Classification Major: Social Sciences

Classification Minor: Psychological and Cognitive Sciences

Human Cooperation when Acting Through Autonomous Machines

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

[a] US Army Research Laboratory, Playa Vista, CA 90094-2536, United States

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

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

Corresponding author: Celso M. de Melo, US Army Research Laboratory, Playa Vista, CA 90094-2536, +1 213-400-1121. E-mail: celso.miguel.de.melo@gmail.com

Keywords: Autonomous Vehicles, Cooperation, Social Dilemmas

Abstract

Recent times have seen an emergence of intelligent machines that act autonomously on our behalf, such as autonomous vehicles. Despite promises of increased efficiency, it is not clear whether this paradigm shift will change how we decide when our self-interest (e.g., comfort) is pitted against the collective interest (e.g., environment). Here we show that acting through machines changes the way people solve these social dilemmas and we present experimental evidence showing that participants program their autonomous vehicles to act more cooperatively than if they were driving themselves. We show this happens because programming causes selfish short-term rewards to become less salient, leading to considerations of broader societal goals. We also show that the programmed behavior is influenced by past experience. Finally, we report evidence that the effect generalizes beyond the domain of autonomous vehicles. We discuss implications for designing autonomous machines that contribute to a more cooperative society.

Significance Statement

Autonomous machines that act on our behalf – such as robots, drones, and autonomous vehicles – are quickly becoming a reality. These machines will face situations where individual interest conflicts with collective interest, and it is critical we understand if people will cooperate when acting through them. Here we show, in the increasingly popular domain of autonomous vehicles, that people program their vehicles to be more cooperative than they would if driving themselves. This happens because programming machines causes selfish short-term rewards to become less salient, and that encourages cooperation. Our results further indicate that personal experience influences how machines are programmed. Finally, we show this effect generalizes beyond the domain of autonomous vehicles and, discuss theoretical and practical implications.

\body

Introduction

Recently, Google launched a service that allows a bot that sounds like a human to make calls and conduct business on people's behalf. This is just the latest example of autonomous intelligent technology that can act on our behalf. Recent times have seen increasing interest in these kinds of autonomous machines including robots, personal home assistants, drones, and self-driving cars (1-4). These machines are changing the traditional ways we engage with others and it is important to understand those changes. Recent research confirms that interacting through machines can affect the decisions we make with others (5). These findings showed that in the ultimatum, impunity, and negotiation games, people were less likely to accept unfair offers if asked to program a machine to act on their behalf, when compared to direct interaction with proposers. This kind of experimental work has theoretical implications for our understanding of human decision making, and practical implications for the design of autonomous machines. However, these effects are still not well understood and, in particular, it is unclear if cooperation is affected when people engage with others through an autonomous machine. Here we present experimental evidence that sheds light on this question and, at the same time, demonstrate the practical implications in the increasingly popular domain of autonomous vehicles (AVs).

Autonomous vehicles promise to change the way we travel, move goods, and think of transportation in general (1). Not only are AVs supposed to be safer, many argue they are an important step toward a carbon-free economy (6). Three reasons typically justify this claim: reduction in total number of travelled miles, reduction in traffic congestion, and the trend towards electric-powered AVs. However, this analysis mostly neglects cognitive factors that influence the way AVs will be programmed to drive. There has been some interest in moral dilemmas AVs will face on the road, where decisions involving a cost in human life are made autonomously (2). However, less attention has been spent on the possibly more common case of social dilemmas that pit individual interests (e.g., comfort) against collective interests (e.g., the environment) that do not involve human life. Research shows that people can make distinct decisions in these dilemmas, when compared to situations involving sacred and moral values, such as when human life is involved (7, 8). In social dilemmas, the highest reward occurs when the individual defects and everyone else cooperates; however, if everyone defects, then everyone is worse off than if they had all cooperated (6). Autonomous machines – including self-driving cars – are bound to face these kinds of dilemmas often and, thus, we ask: Will people cooperate when engaging through these machines with others?

Theoretical Framework

When programming an autonomous machine, people are asked to think about the situation and decide ahead of time. This draws a parallel with the strategy method, often employed in experimental economics, whereby participants are elicited complete decision profiles by reporting how they would decide for each possible eventuality in the task. A meta-review of behavior in the ultimatum game revealed that, when asked to decide under the strategy method, participants were more likely to behave fairly, when compared to real-time interaction with their counterparts (10). This may occur because people will, on the one hand, consider the counterparts' perspective (11) and, on the other hand, rely on social norms – such as fairness – to provide a measure of consistency when considering the decisions for all possible outcomes (12). This research, thus, suggests that people should be more inclined to cooperate when programming an agent, than in real-time direct interaction.

A compatible point of view is that programming a machine leads people to focus less on the selfish short-term reward, in contrast to what would happen if the decisions were done on a moment-to-moment basis. In this sense, research has shown that the way a social dilemma is presented to the individual – or framed – can impact the saliency of the selfish reward (13, 14). Complementing the view that social dilemmas involve a conflict between individual and collective interests, researchers have also proposed that social dilemmas present an inherent temporal conflict between a decision that leads to an immediate short-term reward (i.e., defection) vs. a decision that leads to a larger long-term reward (i.e., cooperation) (15, 16). When someone defects, thus, it is because people are temporally discounting distant rewards (17-19) and the short-term reward is perceived to be more salient. To avoid the temptation of the short-term reward, people can adopt a course of action that precludes the short-term reward (i.e., a pre-commitment) in favor of the higher long-term reward (20-22) – in our case, programming the machine accordingly; moreover, once people make a commitment, they tend to behave in a manner that is consistent with that commitment (23, 24). Our hypothesis, thus, is that programming machines causes short-term rewards to become less salient and that leads to increased cooperation.

This hypothesis is also in line with predictions coming from work on construal level theory (25). According to the theory, when people mentally represent or construe a situation at a higher level, they focus on more abstract and global aspects of the situation; in contrast, when people construe a situation at a lower level, they focus on more specific context-dependent aspects of the situation. Building on this theory, Agerström and Björklund (26, 27) 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 (28) for a dissenting view. In separate studies, they showed that people made harsher moral judgments – in dilemmas such as the "recycling dilemma", which never involved costs in human life – of others' distant-future morally questionable behavior (26) and, that people were more likely to commit to moral behavior when thinking about distant versus near future events (27). In one of their studies (27; Experiment 4), they further showed that the saliency of moral values mediated the effect of temporal distance on moral behavior. Kortenkamp and Moore (29) further 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. In a negotiation setting, Henderson et al. (30) and De Dreu et al. (31) showed that individuals under high construal level negotiated more mutually beneficial and integrative agreements. Finally, Fujita et al. (32) noted that high-level thinking can enhance people's sense of control; increased self-control, in turn, can lead individuals to pre-commit to the more profitable distant reward (20-22). Therefore, the hypothesis is that programming machines leads to increased cooperation, because it motivates more abstract thinking, which is likely to decrease the saliency of selfish short-term rewards.

