

# Interacting with Artifacts: Trust and Punishment

Lévan Sardjevéladzé and Edouard Machery

## Abstract

Previous studies have found numerous behavioral and neuropsychological differences between people's interactions with humans and their interactions with computers in social dilemmas. A common explanation of these differences is that we adopt the intentional stance when we interact with humans, but not when we interact with computers. Although this explanation is plausible, in this paper, we provide some evidence that in some contexts, people are willing to adopt the intentional stance in interacting with computers in social dilemmas. As a result, they interact with computers—cooperate with them and, surprisingly, punish them—as they would interact with humans. Using a single-round, extensive form trust game with a punishment stage, we compared the performance of participants who wrongly believed that they were interacting with a human partner and the performance of participants who knew that they were interacting with a computer. In both conditions, participants behaved differently from the predictions of the standard economic model: Players cooperated and punished their partner. Surprisingly, we found a strikingly similar pattern of cooperation and punishment when participants incorrectly believed that they were playing with a human partner and when they knew that they were playing with a computer. Additionally, gender was found to affect participants' decision to trust their partner. We discuss whether and when people adopt the intentional stance in interacting with artifacts such as computers.

**Key words:** theory of mind, artifacts, punishment, trust, cooperation, gender difference

## 1. Introduction

From an early age on, we distinguish between agents and artifacts (e.g., Gergely & Csibra, 2003). We typically explain and predict the behavior of agents by ascribing to them beliefs, desires, intentions or emotions. By contrast, we typically explain and predict the behavior of artifacts by means of other principles. That is, we typically adopt

the intentional stance in interacting with agents, but not with artifacts. Consistent with this fact, psychologists and experimental economists have found that participants behave differently when they interact with computers and with humans in experimental games involving social dilemmas.

### *1.1 Cooperating with Humans and with Computers: Behavioral Differences*

Using prisoner's dilemma games (PDG), psychologists and experimental economists have found that people's tendency to cooperate in social dilemmas is weaker when they are knowingly interacting with computers than when they believe that they are interacting with humans. In a PDG, two players are endowed with some initial monetary endowment. Then, they decide to cooperate or to defect. Their payoff is a function of their choice and of the choice of their partner. The payoffs of a given player are ordered in a decreasing manner as follows: she defects and her partner cooperates, both players cooperate, both players defect, she cooperates and her partner defects. Rational, self-interested players would play the single Nash Equilibrium: Both players would defect. A PDG is repeated if the two players play successive rounds. If the number of rounds is known in advance, rational, self-interested players would defect during the whole game. If the number of rounds is unknown, the folk theorem shows that any strategy that stops cooperating entirely if the partner defects, including a strategy that cooperates with a cooperative partner, but defects for the rest of the game following a defection, is a Nash equilibrium (Fudenberg & Maskin, 1986).

Rilling et al. (2002) examined participants' behavior in a repeated PDG with a known number of rounds. In one condition, participants believed that they were playing against a human player, while in another condition they were told that they were playing against a "preprogrammed computer strategy that does not play a fixed sequence of choices, [but that] responds to [their] choice from earlier round with specified probabilities" (Rilling et al., 2002, 404). They found that mutual cooperation was significantly less common in the latter condition than in the former condition.

Rilling, Sanfey, Aronson, Nystrom and Cohen (2004) also found that in the single-round PDG, participants were significantly more likely to cooperate when they believed that they were playing with a human partner than when they knew that they

were playing with a computer. However, this finding should be interpreted with caution. The two conditions (playing with a human partner and playing with a computer) were different. Participants took part in several games. When participants believed that they were playing with a human partner, it was clear that they were playing with a different partner in each game. By contrast, when participants knew that they were playing with a computer, they might have thought that they were playing with the same computer in all the successive games. Thus, participants might have tried to learn the best strategy to exploit their partner in the latter condition, but not in the former condition. Additionally, the payoffs were different in both conditions. When participants were knowingly playing with a computer, they had a greater incentive to defect.

### *1.2 Punishing Humans and Computers: Behavioral Differences*

Evidence suggests that people are substantially less inclined to punish when they interact with an artifact, such as a computer, than when they interact with a human. In Sanfey Rilling, Aronson, Nystrom and Cohen (2003), participants played an ultimatum game (UG). In this game, player 1 is endowed with some monetary endowment. She chooses whether and how much to share with player 2. Player 2 accepts or rejects player 1's offer. If player 2 accepts the offer, she gets the monetary amount specified by player 1 and player 1 keeps the remaining amount. If player 2 rejects the offer, player 1 and player 2 end up with nothing. If player 2 were purely self-interested, she would accept any positive offer. Expecting this, a self-interested player 1 would make the smallest possible positive offer. In fact, in many cultures, but not in all (Henrich et al., 2005), player 1's modal offer is around 50% of the initial endowment and player 2 rejects one low offer out of two (where a low offer corresponds to 30% of the initial endowment). Rejection is often interpreted as a type of punishment of the unfair offer made by player 1. In Sanfey et al. (2003), participants played the role of player 2. Sanfey et al. (2003) found that low offers made by a human were rejected by participants significantly more often than low offers made by a computer. Rilling et al. (2004) report similar findings.

Using a different game (a trust game, described in section 1.5), de Quervain et al. (2004) have provided convergent evidence. When participants were told that their partner's lack of reciprocation was the result of a random procedure, they reported that

their partner did not behave unfairly. Consistent with their reports, participants reported almost no desire to punish their partner and almost no punishment was imposed.

A clear pattern emerges from the body of findings just reviewed: In social dilemmas, people behave differently when they believe that they are interacting with a human partner and when they believe that they are interacting with a computer. What accounts for these differences? Plausibly, people adopt the intentional stance while interacting with a human partner, but not while interacting with a computer.

