PNAS
www.pnas.org

Supplementary Information for

Human Cooperation when Acting Through Autonomous Machines

Celso M. de Melo, Stacy Marsella, and Jonathan Gratch

Celso M. de Melo
Email: celso.miguel.de.melo@gmail.com

This PDF file includes:

Supplementary text
Fig. S1
Table S1
Captions for movie S1
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Movies S1

www.pnas.org/cgi/doi/10.1073/pnas.1817656116

Supplementary Information Text

Participant Samples

All participants were recruited from Amazon Mechanical Turk. In every experiment, we only sampled participants from the United States with an excellent performance history (95% approval rate on previous Mechanical Turk's tasks). To estimate sample sizes we used G*Power 3. In Experiment 1, based on earlier work (1), we predicted a medium effect size (Cohen's $f = 0.50$). Thus, for $\alpha = .05$ and statistical power of .80, the recommended total sample size was 102 participants. In practice, because some participants did not complete the task in the alloted time, we only recruited 98 participants (~49 participants per condition). Regarding gender, 43.9% of the participants were males. Age distribution was as follows: 18 to 21 years, 1.0%; 22 to 34 years, 60.2%; 35 to 44 years, 22.4%; 45 to 54 years, 10.2%; 55 to 64 years, 6.1%. In Experiments 2 and 2b, based on the previous experiment, we predicted small to medium effect sizes (Cohen's $f = 0.20$). For a statistical power of .80, $\alpha = .05$, the recommended total sample size was 366 participants. In practice, we recruited 334 participants for Experiment 2 (~41 participants per condition) and 351 participants for Experiment 2b (~44 participants per condition). Gender distribution was as follows: Experiment 2, 56.6% males; Experiment 2b, 44.2% males. Age distribution for Experiment 2 was: 18 to 21 years, 1.5%; 22 to 34 years, 57.5%; 35 to 44 years, 24.9%; 45 to 54 years, 9.9%; 55 to 64 years, 6.0%; over 65 years, 0.3%. Age distribution for Experiment 2b was: 18 to 21 years, 2.0%; 22 to 34 years, 44.4%; 35 to 44 years, 29.9%; 45 to 54 years, 14.2%; 55 to 64 years, 8.0%; over 65 years, 1.4%. Similarly, to the previous experiments, we estimated a total sample size of 366 participants for Experiment 3. In practice, we recruited 339 participants (~42 participants per condition). There were 47.8% males and ages were distributed as follows: 18 to 21 years, 3.5%; 22 to 34 years, 49.6%; 35 to 44 years, 28.0%; 45 to 54 years, 11.8%; 55 to 64 years, 6.2%; over 65 years, 0.9%. Finally, analogously to Experiment 1, the estimated sample size for Experiment 4 was 102 participants. In practice, we recruited 103 participants. There were 56.3% males and age was distributed as follows: 18 to 21 years, 1.9%; 22 to 34 years, 58.3%; 35 to 44 years, 22.3%; 45 to 54 years, 7.8%; 55 to 64 years, 4.9%; over 65 years, 4.9%.

Subjective Scales

To understand possible mediators for the effects of autonomy on cooperation, we created scales for focus on self, short-term reward saliency, fairness, and high-construal reasoning scales. We are not aware of any existent standard scales for these constructs that would fit to our domain, so we developed new scales based on the literature described in the Theoretical Foundation section. For every scale, participants were asked to rate on a 7-point scale (1, *not at all*, to 7, *very much*) how much each of the statements applied to them, in the context of the decisions made. The focus on self scale consisted of the following statements: I was mostly worried about myself; I wanted to make decisions that were best for all (reversed); I was concerned with the other participants' well-being (reversed); I focused on my own payoff; I was not concerned with the other participants; I felt connected with the other participants (reversed). The short-term reward saliency consisted of the following statements: I was focused on the impact of my

decisions on the environment (reversed); I wanted to minimize the negative impact on society (reversed); I was focused on the monetary payoff; I wasn't worried about how many lottery tickets I or the others made (reversed). The fairness scale consisted of the following statements: I tried to act as society expects me to behave in these situations; I wanted to be fair; I was not worried about what others would think of me (reversed); I was mindful of social norms; I was not worried about fairness (reversed). Finally, the high-construal reasoning scale consisted of the following statements: I was focused on the moment-to-moment aspects of the interaction (reversed); I was mostly focused with the "big picture"; I was thinking in more general terms, at a higher-level, about the situation.