Finally, building on this work on construal level theory, Giacomantonio et al. (33) 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 (34) – would be more likely to cooperate, whereas individuals with a selfish orientation – or pro-selves – would be more likely to defect. In support of this thesis, they show that, in the ultimatum game and negotiation, prosocials were more likely to cooperate under high construal, but pro-selves more likely to compete. This work, thus, suggests an alternative hypothesis, whereby pro-socials would be more likely to program machines to cooperate, when compared to direct interaction; and, complementary, pro-selves would be more likely to program machines to defect, when compared to direct interaction.

Overview of Experiments

To study people's behavior when programming machines and the underlying mechanism, we ran five experiments where participants engaged in a social dilemma – the *n*-person prisoner's dilemma – as either the owners of an autonomous vehicle or as the drivers themselves. Economic games, such as the prisoner's dilemma, are abstractions of prototypical real-life situations and are ideal for studying the underlying psychological mechanisms in controlled experiments (35). The decisions participants make, moreover, always had real financial consequences, as the points they earned in the games would be converted to tickets for lotteries – one per experiment – for a $30 prize. Finally, to prevent any reputation concerns (36), the experiments were always fully anonymous – i.e., the participants were anonymous to each other and to the experimenters (see the Methods section for details on how this was accomplished).

The prisoner's dilemma is presented to the participants as an environmental social dilemma (16), where they choose between two options: turning on the air conditioner (AC) – defection – which increased comfort at the expense of the environment, as more gas was consumed; and, turning off the AC – cooperation – which sacrificed comfort in favor of the environment, as less gas was consumed. The critical comparison was between the cooperation rate – i.e., how often the AC was turned off – when acting through an AV vs. driving themselves.

When studying machines that act on behalf of humans, it is also important to clarify how much autonomy is given to these machines. On one extreme, the decisions made by the machine can be fully specified by the humans; on the other extreme, the machine could make the decision by itself with minimal input from humans. The degree of autonomy is an important factor that influences the way people behave with machines (37). Effectively, research suggests that the degree of thought and intentionality behind a decision can impact people's reactions (38). In this paper, we focus on machines that have minimal autonomy. Thus, our machines make decisions that are fully specified by the humans they represent. The rationale is that, if engaging with others through machines that have minimal autonomy affects social behavior, then the effect is likely to exist (and possibly be stronger) with increased autonomy.

In Experiment 1, we tested and confirmed that people cooperated more when programming their AVs, when compared to direct interaction with others. This effect was not moderated by the participants' social value orientations. In Experiment 2, we showed that programming caused selfish short-term rewards to become less salient and this led to increased cooperation. Experiment 2b revealed that this effect was robust even when participants could re-program their vehicles during the interaction. Experiment 3 showed that participants adjusted their programs based on the (cooperative or competitive) behavior they experienced from their counterparts. Finally, in Experiment 4, we showed that the effect also occurred in an abstract social dilemma, thus suggesting it generalizes beyond the domain of autonomous vehicles.

Experiment 1: Do People Program AVs to Cooperate More?

This experiment followed a simple 2-level between-participants design: *autonomy* (programming vs. direct interaction). Participants engaged with three counterparts in the 4-person prisoner's dilemma. The payoff table for this game is shown in Fig. 1-a (top center). Participants made simultaneous decisions without communicating with each other. For instance, if the participant chooses to turn off the AC and the remaining players choose to turn it on, then the participant gets 4 points, whereas the other players get 12 points (this is the 2nd column in the payoff table). If everybody chooses to turn off the AC, then everybody gets 16 points (5th column). However, if everyone decides to turn on their AC, then everyone gets only 8 points (1st column). Participants engaged in 10 rounds of this dilemma with the same counterparts. For this first experiment we wanted to exclude any strategic considerations that could occur due to repeated interaction. To accomplish this, participants were informed that they would not learn about the decisions others made until the end of the last round, when they found out how much they earned in total. Before engaging in the task, participants were quizzed to make sure that they understood the instructions and they were not able to proceed until they answered all questions correctly.

In the case of direct interaction, participants would see their car driving down a road and come to a stop; at that point, they were given the opportunity to make the decision on whether to turn on the AC. After making their decision, the car would continue driving and the procedure was repeated for every round. In the autonomous vehicle case, participants programmed their car, before the first round, to make whichever decision they wanted in each round. Then, a simulation would start and show the car driving down a road and come to a stop. At that point, without intervention from the participant, the car would play the decision it was programmed to make autonomously. This procedure was repeated for every round. (Please see the SI for a video showing the software used in every experiment and demonstrating all experimental conditions, Movie S1.)

Before starting the task, participants had to wait approximately 30 seconds while "they waited for other participants to join"; additionally, after submitting their decisions in each round, participants also had to wait while "other participants finished making their decisions". However, in order to increase experimental control, participants always engaged with a computer script that simulated the other participants. Similar experimental manipulations have been used in other experiments studying behavior involving intelligent machines (5, 39, 40). Participants were fully debriefed about this experimental procedure at the end of the experiment. All the experimental methods used in the experiments presented in this paper were approved by the University of Southern California IRB (ID# UP-14-00177) and the US Army Research Lab IRB (ID# ARL 18-002).

The main measure was cooperation rate, averaged across the 10 rounds. Prior to starting the task, we also collected the participants' social value orientation (SVO) using the slider scale (41). We recruited 98 participants from an online pool – Amazon Mechanical

Turk. Previous research shows that studies performed on Mechanical Turk can yield high-quality data and successfully replicate the results of behavioral studies performed on traditional pools (42). Since some gender effects on cooperation have been reported in the literature (43, 44), we began by running a gender × autonomy between-participants ANOVA; however, we did not find any statistically significant main effects of gender or gender × autonomy interactions in any of the experiments and, so, we will not address it further (but see the SI for a full analysis of gender). To compare cooperation rate for programming vs. direct interaction, we ran an independent samples *t*-test. The results confirmed that participants were statistically significantly more likely to turn off the AC – i.e., cooperate – when programming their AV to act on their behalf (*M* = .64, *SE* = .06) than when driving themselves (*M* = .47, *SE* = .05), *t*(96) = 2.33, *p* = .022, *r* = .23. To test if social value orientation was moderating this effect (33), we ran a SVO × autonomy between-participants ANOVA, which revealed no significant SVO × autonomy interaction, *F*(1, 94) = 1.636, *p* = .202. Thus, we found no evidence supporting the contention that prosocials would program AVs to cooperate more, whereas pro-selves would program AVs to cooperate less. Similar patterns were found in all experiments and, thus, we do not address SVO further (but see the SI Appendix for a complete analysis of SVO in all experiments).

