vol. 100, pp. 166–193, 2000.

[19] D. Levine, “Modeling altruism and spitefulness in experiments,” Rev Econ Dynam Control, vol. 1, pp. 593-622, 1998.

[20] M. Rabin, “Incorporating fairness into game theory and economics,” Am Econ Rev, vol. 83, pp. 1281–1302, 1993.

[21] A. Falk and U. Fischbacher, “A theory of reciprocity,” Games Econ Behav, vol. 54, pp. 293-315, 2006.

[22] C. Nass, B. Fogg, and Y. Moon, “Can computers be teammates?,” I J Hum-Comp Stud, vol. 45, pp. 669-678, 1996.

[23] H. Gray, K. Gray, and D. Wegner, “Dimensions of mind perception,” Science, vol. 315, pp. 619, 2007.

[24] N. Haslam, “Dehumanization: An integrative review,” Pers Soc Psychol Rev, vol. 10, pp. 252-264, 2006.

[25] M. Blanco, D. Engelmann, and H. Normann, “A within-subject analysis of other-regarding preferences,” Games Econ Behav, vol. 72, pp. 321-338, 2011.

[26] J. Brandts and G. Charness, “The strategy versus the direct-response method: A first survey of experimental comparisons,” Exp Econ, vol. 14, pp. 375-398, 2011.

[27] A. Damasio, Descartes’ error: Emotion, reason and the human brain. New York, NY: Putnam, 1994.

[28] G. Loewenstein and J. Lerner, “The role of affect in decision making,” in R Handbook of Affective Sciences, . Davidson, K. Scherer, and H. Goldsmith, Eds. New York, NY: Oxford University Press, 2003, pp. 619-642.

[29] A. Choi, C. de Melo, P. Khooshabeh, W. Woo, and J. Gratch, “Physiological evidence for a dual process model of the social effects of emotion in computers,” I J Hum-Comp Stud, vol. 7, pp. 41-53, 2015.

[30] C. de Melo, P. Carnevale, and J. Gratch, “The importance of cognition and affect for artificially intelligent decision makers,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2014.

[31] R. Picard, Affective computing. The MIT Press, Cambridge, MA, 1997.

[32] C. Breazeal, “Toward sociable robots,” Robot Auton Syst, vol. 42, pp. 167-175, 2003.