To analyze the subjective scales, we first reduced the dimensionality of the scales using principal component analysis with varimax rotation. The focus on self scale was collapsed into a single factor, explaining 62.0% of the variance, with main loadings on "I wanted to make decisions that were best for all" (reversed) and "I was concerned with the other participants' well-being" (reversed). The short-term reward saliency scale was collapsed into a single factor, explaining 65.0% of the variance, with main loadings on "I was focused on the impact of my decisions on the environment" (reversed) and "I wanted to minimize the negative impact on society" (reversed). The fairness scale was collapsed into a single factor, explaining 58.2% of the variance, with main loadings on "I tried to act as society expects me to behave in these situations" and "I wanted to be fair". Finally, the high-construal reasoning scale was collapsed into two factors and we focused on the first factor, explaining 48.8% of the variance, with main loadings on "I was mostly focused with the 'big picture'" and "I was thinking in more general terms, at a higher-level, about the situation".

Social Value Orientation

Building on work by construal level theorists suggesting that the adoption of high construal abstract thinking can encourage cooperation (2-6), Giacomantonio et al. (7) proposed that, rather than simply promoting cooperation, abstract thinking reinforces one's values. Thus, under high construal thinking, prosocials – as measured by a social orientation scale (8) – would be more likely to cooperate, whereas individuals with a selfish orientation – or pro-selves – would be more likely to defect. In this Appendix, we test this hypothesis in all five experiments. Social value orientation (SVO) was measured, prior to engaging in the decision task, using the slider scale (9). To test the hypothesis, we focus on the SVO × autonomy (programming vs. direct interaction) interaction. For the analysis, we only focus on the prosocial and pro-self orientations, which constitute the majority of the participants, and exclude other orientations (e.g., competitive).

Experiment 1. The sample size for this experiment was 98 participants (60.2% prosocials). To analyze the data, we ran a SVO × autonomy ANOVA. The results revealed a main effect of SVO with prosocials (M = .66, SE = .05) cooperating more than pro-selves (M = .39, SE = .06), $F(1, 94) = 13.742$, $p < .001$, partial $\eta^2 = .128$. However, more importantly, there was no significant SVO × autonomy interaction, $F(1, 94) = 1.636$, $p = .202$.

Experiment 2. We excluded from this analysis 3 participants that were classified to have a competitive orientation, leaving a total of 331 participants (54.7% prosocials). To

analyze the data, we ran a SVO × autonomy × saliency × focus ANOVA. The results revealed no significant SVO × autonomy interaction, $F(1, 315) = .993$, $p = .320$. There were no other relevant effects involving SVO.

Experiment 2b. We excluded from this analysis 2 participants, one classified as altruist, and another as competitive, leaving a total of 349 participants (61.6% prosocials). To analyze the data, we ran a SVO × autonomy × saliency × focus ANOVA. The results revealed no significant SVO × autonomy interaction, $F(1, 333) = .066$, $p = .797$. There was a trend for a main effect of SVO in the expected direction, $F(1, 333) = 2.883$, $p = .090$, partial $\eta^2 = .009$. There were no other relevant effects involving SVO.

Experiment 3. We excluded from this analysis 3 participants, one classified as altruist, and two as competitive, leaving a total of 336 participants (55.1% prosocials). To analyze the data, we ran a SVO × autonomy × counterpart strategy × persistency ANOVA. The results revealed no significant SVO × autonomy interaction, $F(1, 320) = .295$, $p = .587$. There were no other relevant effects involving SVO.

Experiment 4. The sample size for this experiment was 103 participants (47.6% prosocials). To analyze the data, we ran a SVO × autonomy × counterpart strategy × persistency ANOVA. The results revealed no significant SVO × autonomy interaction, $F(1, 320) = .289$, $p = .592$. There were no other relevant effects involving SVO.

Discussion. In sum, our data did not support the hypothesis that prosocials would program machines to cooperate more and pro-selves to cooperate less. This result is compatible with the results in Experiments 2 and 2b showing that an experimental manipulation of social value orientation, via instructions as has been done in the past (9), did not interact with the autonomy effect either.

Gender and Cooperation

Gender effects on cooperation have been reported in the literature. Eckel and Grossman (10) review evidence from public goods, ultimatum, and dictator games and conclude that, when exposure to risk is minimal (e.g., as a responder in the ultimatum game), women make less individualistic choices than men. Simpson (11) argues and presents experimental evidence supporting that women tend to defect more due to fear, whereas men defect more due to greed. Van Vugt, De Cremer, and Janssen (12) also show that, when intergroup competition is heightened, men tend to behave more competitively than women. For these reasons, we looked at whether gender had an effect on cooperation in our case and, in particular, whether gender interacted with the autonomy effect. In general, we found no evidence for gender effects, which may be due to the fact that the prisoner's dilemma can motivate, albeit for different reasons, both men and women to defect (11).