Experiment 2: What is the Mechanism?

The previous experiment showed that people programmed AVs to cooperate more than when driving themselves. In Experiment 2, we wanted to get insight on the mechanism behind this effect. As described in the Theoretical Foundation section, we compare complementary mechanisms based on focus on self, saliency of the short-term reward, fairness, and high-construal reasoning. To accomplish this, we introduced a second experimental factor, *saliency*, that instructed the participant that turning on the AC was more likely than the other option to lead to a higher payoff for themselves (Fig. 1-a); in the other case, in contrast, participants were informed that turning on the AC was less likely than the other option to be environmentally-friendly (Fig. 1-b). To compare saliency to more traditional manipulations of social motivation (45), we introduced a third factor, *focus*, that simply instructed participants to either focus on how the decision would "affect yourself" (Fig. 1-a) versus "yourself and the others" (Fig. 1-b). Finally, we administered four subjective scales, after completing the task, that represented four possible mediators: focus on self (e.g., "I was mostly worried about myself"), short-term reward saliency (e.g., "I was focused on the monetary payoff"), fairness (e.g., "I wanted to be fair"), and high-construal reasoning (e.g., "I was thinking in more general terms, at a higher-level, about the situation"). Please see the SI for full details on these scales.

The experiment, thus, followed a 2 × 2 × 2 between-participants factorial design: *autonomy* (programming vs. direct interaction) × *saliency* (monetary reward vs. environment) × *focus* (self vs. collective). The autonomy factor and the experimental procedure were similar to Experiment 1 and we used an identical financial incentive. We recruited 334 participants from Amazon Mechanical Turk. To analyze cooperation rate, we ran an autonomy × saliency × focus between-participants ANOVA. The average cooperation rates are shown in Fig. 2-a. Once again, participants were more likely to turn off the AC – i.e., cooperate – when programming their AVs ($M = .58$, $SD = .03$) than when driving themselves ($M = .46$, $SE = .03$), $F(1, 326) = 8.56$, $p = .004$, partial $\eta^2 = .026$. In addition, participants cooperated more when instructed to focus on the collective ($M = .57$, $SE = .03$) than the self ($M = .47$, $SE = .03$), $F(1, 326) = 6.22$, $p = .013$, partial $\eta^2 = .019$; there was, however, no statistically significant autonomy × focus interaction ($F(1, 326) = 1.02$, ns). This suggests that, even though cooperation is impacted by experimentally manipulating social motivation, as has been shown in the past (45), this factor did not appear to explain the effect of autonomy. Finally, people cooperated more when the environment was more salient ($M =$

.64, *SE* = .03) than the monetary reward (*M* = .40, *SE* = .03), $F(1, 326) = 41.76$, $p < .01$, partial $\eta^2 = .114$; moreover, contrasting to the previous case, this time there was a trend for an autonomy × saliency interaction, $F(1, 326) = 2.95$, $p = .087$, partial $\eta^2 = .009$: people cooperated more when programming AVs than in direct interaction, but only when the monetary reward was salient.

To complement and reinforce this analysis, we ran a multiple mediation analysis (46) where we tested if focus on self, short-term reward saliency, fairness, and high-construal reasoning mediated the effect of autonomy on cooperation. The independent variable, autonomy, was binary coded: 0, for direct interaction; 1, for programming. The dependent variable was cooperation rate. Fig. 2-b shows the mediation analysis results. (See SI for details on bootstrapping tests for statistical significance of the indirect effects, Table S1.) The results indicated that the total indirect effect was statistically significant (.073, $p = 0.04$) and, more importantly, the indirect effect for short-term reward saliency was the only one which was statistically significant (.063, $p = 0.03$), i.e., the corresponding 95% bootstrapping confidence interval did not include zero. The analysis also indicated that the total effect (.122, $p = .003$) became statistically non-significant once we accounted for the effect of the mediators – i.e., the direct effect was statistically non-significant (.048, *ns*). In sum, the mediation analysis suggests that the effect of autonomy on cooperation was proximally caused by reduced saliency of the short-term reward.

Experiment 2b: What if Re-Programming was Permitted?

In the previous experiment, participants programmed their autonomous vehicles in the beginning and, then, were never given the opportunity to re-program during the task; in other words, participants were required to pre-commit to their choice at the start of the task. Research indicates that pre-commitment can be effective in minimizing the temptation to choose the short-term reward (20-22). However, does the effect still occur if people are not required to pre-commit? The answer to this question matters because, in practice, manufacturers are unlikely to force customers to pre-commit their decisions and never give them the chance to change their decisions based on their driving experience. Experiment 2b, thus, replicates the previous experiment but with one important change: participants could re-program their vehicles, if so desired, at the start of each round.

To remove strategic considerations, the previous experiment had participants engage with the same counterparts throughout all rounds while, at the same time, preventing them from learning what the counterparts' decided in each round. In this experiment, we wanted to use a more natural paradigm that still minimized strategic reasoning. In Experiment 2b, first, participants engaged with different counterparts in each round. Second, participants learnt, at the end of each round, about the counterparts' decisions. The decisions the counterparts made were neither too cooperative, nor too competitive: in half of the rounds, two counterparts defected (Rounds 1, 3, 6, 8, and 10) and, in the other half, two counterparts cooperated (Rounds 2, 4, 5, 7, and 9).

This experiment, thus, followed the same design and procedure as Experiment 2, with three exceptions. First, participants in the programming condition where given the opportunity to re-program the vehicle at the start of every round, following the first round where everyone was required to program the vehicle. When re-programming, participants would see a screen with all the current decisions chosen by default and would be allowed to change the decisions for every round that hadn't occurred yet. Second, participants were told that each round would be played with different counterparts and that they would never engage with the same counterpart for more than one round. Finally, participants were given information about the counterparts' decisions at the end of each round.