### *1.3 Cooperating with Humans and with Computers: Neuropsychological Differences*

Neuropsychological data is consistent with this hypothesis: Brain activation is different when participants are knowingly interacting with a computer and when they believe that they are interacting with a human partner. In McCabe, Houser, Ryan, Smith and Trouard's (2001), participants took part in several games that involved choosing between defection and cooperation (including a trust game).<sup>1</sup> McCabe et al. compared the brain activation of the most cooperative participants to the brain activation of the least cooperative participants. It was found that for the former, the medial prefrontal regions were more strongly activated when they believed that they were playing with a human than when they knew that they were playing with a computer. No difference was found for the least cooperative participants. To explain these findings, McCabe et al. proposed that when playing with a computer or when playing non-cooperatively, participants adopted simpler strategies than when they cooperated with human partners.

Rilling et al. (2002) compared the brain activation of participants who incorrectly believed that they were playing a repeated PDG with a human partner and the brain activation of participants who knew that they were playing with a computer. They found that mutual cooperation in the latter condition activated a neural network that overlapped with the neural network activated by mutual cooperation in the former condition. In both conditions, the ventromedial and orbital cortex was activated. However, in the former condition, but not in the latter condition, the rostral anterior cingulate and the anteroventral striatum were activated. Rilling et al. link the ventromedial cortex and the anteroventral striatum to reward processing and argue that the activation of these two

---

<sup>1</sup> McCabe et al. (2001) did not report their behavioral findings.

areas result from the satisfaction resulting from mutual cooperation. Consistent with Rilling et al.'s (2002) finding that participants cooperate less with computers than with humans (see above), the neural network involved in reward processing was more activated when participants believed they were interacting with a human than when they knew they were interacting with a computer.

Rilling et al. (2004) found that a neural network linked to mindreading, including the right posterior temporal sulcus and the anterior paracingulate cortex, was activated when participants in a UG and in a single-round PDG believed that they were playing with a human partner.<sup>2</sup> In the PDG, when participants knew that they were playing with a computer, an overlapping neural network was activated, including the anterior paracingulate cortex and the right posterior superior temporal sulcus. However, activation in these areas was stronger when participants believed that they were playing with a human than when they knew that they were playing with a computer.<sup>3</sup> In the UG, activation was weak when participants knew that they were playing with a computer. Consistent with Rilling et al.'s (2004) finding that participants cooperated more when they believed that they were interacting with a human than when they knew that they were interacting with a computer, the neural network involved in mindreading was less activated in the latter condition than in the former.

#### *1.4 Punishing Humans and Computers: Neuropsychological Differences*

In Sanfey et al. (2003), brain activation was measured when participants, who played player 2 in a UG, were told how much money they were offered. They found that the activity of the bilateral anterior insula, which is associated with disgust and anger, was lower for low offers, when participants knew that the unfair offer was made by a computer program, than when participants believed the unfair offer was made by a human. Consistent with Sanfey et al.'s (2003) finding that participants punished less when they knew that they were interacting with a computer than when they believed that

---

<sup>2</sup> But see Saxe, Carey and Kanwisher (2004) on the anterior paracingulate cortex

<sup>3</sup> Using a zero-sum game ("stone, paper, scissors"), Gallagher, Jack, Roepstorff and Frith (2002) found similarly that the bilateral anterior paracingulate cortex was less activated when participants knew that they were playing with a computer than when they incorrectly believed that they were playing with a human partner.

they were interacting with a human (see above), the brain areas associated with negative emotions were less activated in the former case than in the latter case.

De Quervain et al. (2004) also found that the activation of the caudate nucleus was lower than the baseline (mean activation of the caudate nucleus in the experiment), when participants were told that their partner's lack of reciprocation was the result of a random procedure. De Quervain et al. (2004) note that the caudate nucleus is involved in processing the rewards resulting from intentional actions. Consistent with de Quervain et al.'s (2004) finding that participants punish less when they know that they are interacting with a computer than when they believe that they are interacting with a human, participants seemed to have anticipated less satisfaction from punishing in the former condition than from punishing in the latter condition.

To summarize, behavioral findings show that in social dilemmas, people behave differently when they interact with computers and with humans. People tend to be less cooperative and less punitive in the former case than in the latter case, plausibly, because they adopt the intentional stance in the former case, but not in the latter. Neuropsychological findings are consistent with this explanation. When interacting with a computer, brain areas involved in control, mindreading and reward processing are either not activated or less activated than when participants are interacting with a human partner. Moreover, brain areas involved in negative emotions are less activated when playing with a non-cooperative computer than when playing with a non-cooperative human. These neuropsychological findings constitute a coherent pattern. Plausibly, because people do not adopt the intentional stance while interacting with computers, they rely on simpler strategies to decide how to behave, they feel no (or less) anger when their computer partner fails to cooperate and they anticipate no (or less) reward from punishing their partner.

### *1.5 Adopting the Intentional Stance in Interactions with Computers*

Previous studies have highlighted the differences between our interactions with humans and with computers. We contend that these studies paint only a partial picture of these interactions. Previous studies cast an incomplete light on how people interact with computers because of the way computers were described to participants. In most studies,

participants were told how the computer would behave (McCabe et al., 2001; Rilling et al., 2002, 2004). For instance, in Rilling et al. (2004), participants to the UG were told that computer offers were randomly generated. In other studies, participants were told that their computer partner would behave in a preprogrammed way (Rilling et al., 2002, 2004). For instance, in Rilling et al. (2004), participants to the PDG were told that computer decisions would be made “according to programmed probabilities.”<sup>4</sup> These elements of information might prevent participants to adopt the intentional stance while interacting with a computer, because the computer is characterized as a non-intelligent entity. It is unclear why people would need to use their mindreading capacities in interacting with a computer, if they have been told that it behaves randomly or according to fixed probabilities. If participants do not represent their computer partner in intentional terms, they are unlikely to feel any moral pressure to cooperate with it. This would explain why they cooperate less. Moreover, they are unlikely to feel any negative emotion toward non-cooperative computer partners and to feel any satisfaction in anticipating punishing them. This would explain why they punish less when they know that they are interacting with a computer than when they believe that they are interacting with a computer.