Experiment 1. We ran a gender × autonomy ANOVA, which revealed a trend for a main effect of gender, with women ($M = .61$, $SE = .05$) cooperating more than men ($M = .48$, $SE = .06$), $F(1, 94) = 2.812$, $p < .097$, partial $\eta^2 = .029$. However, more importantly, there was no significant gender × autonomy interaction, $F(1, 94) = .005$, $p = .944$.

Experiment 2. We ran a gender × autonomy × saliency × focus ANOVA, which revealed a trend for a gender × autonomy interaction, $F(1, 318) = 2.840$, $p = .093$, partial

η^2 = .009, with women tending to cooperate slightly more than men, but only when programming the autonomous machine.

Experiment 2b. We ran a gender × autonomy × saliency × focus ANOVA but, this time, there was no gender × autonomy interaction, $F(1, 335) = .564$, $p = .453$.

Experiment 3. We ran a gender × autonomy × counterpart strategy × persistency ANOVA, which revealed, once again, no gender × autonomy interaction, $F(1, 323) = .101$, $p = .751$.

Experiment 4. We ran a gender × autonomy ANOVA, which revealed no gender × autonomy interaction, $F(1, 103) = .254$, $p = .615$.

Repeated-Measures Analysis of Cooperation Rate in Experiment 3

To understand how cooperation rate unfolded across rounds, we ran an autonomy × counterpart behavior × persistency × round mixed ANOVA. The analysis revealed a main effect of round, $F(9, 2979) = 3.98$, $p < .001$, partial $\eta^2 = .012$, with cooperation in the last rounds being lower than in the initial rounds. This may have happened because participants were punishing competitors as the task progressed. Effectively, there was a strategy × round interaction, $F(9, 2979) = 4.73$, $p < .001$, partial $\eta^2 = .014$: competitors experienced considerably less cooperation in the last round ($M = .36$, $SE = .04$) than in the initial round ($M = .54$, $SE = .04$); in contrast, cooperators experienced an increase in cooperation, though only a slight one, in the last round ($M = .57$, $SE = .04$) when compared to the initial round ($M = .52$, $SE = .04$).

To understand if participants were behaving differently when engaging with the same counterparts in every round vs. different counterparts in each round, we first note that there was no statistically significant main effect for this factor, $F(1, 331) = 1.10$, $p = .295$. This suggests that participants were punishing new counterparts, with whom they had no history, based on the experience from previous rounds with different counterparts. Nevertheless, a more detailed analysis reveals subtle differences in the way people behaved with same vs. different counterparts. There was a trend for a persistence × round interaction, $F(9, 2979) = 1.72$, $p = .080$, partial $\eta^2 = .005$, with cooperation tending to be lower in later rounds when engaging with the same counterparts in every round, when compared to interaction with different counterparts in every round. This was likely driven by the fact that participants were being harsher with counterparts that were consistently being competitive; effectively, cooperation rate with same competitors ($M = .41$, $SE = .03$) was lower than with different competitors ($M = .49$, $SE = .04$).

Finally, we wanted to understand if these rewarding and punishing effects were moderating the effect of autonomy. To get insight, we ran independent samples t tests comparing cooperation rate when programming vs. direct interaction, for each counterpart behavior × persistency combination. This revealed that, when engaging with the same competitors in every round, participants cooperated less when interacting directly than when programming, $t(86) = 2.333$, $p = .022$, $r = .244$; however, when engaging with different competitors in each round, there was only a weak trend for this effect, $t(81) = 1.026$, $p = .308$, $r = .133$. On the other hand, when engaging with different cooperators, there was a weak trend for increased cooperation when programming, $t(82) = 1.094$, $p = .277$; but, with different cooperators, there was no difference, $t(82) = .219$, $p = .827$. This suggests that interaction history can influence the autonomy effect and that

may explain why there was only a trend for a main effect of autonomy in this experiment ($F(1, 331) = 2.64$, $p = .105$, partial $\eta^2 = .008$).

Fig. S1. The experimental conditions in the abstract version of the *n*-person prisoner's dilemma (Experiment 3). (**a**) The condition where participants made the decision in real-time round-by-round. (**b**) The condition where participants program the computer agent to act on their behalf.