We recruited 351 participants from Amazon Mechanical Turk. Similarly to Experiment 2, participants cooperated more when programming their AVs (*M* = .54, *SD* =

.02) than when driving themselves (*M* = .47, *SE* = .02), *F*(1, 343) = 4.59, *p* = .033, partial η^2 = .013. Participants also cooperated more when the environment was more salient (*M* = .63, *SE* = .02) than the monetary reward (*M* = .37, *SE* = .03), *F*(1, 343) = 56.88, *p* < .01, partial η^2 = .142. There was also a replication of the trend for an autonomy × saliency interaction, *F*(1, 343) = 2.96, *p* = .086, partial η^2 = .009: people cooperated more when programming AVs than in direct interaction, but only when the monetary reward was salient. In this experiment, however, there was no statistically significant effect of focus on the collective (*M* = .51, *SE* = .03) versus the self (*M* = .50, *SE* = .02), *F*(1, 343) = .720, *ns*. The failure to replicate this effect suggests that the effect size of focus is small in this setting when compared to the effect of saliency. Overall, the experiment reinforces that people are likely to focus less on the short-term reward and cooperate more when programming their vehicles, even when given the opportunity to change their minds throughout the interaction (i.e., in the absence of a pre-commitment).

Experiment 3: What if Others Behaved Cooperatively or Competitively?

In the previous experiments we created conditions to minimize strategic interactions that may occur due to repeated interaction; in this experiment, by contrast, we explicitly study participants' behavior when they engage with counterparts that either behaved cooperatively or competitively in repeated interaction. Based on recent work showing people adjust their negotiation programs based on personal experience (47), we expected people would adapt their programming behavior according to their experience in the repeated dilemma. Furthermore, we wanted to study if participants would differentiate their behavior when the interaction was with different counterparts in every round – i.e., repeated one-shot games – or always with the same counterparts – i.e., a single repeated game. This is relevant since, in practice, people are more likely to engage in their driving with different drivers and, thus, it is important to understand if their past experience spills to new drivers with whom there is no interaction history.

The experiment followed a 2 × 2 × 2 between-participants factorial design: *autonomy* (programming vs. direct interaction) × *counterpart behavior* (cooperative vs. competitive) × *persistency* (always same counterparts vs. different counterparts in every round). In the cooperative condition, in half of the rounds, two counterparts would cooperate (Rounds 1, 3, 4, 7, and 8); in the other half, every counterpart would cooperate (Rounds 2, 5, 6, 9, and 10). In the competitive condition, in half of the rounds, two counterparts would defect (Rounds 1, 3, 4, 7, and 8); in the other half, every counterpart would defect (Rounds 2, 5, 6, 9, and 10). We recruited 339 participants from Amazon Mechanical Turk. The cooperation rate for each condition, across rounds, is shown in Fig. 3. To analyze the data we ran a between-participants ANOVA that confirmed a trend for a main effect of autonomy, $F(1, 331) = 2.64$, $p = .105$, partial $\eta^2 = .008$, with people cooperating more when programming their AVs ($M = .54$, $SD = .02$) than when driving themselves ($M = .48$, $SE = .02$); this effect becomes statistically significant if we look only at the last round $F(1, 331) = 5.17$, $p = .024$, partial $\eta^2 = .015$. The results confirmed an effect of counterpart strategy with participants cooperating more with cooperative ($M = .57$, $SE = .02$) than competitive counterparts ($M = .45$, $SE = .02$), $F(1, 331) = 12.96$, $p < .001$, partial $\eta^2 = .038$. However, there was no significant autonomy × counterpart strategy interaction – $F(1, 331) = 1.30$, $p = .255$ – which suggests that counterparts' behavior was just as likely to influence participants' programming as driving behavior. Somewhat surprisingly, we found no effect of persistency – $F(1, 331) = 1.10$, $p = .295$ – which suggests that people were still adjusting their behavior with new counterparts, with whom they had no history, based on the experience from previous rounds with different counterparts. A more detailed analysis of how cooperation rate unfolded across rounds is presented in the SI.

Experiment 4: Does the Effect Generalize?

This experiment sought to study whether the effect identified in the previous experiments generalize beyond this environment preservation task involving autonomous vehicles. To test that, in this final experiment, we had participants engage in an abstract social dilemma, without any mention of AVs. Similarly to Experiment 1, this experiment followed a 2-level between-participants design: *autonomy* (programming vs. direct interaction). The experimental procedure was similar to the one followed in Experiment 2b, except that in this case we did not consider the 'saliency' and 'focus' factors. The important difference was that the *n*-person social dilemma was presented abstractly and there was no mention of "self-driving cars" or "environment". In this case, the participant made a choice between "Option A" – corresponding to cooperation – and "Option B" – corresponding to defection (see SI for screenshot, Fig. S1). The payoff matrix was exactly the same as in the previous experiments. To justify the programming condition, we mentioned that the purpose of the experiment was to study how people make decisions through "computer agents" that act on their behalf. Thus, in this experiment, instead of programming self-driving cars, participants programmed a computer agent. Similarly to Experiment 2b, participants were allowed to re-program the agents at the start of each round. We recruited 103 participants from Amazon Mechanical Turk. To analyze cooperation rates, we ran an independent samples *t*-test for programming vs. direct interaction. In line with the previous experiments, the results showed that participants cooperated more when programming their agents (*M* = .45, *SE* = .05) than when engaging directly with others (*M* = .31, *SE* = .04), *t*(101) = 2.21, *p* = .029, *r* = .22.

General Discussion

As autonomous machines that act on our behalf become immersed in society, it is important we understand whether and, if so, how will people change their behavior and what are the corresponding social implications. Here we show that this analysis is incomplete if we ignore the cognitive factors shaping the way people decide through these machines. In the increasingly relevant domain of autonomous vehicles, we showed that programming vehicles ahead of time, led to an important change in the way social dilemmas are solved with people being more cooperative – i.e., more willing to sacrifice individual for the collective interest – than if driving the vehicle directly. The paper shows that this effect prevailed even if participants were allowed to re-program their machines (Experiment 2b), and independently of whether others behaved cooperatively or competitively (Experiment 3). We also showed that this effect was not limited to the domain of autonomous vehicles. In line with earlier work showing that people program machines to act more fairly in abstract games (5), participants still showed increased cooperation when programming a machine in an abstract version of the dilemma that had no mention of AVs (Experiment 4).

The paper provides insight into the mechanism driving the effect, emphasizing the role programming machines plays in lowering the saliency of the short-term reward when deciding whether to cooperate (Experiment 2). This is in line with research showing that introducing an opportunity to make a pre-commitment – in this case, by programming the machine ahead of time – can reduce the temptation to choose the short-term reward (20-22). However, the results also clarify that the effect still occurs in the absence of pre-commitment (Experiment 2b), thus, suggesting that people are motivated to maintain the commitment they made in the initial round (23, 24). The findings also align with research indicating that, when people are asked to report their decisions ahead of time, they tend to behave more fairly than when making the decision in real-time (10-12), and with prior work showing that cooperation can be encouraged by the way a social dilemma is framed (13, 14). Finally, these findings are compatible with earlier work by construal level theorists suggesting that the adoption of high construal abstract thinking – as is motivated by programing an autonomous machine –

increases the saliency of moral values (26, 27, 29-31) and self-control (32), which can encourage cooperation.