In our experiment, participants were told neither how the computer would behave nor that its behavior had been preprogrammed. Instead, they were told that they would be interacting with “a computer endowed with an artificial intelligence program.” We hypothesize that this piece of information might prime participants to adopt the intentional stance when interacting with a computer. Following Dennett (1978, 1987), we propose that people adopt the intentional stance to decide how to interact with artifacts, when the behavior of these artifacts, is in some way, intelligent. In such cases, people behave with respect to artifacts as they would with respect humans. Thus, if people adopt the intentional stance in interacting with computers in social dilemmas, they should behave similarly when knowingly playing with a computer and when believing that they are playing with a human partner.

To test the hypothesis that people adopt the intentional stance in order to decide how to interact with intelligent computers, we ran a single-round, extensive-form trust

---

<sup>4</sup> It is unclear what participants in Sanfey et al.’s (2003) experiment were told.

game with a punishment stage (Figure 1). To study participants' interaction with computers in this trust game, participants were assigned to one of two conditions. In the human player condition, participants incorrectly believed that they were playing with a human player, while in the computer condition, they knew that they were playing with a computer. In both cases, they played with the same computer program.

Our version of the trust game involved two players, player 1 and player 2. Participants always played player 1. Player 1, but not player 2, was endowed with some initial monetary endowment (10 écus).<sup>5</sup> In the first stage of the game, participants were invited to give a fraction of their endowment to player 2. The monetary amount given by the participant ("the offer") was tripled and was given to the second player. For instance, if a participant gave 2 écus, player 2 received 6 écus and the participant ended up with 8 écus (her initial endowment of 10 écus minus 2 écus). Moreover, participants were invited to ask player 2 to give back a fraction of the amount of money player 2 received ("the request").

In the second stage of a typical trust game, player 2 decides whether and how much she wants to give back to player 1. The monetary amount that is given back is added to player 1's remaining monetary endowment (without being tripled). In the two conditions of our version of the trust game, player 2 gave back nothing.

In the last stage of the game, participants were given the opportunity to punish their partner at their own cost. The cost of punishment was set at half an écu for each écu taken from player 2.

Put Figure 1 about here.

If it were common knowledge that the two players in a trust game with a punishment stage were rational, self-interested agents, player 1 would not punish in the third stage of the game, because the game involves a single round. Thus, in the first stage of the game, player 1 would expect player 2 to view the threat of punishment as empty. As a result, player 1 would expect player 2 to give back no money in the second stage of the stage. Thus, in the first stage of the game, player 1 would not trust player 2 and would not offer any money to player 2. Yet, if player 2 had been known to be trustworthy, player 1 would have offered some money and both players would have been better off.

---

<sup>5</sup> See section 2 on what écus are.

In fact, in single-round trust games without punishment, 50% of the participants playing player 1 trust their partner (first right option in Figure 1) and 75% of the participants playing player 2 reciprocate (Berg, Dickhaut & McCabe, 1995; McCabe & Smith, 2000; McCabe et al., 2001; Fehr & Rockenbach, 2003). In previous single-round trust games with punishment, participants tended to trust their partners and to punish her if she failed to reciprocate (de Quervain et al., 2004).

In line with previous findings about people's behavior in the trust game, we predicted that participants' behavior would differ from the behavior predicted by the standard economic model. That is, we predicted that when they incorrectly believed that they were interacting with a human partner, participants' decisions to cooperate and to punish would differ from the decisions a rational, self-interested agent would make.

However, in contrast with previous findings about the behavioral differences between people's interactions with humans and with computers, we also predicted that there would be no difference between participants' behavior in our two conditions: That is, we predicted that participants would cooperate and punish similarly when they incorrectly believed that they were interacting with a human partner and when they knew that they were interacting with an intelligent computer. The reason is that we agree with Dennett that people are disposed to adopt the intentional stance when they interact with intelligent artifacts.

We were also interested in finding out whether gender might affect participants' cooperative and punishing behavior. A trust game involves taking some financial risk and gender is known to modulate risk aversion in some contexts. Thus, we examined whether males would adopt a more risky financial behavior, that is, whether they would offer a larger fraction of their initial endowment to their partner.

## **2 Method**

### *2.1 Participants*

Participants were players in the French-speaking on-line virtual reality game, *Les Royaumes Renaissants* ([www.lesroyaumes.com/](http://www.lesroyaumes.com/)).<sup>6</sup> Participants were recruited, when they logged in the game. They were asked whether they were interested in taking part in an experiment. If they accepted this invitation, they were sent to a different website. The design of this website made it clear that they had left the on-line game. 1,415 participants were thus recruited. On the website of the experiment, the structure of the experiment was described to participants. Participants were then asked a series of questions. Participants who failed to answer correctly these questions or who answered too slowly did not take part in the experiment. Thus, we controlled that participants understood the experiment and were motivated to take part in the experiment. Out of the 1,415 volunteers, 723 participants answered the questions correctly and in a timely manner. Out of these 723 participants, 369 participants took part in the experiment described in this article (74.8% male; mean age: 26;1 (SD = 11;8); age range: 11-65).<sup>7</sup>

## 2.2 Procedure

Participants were randomly assigned to the human player condition or to the computer condition. 206 participants were assigned to the human player condition and 163 to the computer condition. In the human player condition, participants were told that they were paired with another player from the on-line game *Les Royaumes Renaissants*. In the computer condition, participants were told that they were paired with “a computer endowed with an artificial intelligence program.”<sup>8</sup> Importantly, contrary to other experimental games involving interactions with computers (e.g., McCabe et al., 2001; Rilling et al., 2002, 2004), participants were not told how the computer would play. In both conditions, participants were in fact paired with the same computer program.