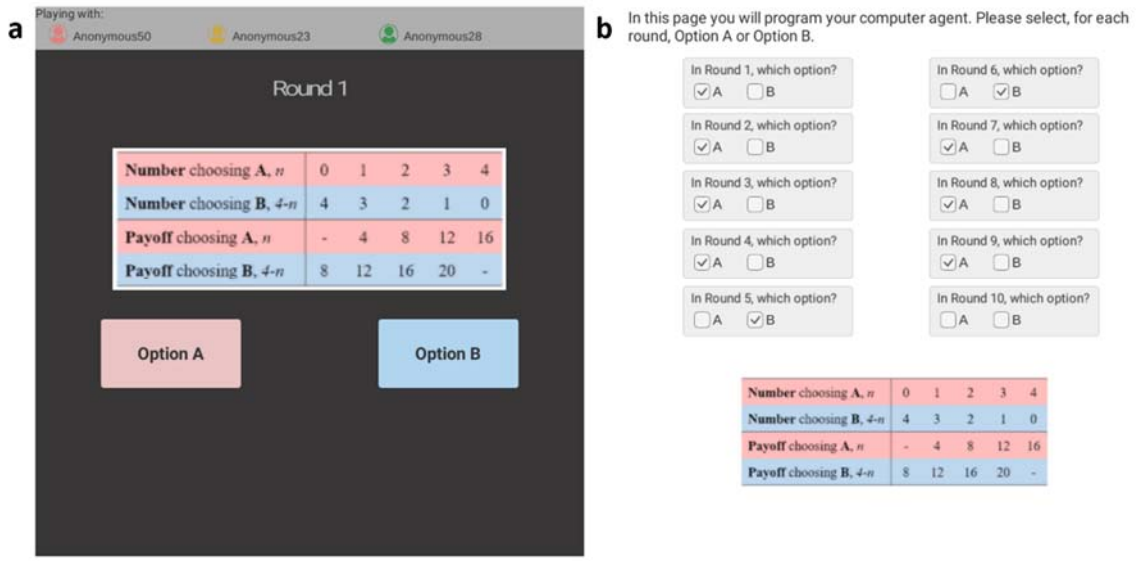

Table S1. Bootstrapping Analysis of the Statistical Significance of the Indirect Effects.

Indirect Effect	Point Estimate	Standard Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Total	.073	.024	.027	.122
Focus on Self	.001	.006	-.006	.019
Short-Term Reward Saliency	.063	.019	.029	.105
Fairness	.005	.007	-.003	.024
High-Construal Reasoning	.004	.005	-.001	.020

Movie S1. The software used in the experiments.

References

1. de Melo C, Marsella S, Gratch J (2017) Social decisions and fairness change when people's interests are represented by autonomous agents. *Auton Agents Multi Agent Syst*, 10.1007/s10458-017-9376-6.
2. 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 Appl Soc Psych* 31: 49-59.
3. Agerström J, Björklund F (2009) Moral concerns are greater for temporally distant events and are moderated by value strength. *Soc Cogn* 27: 261-282.
4. Kortenkamp K, Moore C (2006) Time, uncertainty, and individual differences in decisions to cooperate in resource dilemmas. *Pers Soc Psychol Bull* 32: 603-615.
5. Henderson M, Trope Y, Carnevale P (2006) Negotiation from a near and distant time perspective. *J Pers Soc Psychol* 91: 712-729.
6. De Dreu C, Giacomantonio M, Shalvi S, Sligte D (2009) Getting stuck or stepping back: Effects of obstacles in the negotiation of creative solutions. *J Exp Soc Psychol* 45: 542-548.
7. Giacomantonio M, De Dreu C, Shalvi S, Sligte D, Leder S (2010) Psychological distance boosts value-behavior correspondence in ultimatum bargaining and integrative negotiation. *J Exp Soc Psychol* 46: 824-829.
8. Van Lange P (1999) The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientations. *J Pers Soc Psychol* 77: 377-349.
9. Murphy R, Ackerman K, Handgraaf M (2011) Measuring social value orientation. *Judgm Dec Mak* 6: 771-781.Smith V (2003) Constructivist and ecological rationality in economics. *Am Econ Rev* 93: 465-508.
10. Eckel C, Grossman P (1999) Differences in the economic decisions of men and women: Experimental evidence. In C Plott, V Smith (Eds.) *Handbook of experimental results*. Amsterdam: Elsevier.
11. Simpson B (2003) Sex, fear, and greed: A social dilemma analysis of gender and cooperation. *Soc Forces* 82: 35-52.
12. Van Vugt M, De Cremer D, Janssen D (2007) Gender differences in cooperation and competition. *Psychol Sci* 18: 19-23.