This paper focused on the case where the decision was made by the owner of the autonomous machine. This allowed us to identify an important effect and mechanism for cooperation that is made salient in virtue of having to program the machine. However, in practice, this decision may be distributed across multiple stakeholders with competing interests, including Government, manufacturers, and owners. In the case of a moral dilemma, Bonnefon et al. (2) showed that people would prefer other people's AVs to maximize preservation of life (even if that meant sacrificing the driver), whereas their own vehicle to maximize preservation of the driver's life. As these issues are debated, it is important we understand that in the possibly more prevalent case of social dilemmas – where individual interest is pitted against collective interest in situations that do not involve a cost in human life – autonomous machines have the potential to shape how the dilemmas are solved and, thus, these stakeholders have an opportunity to promote a more cooperative society.

Methods

This section describes experimental methodology details that are not described in the main body of the text. All methods used were approved by the University of Southern California IRB (ID# UP-14-00177) and the US Army Research Lab IRB (ID# ARL 18-002). Participants gave written informed consent and were debriefed, at the end, about the experimental procedures. Please see the SI for details on participant samples and the subjective scales used in Experiment 2.

Financial Incentives. Participants were paid $2.00 for participating in the experiment. Moreover, they had the opportunity to earn more money according to their performance in the tasks. Each point earned in the task was converted to a ticket for a lottery worth $30.00. This was real money and was paid to the winner through Amazon Mechanical Turk's bonus system. We conducted a separate lottery for each experiment.

Full Anonymity. All experiments were fully anonymous for participants. To accomplish this, counterparts were referred to as "anonymous" and we never collected any information that could identify participants. To preserve anonymity with respect to experimenters, we relied on the anonymity system of the online pool we used – Amazon Mechanical Turk. When interacting with participants, researchers are never able to identify the participants, unless we explicitly ask for information that may serve to identify them (e.g., name, email, or photo), which we did not. This experimental procedure is meant to minimize any possible reputation effects (30), such as a concern for future retaliation for the decisions made in the task.

Acknowledgments

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

References

1. Waldrop M (2015) No drivers required. *Nature* 518: 20-23.
2. Bonnefon J-F, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. *Science* 352: 1573-1576.
3. Floreano D, Wood R (2015) Science, technology and the future of small autonomous drones. *Nature* 521: 460-466.
4. Stone R, Lavine M (2014) The social life of robots. *Science* 346: 178-179.

5. 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.

6. Alexander-Kearns M, Peterson M, Cassady A (2016) *The impact of vehicle automation on carbon emissions*. Retrieved from Center for American Progress Website: https://www.americanprogress.org/issues/green/reports/2016/11/18/292588/the-impact-of-vehicle-automation-on-carbon-emissions-where-uncertainty-lies/.

7. Tetlock P, Kristel O, Elson B, Green M, Lerner J (2000) The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *J Pers Soc Psychol* 78: 853-870.

8. Dehghani M, Carnevale P, Gratch J (2014) Interpersonal effects of expressed anger and sorrow in morally charged negotiation. *Judgm Decis Mak* 9: 104-113.

9. Dawes R (1980) Social dilemmas. *Annu Rev Psychol* 31: 169-93.

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

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

12. Rauhut H, Winter F (2010) A sociological perspective on measuring social norms by means of strategy method experiments. *Soc Sci Res* 39: 1181-1194.

13. Pruitt D (1970) Motivational processes in the decomposed prisoner's dilemma game. *J Pers Soc Psychol* 14: 227-238.

14. Liberman V, Samuels S, Ross L (2004) The name of the game: predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Pers Soc Psychol Bull* 30: 1175-1185.

15. Dewitte S, De Cremer D (2001) self-control and cooperation: Different concepts, similar decisions? A question of the right perspective. *J Psychol* 135: 133-153.

16. Joireman J (2005) Environmental problems as social dilemmas: The temporal dimension. In: Strathman, A., & Joireman, J. (Eds.), *Understanding behavior in the context of time*, pp. 289-304, London: Lawrence Erlbaum Associates.

17. Frederick S, Loewenstein G, O'Donoghue T (2002) Time discounting and time preference: A critical review. *J Econ Lit* 15: 351-401.

18. Mannix E (1990) Resource dilemmas and discount rates in decision making groups. *J Exp Soc Psychol* 27: 379-391.

19. Kortenkamp K, Moore C (2006) Time, uncertainty, and individual differences in decisions to cooperate in resource dilemmas. *Pers Soc Psychol Bull* 32: 603-615.

20. Wertenbroch K (1998) Consumption self-control by rationing purchase quantities of virtue and vice. *Marketing Sci* 17: 317-337.

21. Ariely D, Wertenbroch K (2002) Procrastination, deadlines, and performance: Using precommitment to regulate one's behavior. *Psychol Sci* 13: 219-224.

22. Ashraf N, Karlan D, Yin W (2006) Tying Odysseus to the mast: Evidence from a commitment savings product in the Philippines. *Q J Econ* 121: 673-697.

23. Cialdini R, Goldstein N (2004) Social influence: Compliance and conformity. *Annu Rev Psychol* 55: 591-621.

24. Cialdini RB, Trost MR. 1998. Social influence: social norms, conformity, and compliance. In: The Handbook of Social Psychology, 4th ed, Gilbert D, Fiske S, Lindzey G, pp. 151-192, Boston: McGraw-Hill.

25. Liberman N, Trope Y (2010) Construal-level theory of psychological distance. *Psychol Rev* 117: 440-463.23

26. 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.

27. Agerström J, Björklund F (2009) Moral concerns are greater for temporally distant events and are moderated by value strength. *Soc Cogn* 27: 261-282.

28. Gong H, Medin D (2012) Construal levels and moral judgment: Some complications. *Judgm Decis Mak* 7: 628-638.

29. Kortenkamp K, Moore C (2006) Time, uncertainty, and individual differences in decisions to cooperate in resource dilemmas. *Pers Soc Psychol Bull* 32: 603-615.

30. Henderson M, Trope Y, Carnevale P (2006) Negotiation from a near and distant time perspective. *J Pers Soc Psychol* 91: 712-729.

31. 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.

32. Fujita K, Trope Y, Liberman N, Levin-Sagi M (2006) Construal levels and self-control. *J Pers Soc Psychol* 90: 351-367.

33. 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.

34. 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.