Participants played a single round of the extensive-form trust game with a punishment stage (Figure 1). They knew that they would play a single game. They always played the role of player 1.

---

<sup>6</sup> *Les Royaumes Renaissants* is a roleplaying and strategy game that simulates a society of 20,000 individuals in a stable environment. It is edited by *Celsius Online*. An English version of the game can be seen at Renaissance Kingdoms, [www.renaissancekingdoms.com](http://www.renaissancekingdoms.com).

<sup>7</sup> We analyzed the effect of age on participants' offer, request or punishment by dividing participants into four equal age groups (below or equal to 17, from 17 to 23, from 23 to 31, and above 31). We found that age did not affect participants' offer, request or punishment.

<sup>8</sup> In French: “un ordinateur muni d'un programme d'intelligence artificielle.”

Participants were endowed with an initial endowment of 10 “écus.” “Ecus” is the money used in the virtual reality game, *Les Royaumes Renaissants*. Although écus is not a real money, it is not without value. Players of the game *Les Royaumes Renaissants* buy écus at the rate of approximately 2 euros for 10 écus. Participants were told that they would add the amount of écus earned in the game to their account in the on-line game *Les Royaumes Renaissants*. Thus, the initial endowment was equivalent to two euros. Participants were also told that Player 2 was not given any endowment.

In the first stage of the game, participants were first invited to give a fraction of their endowment to player 2. Participants knew that the monetary amount given by the participant (“the offer”) would be tripled and given to player 2. Once they had made their decision, participants were invited to ask player 2 to give back a fraction of the amount of money player 2 had just received (“the request”). In the second stage of the game, participants were told that their partner had given back nothing. In the last stage of the game, participants were asked whether they wanted to punish their partner at their own cost.<sup>9</sup> They were reminded of the cost of punishment (half an écu for each écu taken from another player). Participants were then invited to leave a comment about the experiment.

### 3. Results and Discussion

#### 3.1 Participants’ Offer

The first dependent variable of interest is participants’ offer to player 2 in the first stage of the game. In the human player condition, participants’ mean offer was 6.23 écus (SD = 3.05). In the computer condition, participants’ mean offer was 6.09 écus (SD = 3.04; see Table 1). Both offers are significantly different from the offer a self-interested player would make, provided she believed she was playing with a rational self-interested player—viz. 0 (human player condition:  $t(205) = 29.35$ ,  $P < .001$ , two-tailed; computer condition:  $t(162) = 25.56$ ,  $P < .001$ , two-tailed).

Put Table 1 about here.

---

<sup>9</sup> The punishment stage of our experiment differs from the punishment stage of Fehr and Rockenbach’s (2003) experiment. In Fehr and Rockenbach’s experiment, participants had to commit themselves to punish their partner if she failed to give back some amount of money. Punishment was thus a credible threat.

An ANOVA with condition (human player versus computer) and gender (male versus female) as factors, yielded a main effect of gender (male > female,  $F(1,365) = 5.47$ ,  $P < .05$ ; Figure 2), and no other main effect or interaction.

Put Figure 2 about here.

Importantly, there was no main effect of condition ( $F(1,365) = .001$ ,  $P > .9$ ; Figure 3): Participants did not give a significantly different amount of écus when they wrongly believed that they were playing with a human partner by comparison to when they knew that they were interacting with a computer.

Put Figure 3 about here.

Using Buchner and colleagues' G\*Power software (Buchner, Erdfelder, & Faul, 1997), we calculated post-hoc the power of our test. Assuming a medium effect size in the population (Cohen, 1992) and an  $\alpha = .05$ , power was superior to .99, meaning that the probability of rejecting the null hypothesis given a substantial population effect was superior to .99.

### *3.3 Participants' Request*

The second dependent variable of interest is participants' request to the second player at the end of the first stage of the game. In the human player condition, participants' mean request was 9.14 écus ( $SD = 6.39$ ). In the computer direction, participants' mean request was 10.38 écus ( $SD = 7.22$ ; see Table 1). To analyze participants' request, we calculated Pearson's correlation between the request and the offer times 1.5. Suppose that participants thought that they and their partner should get the same amount of money from their interaction. Thus, if a participant gave 5 écus, she would expect her partner to give back 7.5 écus. If this captures participants' expectation, the request and the offer times 1.5 should be correlated. We found indeed that in the human player condition and in the computer condition, the request was significantly correlated with the offer times 1.5 (human player condition,  $r = .87$ ,  $P < .001$ ; computer condition,  $r = .88$ ,  $P < .001$ ).

An ANOVA with condition (human player versus computer) and gender (male versus female) as factors, yielded a main effect of gender (male > female,  $F(1,365) = 5.84$ ,  $P < .05$ ; Figure 4), and no other main effect or interaction.

Put Figure 4 about here.

Importantly, there was no main effect of condition ( $F(1,365) = 2.13, P > .1$ ; Figure 5): Participants did not request a significantly different amount of écus when they wrongly believed that they were playing with a human partner by comparison to when they knew that they were interacting with a computer.

Put Figure 5 about here.

### *3.4 Participants' Punishment*

The last dependent variable of interest is the monetary amount participants invested in punishing their partner. We scored this variable as a percentage of the remaining amount of money—i.e., of the initial monetary endowment (10 écus) minus the offer made to the other player in the first stage of the experiment. For instance, if a participant, who had given 2 écus in the first stage of the experiment, decided to invest 2 other écus in punishment, her investment in punishment was scored as 25%. We think that that this way of scoring the amount invested in punishment better reflects the decision each participant has to make: Given that  $x$  écus remains to a participant, how much should she punish her partner? We note that analyzing the absolute amount of écus invested in punishment, rather than the relative amount, does not change the finding presented below.