35. Pruitt D, Kimmel M (1977) Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annu Rev Psychol* 28: 363-392.

36. Hoffman E, McCabe K, & Smith V (1996) Social distance and other-regarding behavior in dictator games. *Am Econ Rev* 86: 653-660.

37. Endsley M (2017) From here to autonomy: Lessons learned from human-automation research. *Hum Factors* 59: 5-27.

38. Blount S (1995) When social outcomes aren't fair: The effect of causal attributions on preferences. *Organ Behav Hum Decis Process* 63: 131-144.

39. Kircher T, Blümel I, Marjoram D, Lataster T, Krabbendam L et al (2009) Online mentalising investigated with functional MRI. *Neurosci Lett* 454: 176-181.

40. 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.

41. Murphy R, Ackerman K, Handgraaf M (2011) Measuring social value orientation. *Judgm Dec Mak* 6: 771-781.

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

43. Simpson B (2003) Sex, fear, and greed: A social dilemma analysis of gender and cooperation. *Soc Forces* 82: 35-52.

44. Van Vugt M, De Cremer D, Janssen D (2007) Gender differences in cooperation and competition. *Psychol Sci* 18: 19-23.

45. Smith V (2003) Constructivist and ecological rationality in economics. *Am Econ Rev* 93: 465-508.

46. Preacher K, Hayes A (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods* 40: 879-891.

47. Mell J, Lucas G, Gratch J (2018) Welcome to the real world: How agent strategy increases human willingness to deceive. In *Proceedings of AAMAS 2018*.

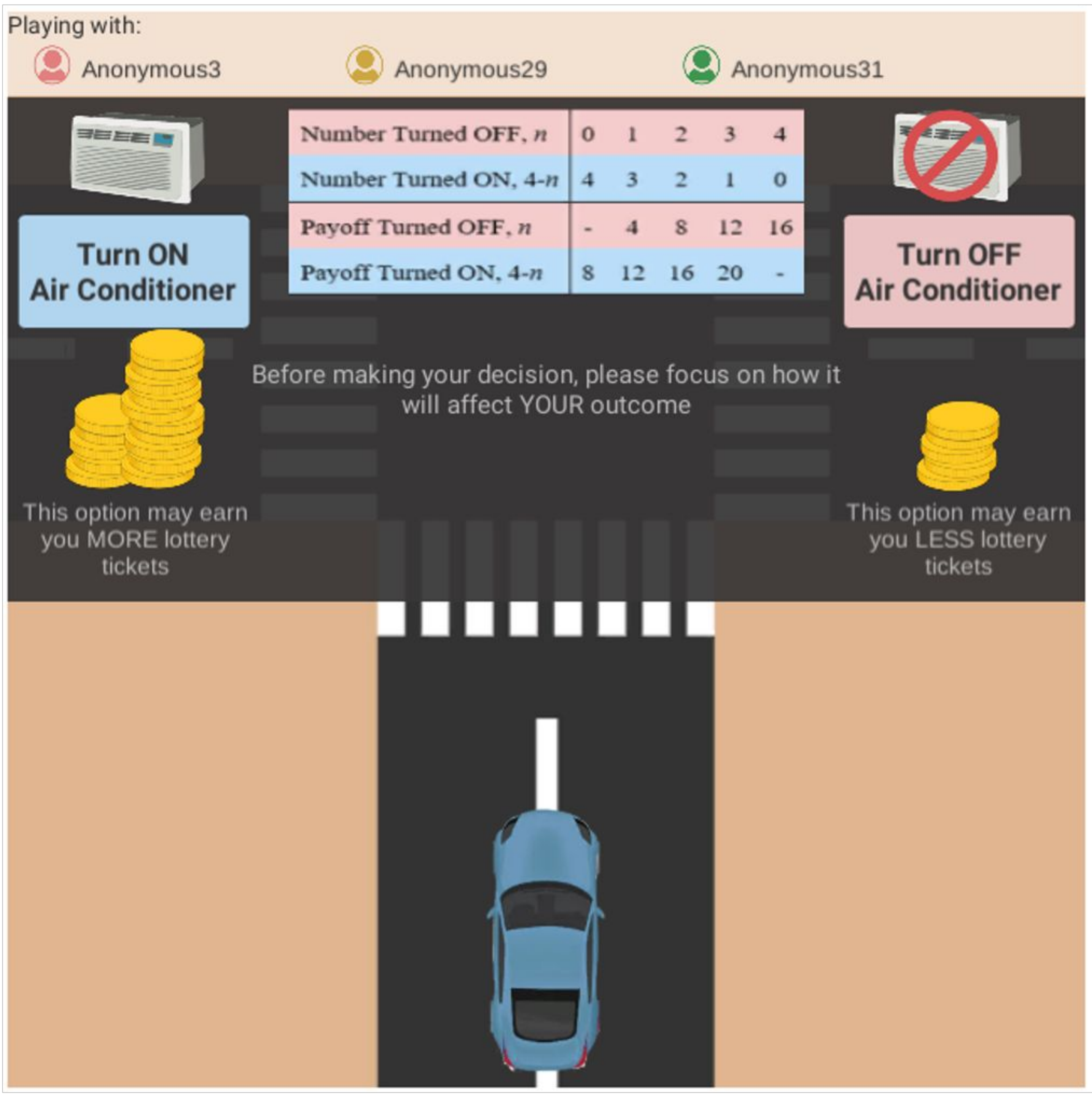
Figure Legends

Fig. 1. The experimental conditions in Experiment 2. (**a**) The condition where participants drive the vehicle directly and are instructed to focus on the outcome for the self, and when the monetary reward is made salient. (**b**) The condition where participants program the vehicle to act on their behalf and are instructed to focus on the outcome for all, and when the environment is made salient.

Fig. 2. The results in Experiment 2. (**a**) Cooperation rates for all conditions. The error bars represent standard errors. (**b**) Analysis of statistical mediation of focus on self, short-term reward saliency, fairness, and high-construal reasoning on the effect of autonomy on cooperation. * $p < .05$.

Fig. 3. Cooperation rates across the 10 rounds in Experiment 3.

a



Playing with:

Anonymous3 Anonymous29 Anonymous31

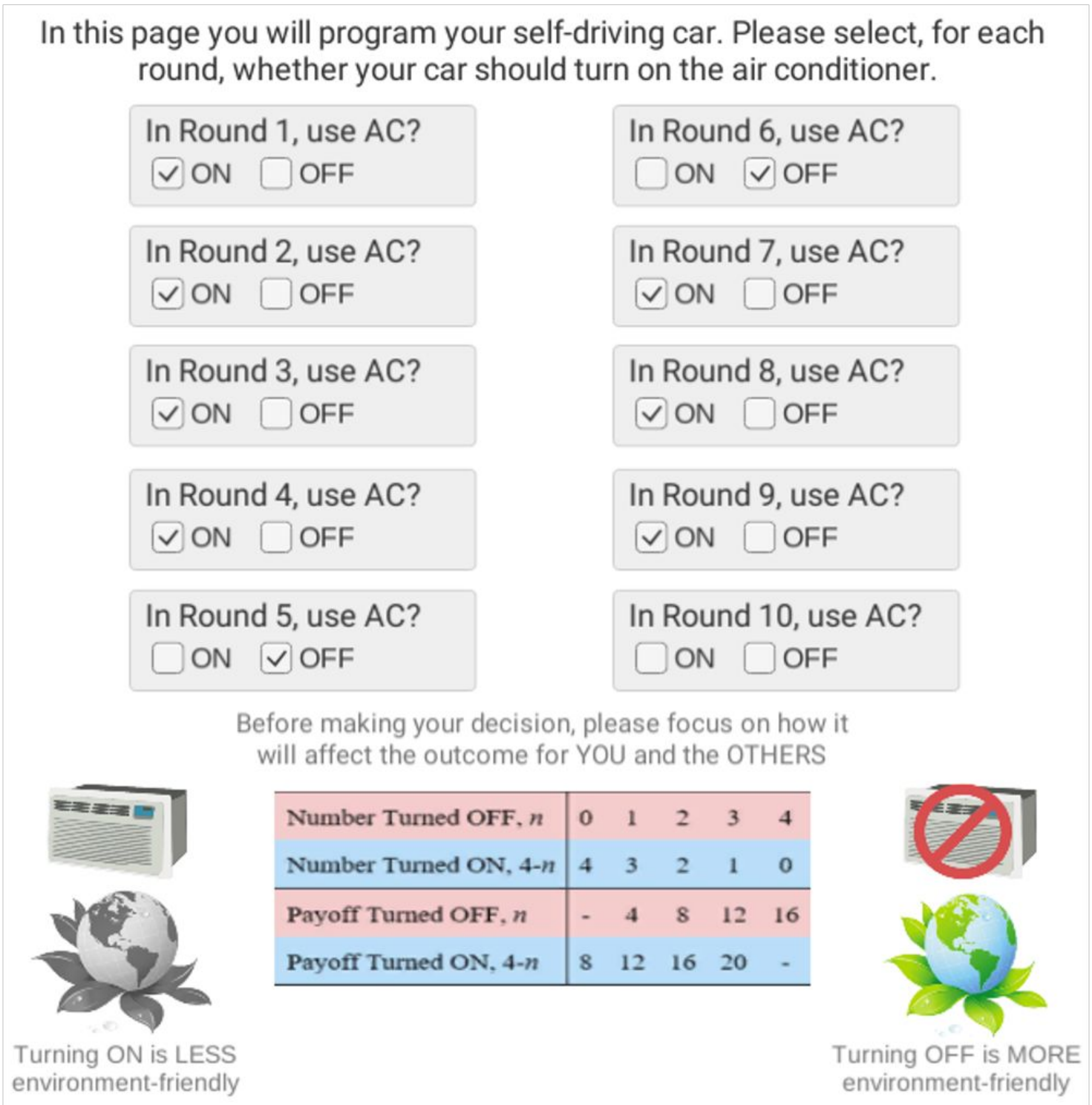
Number Turned OFF, *n*	0	1	2	3	4
Number Turned ON, 4-*n*	4	3	2	1	0
Payoff Turned OFF, *n*	-	4	8	12	16
Payoff Turned ON, 4-*n*	8	12	16	20	-

Before making your decision, please focus on how it will affect YOUR outcome

This option may earn you MORE lottery tickets

This option may earn you LESS lottery tickets

b



In this page you will program your self-driving car. Please select, for each round, whether your car should turn on the air conditioner.