In the human player condition, participants' mean relative punishment was 30.14% ( $SD = 39.86$ ). In the computer condition, participants' mean relative punishment was 30.44% écus ( $SD = 39.46$ ; see Table 1). In both conditions, punishment was significantly different from the punishment a self-interested player would make—viz. 0% (human player condition:  $t(205) = 10.85, P < .001$ , two-tailed; computer condition:  $t(162) = 9.85, P < .001$ , two-tailed).

An ANOVA with condition (human player versus computer) and gender (male versus female) as factors, did not yield any main effect or any interaction. Most important, there was no main effect of condition ( $F(1,365) = .059, P > .8$ ; Figure 6). Surprisingly, participants did not punish differently their partner when they wrongly believed that they were playing with a human partner by comparison to when they knew that they were interacting with a computer.

Put Figure 6 about here.

### 3.5 Comments

We analyzed qualitatively the comments left by participants. Roughly, one participant out of three left a comment about the game (37% in the human player condition and 34% in the player condition). Consistent with the correlation between participants' request and their offer times 1.5, several participants explicitly stated that they expected to get the same amount of money as their partner from the money invested in the game. A few quotations might illustrate this point. A participant in the human player condition wrote: "I tried to play fair, that is, to play WITH my partner in such a way that we each win the maximum (...)." <sup>10</sup> Another player from the same condition wrote: "Morality entails that my partner should give back half. We would have had each 15 écus." <sup>11</sup> A participant from the computer condition wrote: "I chose the case with higher return that ensures equality (...)." <sup>12</sup>

Moreover, participants justified differently their decision to punish when they wrongly believed that they were playing with a human partner by comparison to when they knew that they were interacting with a computer. Comments mentioned negative emotions such as anger and disgust against player 2 more often when participants believed that they were interacting with a human partner than when they believed they were interacting with a computer. 13% of the comments made by participants in the human player condition referred to anger or disgust against player 2 while only 2% of the comments made by participants in the computer condition referred to anger or disgust against player 2. Participants were also more likely to refer to moral considerations in the human player condition than in the computer condition.

The difference between the verbal reports made by participants in the human player condition and in the computer condition clearly shows that participants did not mistake computers for humans, even though they cooperated with computers and punished them.

---

<sup>10</sup> In French: "J'ai tenté de jouer 'réglo', c'est-à-dire de jouer AVEC mon partenaire afin que nous gagnions chacun le maximum (...)." <sup>10</sup>

<sup>11</sup> In French: "La moralité vaut [sic] que mon partenaire me rende la moitié. Nous aurions eu chacun 15 écus (...)." <sup>11</sup>

<sup>12</sup> In French: "ce que j'ai choisis, c'est le cas avec le plus haut rendement qui garantisse l'égalité (...)." <sup>12</sup>

### *3.6 Discussion*

Focusing first on the human player condition, our findings show that in a single-round trust game with a punishment stage, participants were willing to trust their partners, although a self-interested player who believed that she was interacting with another self-interested player would not have given any money to her partner. Participants gave around 60% of their initial monetary endowment. We found an unexpected gender difference: Male participants gave significantly more than female participants. Thus, male participants displayed more trust than female participants.

Participants in the human player condition expected their partner to give back a substantial amount of money. Participants' request was significantly correlated with their offer times 1.5, tentatively suggesting that participants expected that they and their partner should get the same amount of money from their interaction. The comments left by several participants are consistent with this interpretation of the results. Male participants expected their partner to give back a larger amount of money than female participants. This was to be expected, since male participants offered their partner a larger amount of money.

Participants in the human player condition were also willing to punish their partners at their own cost, although they knew that they would not interact again with their partner. On average, participants invested around 30% of their remaining amount of money (initial endowment minus offer in the first stage of the game), even though they knew that they could not derive any benefit from punishing. This finding replicates de Quervain et al's (2004) findings.

Turning now to the comparison between the human player condition and the computer condition, we found that in a single-round trust game with a punishment stage, participants did not behave differently when they incorrectly believed that they were playing with a human partner and when they knew that they were playing with an intelligent computer.

Participants gave a similar amount of money to their partner and asked for a similar amount of money in return, when they incorrectly believed that they were playing with a human partner and when they knew that they were playing with an intelligent computer. This finding is at odds with the body evidence that people cooperate less when

they know that they are interacting with a computer than when they believe that they are interacting with a human partner (Section 1). This finding tentatively suggests that when people are told that they will interact with “a computer endowed with an artificial intelligence program,” as the participants in this experiment were told, they calibrate their interaction with the computer on how they would interact with a human partner.

More puzzling, participants punished similarly when they incorrectly believed that they were playing with a human partner and when they knew that they were playing with a computer. This finding is at odds with previous findings that people do not punish when they interact with a computer (Sanfey et al., 2003; de Quervain et al., 2004).

#### **4. General Discussion**

##### *4.1 Trust and Costly Punishment in Human-Human Interactions*

Our findings provide further evidence that the standard economic model is descriptively inadequate. Previous studies in experimental economics have shown that people are willing to trust other people, in situations where the self-interested agent assumed by the standard economic model would not be trustful (e.g., Berg et al., 1995; McCabe & Smith, 2000; Fehr & Rockenbach, 2003). Our data are consistent with these findings: People are disposed to trust strangers, when trusting strangers might be beneficial.

Additionally, people seem to expect to be recompensed for their trustful behavior. To determine what a fair recompense would be, they seem to follow a simple heuristic: Participants seem to have expected that they themselves and their partner would get an equal amount of money from the amount invested in the interaction. For instance, if a participant had given 2 écus to player 2, she typically expected player 2 to give back 3 écus—so that each player would gain 3 écus from their interaction.