In Round 1, use AC? ☑ ON □ OFF

In Round 2, use AC? ☑ ON □ OFF

In Round 3, use AC? ☑ ON □ OFF

In Round 4, use AC? ☑ ON □ OFF

In Round 5, use AC? □ ON ☑ OFF

In Round 6, use AC? □ ON ☑ OFF

In Round 7, use AC? ☑ ON □ OFF

In Round 8, use AC? ☑ ON □ OFF

In Round 9, use AC? ☑ ON □ OFF

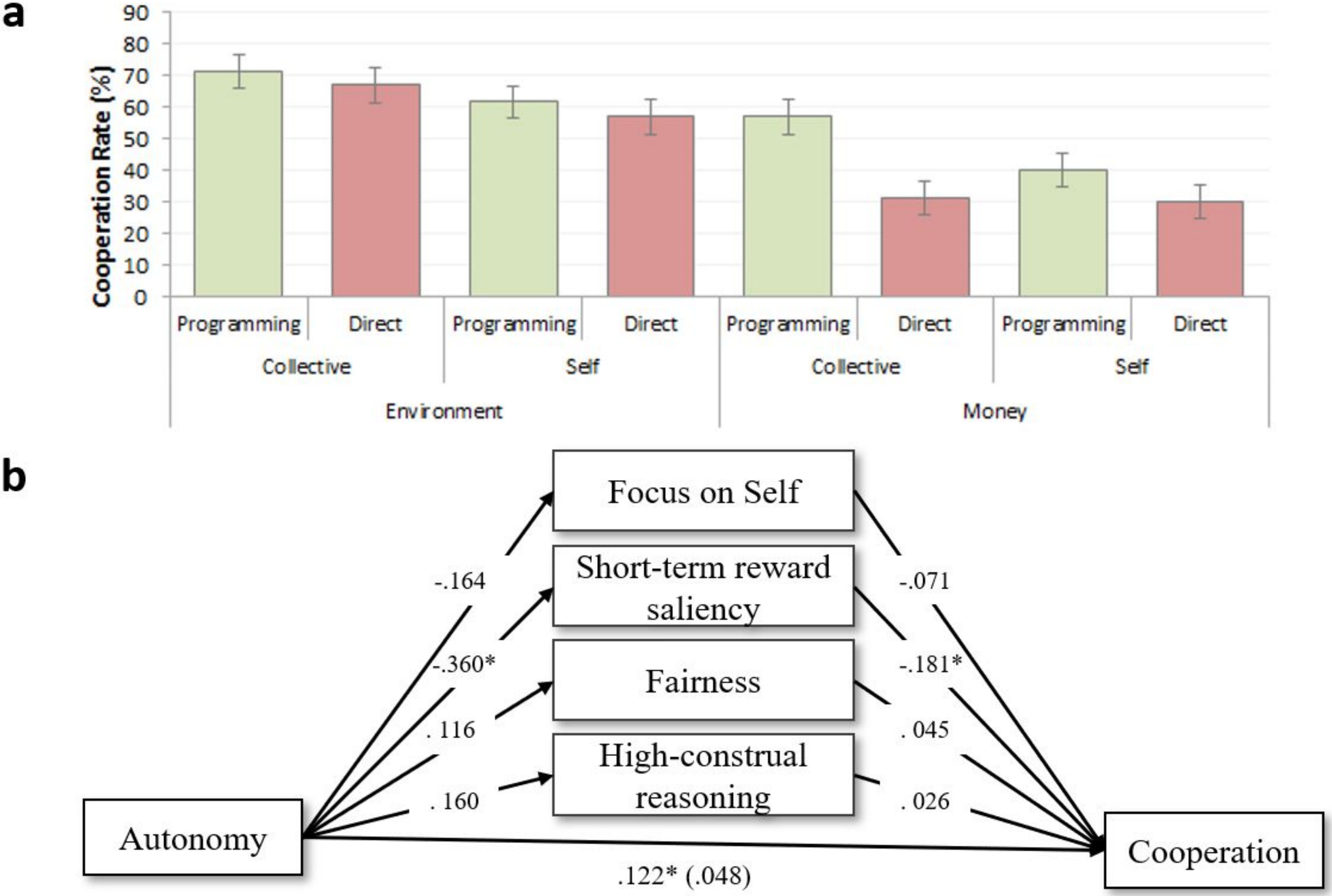
In Round 10, use AC? □ ON □ OFF

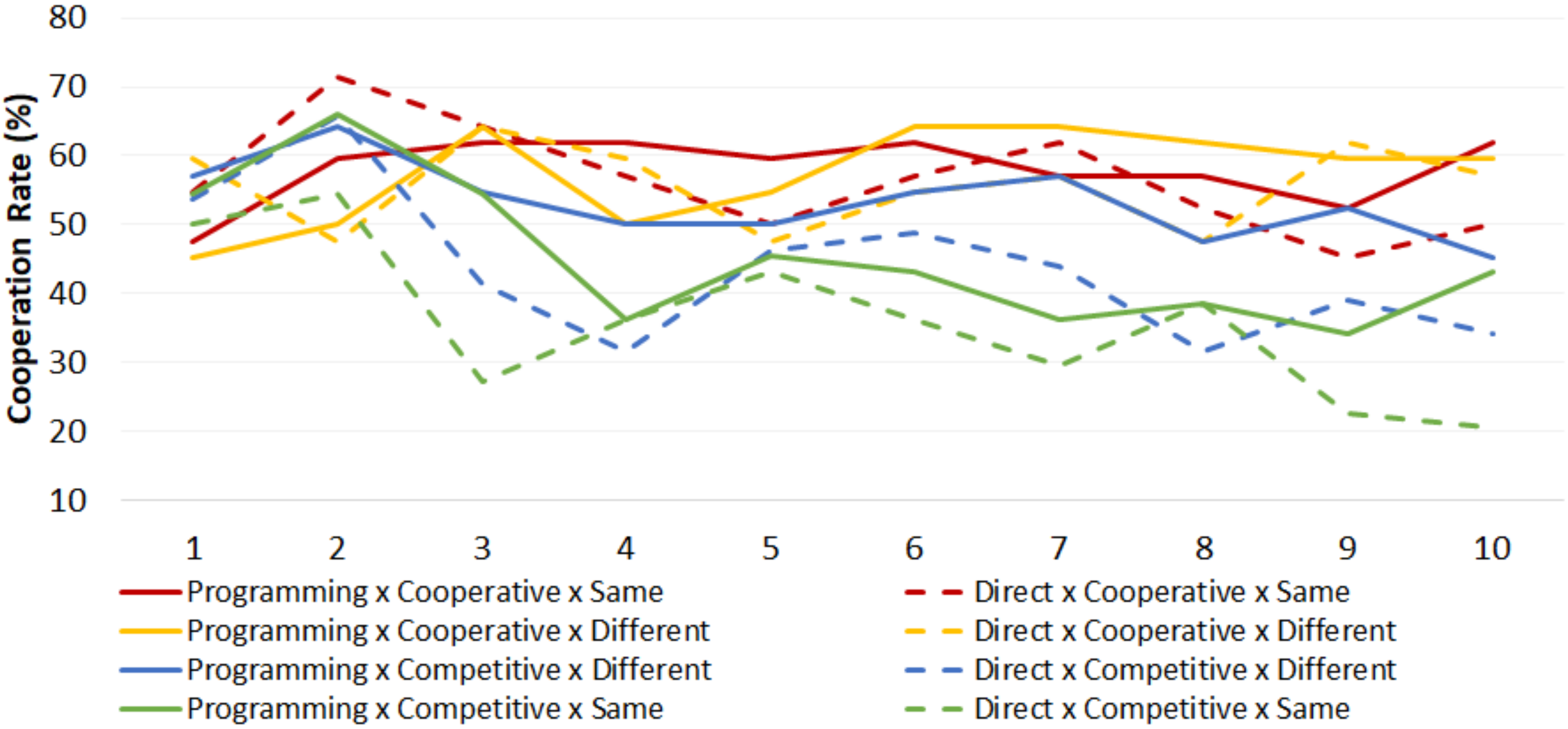
Before making your decision, please focus on how it will affect the outcome for YOU and the OTHERS

Number Turned OFF, *n*	0	1	2	3	4
Number Turned ON, 4-*n*	4	3	2	1	0
Payoff Turned OFF, *n*	-	4	8	12	16
Payoff Turned ON, 4-*n*	8	12	16	20	-

Turning ON is LESS environment-friendly

Turning OFF is MORE environment-friendly


a
Cooperation Rate (%)
0
10
20
30
40
50
60
70
80
90
Programming
Direct
Programming
Direct
Programming
Direct
Programming
Direct
Collective
Self
Collective
Self
Environment
Money
b
Focus on Self
Short-term reward saliency
Fairness
High-construal reasoning
Autonomy
Cooperation
-.164
-.360*
. 116
. 160
-.071
-.181*
. 045
. 026
.122* (.048)


80
70
60
50
40
30
20
10
Cooperation Rate (%)
1
2
3
4
5
6
7
8
9
10
Programming x Cooperative x Same
Programming x Cooperative x Different
Programming x Competitive x Different
Programming x Competitive x Same
Direct x Cooperative x Same
Direct x Cooperative x Different
Direct x Competitive x Different
Direct x Competitive x Same