A large body of evidence has also shown that people are willing to punish norm violators at their own cost, including when they know that they cannot derive any benefit from punishing (Fehr & Gächter, 2000, 2002; Fehr & Fischbacher, 2003; Fehr & Henrich, 2003; Henrich et al., 2006). Our experiment provides convergent evidence for the existence of a disposition to punish at one’s own cost, independently of the benefits one might derive from punishing.

Finally, we found that males were more likely than females to trust their partner. To our knowledge, this is the first report of such a gender difference in trust games. Additional studies are certainly needed to confirm this finding. We speculate that the gender difference may result from the fact that males have a lower risk aversion than females at least in some contexts of financial decision making (Jianakoplos & Bernasek, 1998), if not in all contexts (Schubert, Brown, Gysler & Brachinger, 1999). If males really are less risk averse than females, they might be willing to bet that the partner that they are paired with will be trustworthy.

#### *4.2 Taking the Intentional Stance in Human-Computer Interactions*

Our experiment also casts a new light on how people interact with potentially intelligent artifacts such as computers. We found that participants showed a similar level of trust when they wrongly believed that they had been paired with a human player and when they knew that they had been paired with a computer. A plausible interpretation of this finding is that our participants adopted the intentional stance in order to decide whether to trust a computer that had been described as being “endowed with an artificial intelligence program.” That is, they treated the computer as if it had beliefs and desires and as if it acted rationally on these beliefs and desires (Dennett, 1978, 1987). More precisely, participants treated the computer as if it had the beliefs and desires a human partner would have in this situation. Thus, at the first stage of our extensive-form trust game with a punishment stage, participants in the computer condition expected their computer partner to react to a trusting offer as a human partner would react. As a result, participants offered the same amount of money to their computer partner as they would have offered to a human partner.

Previous studies of human-computer interactions had little to say about the adoption of the intentional stance in interaction with computers, because participants were told how the computer would behave (McCabe et al., 2001; Rilling et al., 2004). Thus, participants did not have to adopt the intentional stance in order to anticipate the behavior of their computer partner.

Neuropsychology provides some consistent evidence with the hypothesis that people adopt the intentional stance in some interactions with computers. As we saw in

Section 1, in comparing brain activation of participants who wrongly believed that they were playing with a human player and brain activation of participants who knew that they were playing with a computer in a UG and in a sequential PDG, Rilling et al. (2004) found that the neural network involved in mindreading, which involves the posterior superior temporal sulcus and the anterior paracingulate cortex, was activated in both groups of participants (but to a lesser extent when participants interacted knowingly with computers). This tentatively suggests that people might be disposed to adopt an intentional stance toward computers, as they would toward humans, but that people might refrain from doing so when they are told how their computer partner will behave.

It is worth noting that adopting the intentional stance when interacting with intelligent computers is not tantamount to mistaking computers for human agents. The comments left by participants were different when participants knew that they were interacting with computers and when they incorrectly believed that they were interacting with humans. One can treat an artifact as an intelligent partner, without believing that this artifact literally has desires and beliefs (Dennett, 1978, 1987).

#### *4.3 Punishment in Human-Computer Interactions*

Contrary to previous results (Sanfey et al., 2003; Riling et al., 2004), we found that participants punished their partner even when they knew it was a computer. Moreover, the amount of money invested in punishment was not significantly different when participants incorrectly believed that they were playing with a human partner and when they knew that they were playing with a computer (Figure 6). What accounts for the difference between the current finding and previous results? We focus on this issue in the remainder of this article.

As we saw above, it makes sense to interact with an intelligent computer in the first stage of the trust game as one would interact with a human partner. Participants in the computer condition might have expected that if their computer partner was psychologically similar to humans, it might respond to a trusting offer by giving back approximately half of the tripled offer. It is much more puzzling that participants in the computer condition did punish their computer partner as they would have punished a human partner. Since they were playing a single-ground trust game, they knew that their

computer partner could not respond in any way to their punishment. Thus, in the last stage of the experimental game, it seems irrational to treat the computer partner as if it were a human partner. The finding that in some contexts, people are willing to punish computers at their own cost even though they know that they cannot derive any benefit from punishing is an anomaly: It is, at least *prima facie*, at odds with what reason would prescribe in these circumstances.

Before going any further, we note that the instructions used in the experiment reported in this article were very similar to the instructions used by other researchers in experiments on punishment in behavioral economics and in neuropsychology. Thus, the difference between our finding and previous results does not result from different instructions. Additionally, it is unlikely that we failed to find any difference between the human player condition and the computer condition because of a lack of statistical power, since the power of the ANOVA test applied to participants' offer was higher than .99. Moreover, we found a statistical difference between male participants' and female participants' offers and requests and, in the human player condition, we replicated previous findings with the trust game.

To explain our finding, one could propose that our participants were drawn from an atypical population. Participants are a subset of the players in the on-line game *Les Royaumes Renaissants*. One might speculate that players in on-line games have a specific type of personality. One might even speculate that they tend to interact with computers as they do with humans.

We are not convinced by this interpretation of our finding. Players in the game *Les Royaumes Renaissants* are not so-called "hard-core gamers." Moreover, participants were not confused about the nature of their partner in the computer condition. As noted above, participants' comments were different across the two conditions. Several participants in the human player condition reported emotional reactions, particularly anger and disgust, and morally condemned the behavior of their partner. Participants in the computer condition much more rarely made such reports.

One could also explain our finding by arguing that punishment was not really costly in the experiment reported in this article. Participants were given an initial endowment in écus—the money used in the game *Les Royaumes Renaissants*. Thus,

participants' offers were made in écus and participants invested écus in punishment. Because écus is not a real money, participants might have viewed it as being without value. This might explain why they were willing to punish a computer.

We do not think that the difference between our finding and previous results can be so explained. Écus play an important role in the on-line game *Les Royaumes Renaissance*: A player's wealth determines to a large extent what she can do. Players in the game *Les Royaumes Renaissance* typically buy écus to carry out their projects. Participants knew that the amount of écus earned in the game would be added to their account in the on-line game *Les Royaumes Renaissance*. Since all the participants were players in the game *Les Royaumes Renaissance*, we doubt that they viewed écus as being without value. Moreover, the fact that participants did not give their whole endowment to their partner in the first stage of the trust game suggests that they did not view écus as being without value. Finally, participants in the human partner condition behaved as they do in similar experiments involving real money (e.g., de Quervain et al., 2004).

We think that the most plausible explanation of participants' disposition to punish their computer partner is the following. Converging evidence shows that emotions, including disgust and anger, play a crucial role in the triggering of the decision to punish other individuals. Kahneman, Schkade, and Sunstein (1998) have shown that participants' judgment about how much an action is "outrageous" correlates almost perfectly with how much participants think the agent should be punished (see also Carlsmith, Darley, & Robinson, 2002). Self-reports support these findings. Fehr and Gächter (2002) asked participants to imagine that they were investing to a common fund in a public goods game, while another participant to this game failed to invest. Participants reported that they would be angry against the non-contributor (see also Pillutla & Murnighan, 1996). Neuropsychological evidence provides some additional evidence for the role of negative emotions in decision-making about punishment. Sanfey et al. (2003) found that low offers increased activity in the bilateral anterior insula, which is associated with disgust and anger. Additionally, participants' mean activity of the anterior insula correlated with the percentage of offers they rejected. These results strongly suggest that the process leading to the decision to punish is, at least to some

extent, based on our negative emotions. When someone behaves unfairly, we experience a negative emotion, such as anger, which influences our decision to punish.

This process seems to be inhibited when people do not adopt the intentional stance in social dilemmas, for instance when people interact with a computer after being told how the computer will behave. Anterior insula activation is weaker when people are receiving unfair offers from computers in an UG (Sanfey et al., 2003). Because in previous experiments on interactions with computers in social dilemmas, participants did not adopt the intentional stance, the decision-making process leading to punishment might have been inhibited, which might explain why people decided not to punish.

We propose that because participants in the computer condition adopted the intentional stance to decide whether or not to trust their computer partner in the first phase of the trust game, the emotion-based decision-making process leading to punishment was not inhibited. As a result, participants in the computer condition punished as they would have done if they had been interacting with a human partner.

One could wonder why, if the explanation proposed here is correct, participants gave different justifications when they punished a human partner by comparison to when they punished a computer partner. Following Haidt (2001), we propose that the decision to punish and the justification of one's punishment decision tap into two different mechanisms. The decision to punish results from a decision-making process that involves negative emotions, such as anger, outrage and, maybe, disgust. The justification of a decision to punish taps into explicit beliefs about what behaviors are appropriate. Because there is a norm prohibiting unfair offers in Western cultures, participants in the human partner condition often justified their punishment decision by referring to the morally inappropriate behavior of partner 2 and to their moral emotions. Because lashing out at artifacts and other non-intentional entities is socially inappropriate in Western cultures, participants in the computer condition rarely justified their punishment decision by referring to moral considerations and to emotions.

To summarize, we propose to explain the difference between our finding about punishment and previous results as follows. People have an emotion-based decision-making process leading to punishment. This process is inhibited when people interact with entities, such as artifacts, without adopting the intentional stance. When people

adopt the intentional stance, as participants in the computer condition of this experiment probably did, the decision-making process leading to punishment is not inhibited. When they are asked to justify their punishment decision, people appeal to the behavioral norms prevalent in their culture. This explains why participants justified their punishment decision differently in the human partner condition and in the computer condition, although participants punished their partner similarly, whether it was a human or a computer.

## References

- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, 122-142.
- Buchner, A., Erdfelder, E., & Faul, F. (1997). How to use G\*Power. URL [http://www.psych.uni-duesseldorf.de/aap/projects/gpower/how\\_to\\_use\\_gpower.html](http://www.psych.uni-duesseldorf.de/aap/projects/gpower/how_to_use_gpower.html).
- Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*, 284-299.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.
- de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. *Science*, *305*, 1254-1258.
- Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology*. Cambridge, MA: MIT Press.
- Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*, 785-791.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*, 980-994.
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137-140.
- Fehr, E., & Henrich, J. (2003). Is strong reciprocity a maladaptation. In P. Hammerstein (Ed.), *Genetic and culture evolution of cooperation* (pp. 55-82). Cambridge, MA: MIT Press.
- Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*, 137-140.
- Fudenberg, D., & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, *54*, 533-554.
- Gallagher, H., Jack, A., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, *16*, 814-821.
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Sciences*, *7*, 287-292.

- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814-834.
- Henrich, J., et al. (2005). 'Economic man' in cross-cultural perspective: Ethnography and experiments from 15 small-scale societies. *Behavioral and Brain Sciences*, *28*, 795-855.
- Henrich et al. (2006). Costly punishment across human societies, *Science*, *312*, 1767-1770.
- Jianakoplos, N. A., & Bernasek, A. (1998). Are women more risk averse? *Economic Inquiry*, *36*, 620-630.
- Kahneman, D., Schkade, D., & Sunstein, C. R. (1998). Shared outrage and erratic rewards: The psychology of punitive damages. *Journal of Risk and Uncertainty*, *16*, 49-86.
- McCabe, K. A., & Smith, V. L. (2000). A comparison of naive and sophisticated subject behavior with game theoretic predictions. *Proceedings of the National Academy of Science*, *97*, 3777-3781.
- McCabe, K. A., Houser, D., Ryan, L., Smith, V. L., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Science*, *98*, 11832-11835.
- Pillutla, M. M., & Murnighan, J. K. (1996). Unfairness, anger, and spite: emotional rejections of ultimatum offers. *Organizational Behavior & Human Decision Processes*, *68*, 208-224.
- Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron*, *35*, 395-405.
- Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004) The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, *22*, 1694-1703.
- Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*, 1755-1758.
- Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Liking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87-124.

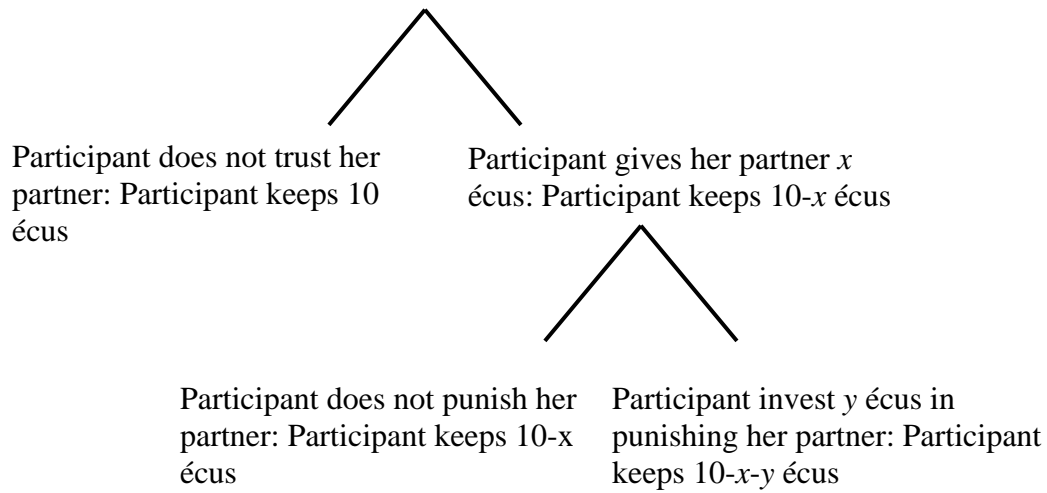
Schubert, R., Brown, M., Gysler, M., & Brachinger, H. W. (1999). Financial decision-making: Are women really more risk-averse? *American Economic Review*, 89, 381-385.

**Table 1:** Mean offer in écus, mean request in écus and mean amount invested in punishment in percentage (standard deviation in parentheses) in the human player condition and in the computer condition.

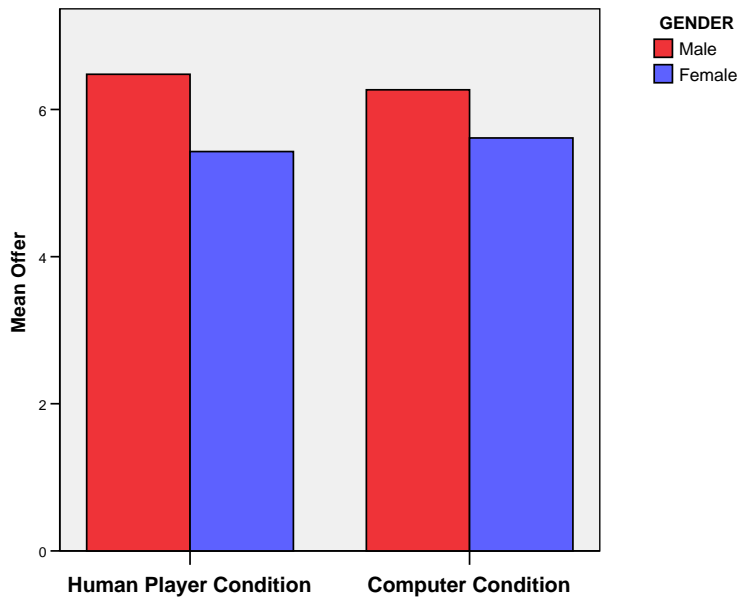
|                   | <b>Human player condition</b> | <b>Computer condition</b> |
|-------------------|-------------------------------|---------------------------|
| <b>Offer</b>      | 6.23 (3.05)*                  | 6.09 (3.05)*              |
| <b>Request</b>    | 9.14 (6.39)*                  | 10.38 (7.22)*             |
| <b>Punishment</b> | 30.44 (39.46)*                | 30.14 (39.46)*            |

\* Significantly different from the behavior of rational, self-interested agents ( $P < .005$ ).

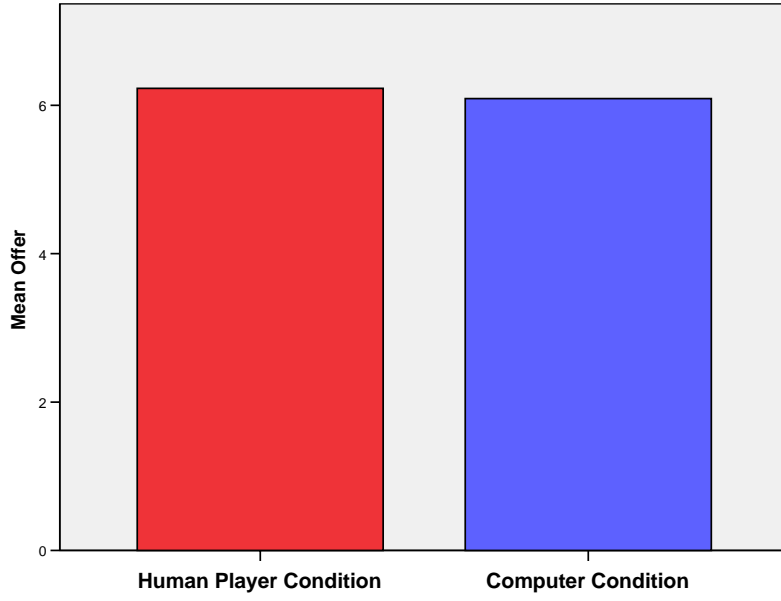
**Figure 1:** Diagram of the trust game with a punishment stage. The participant plays first, either by moving left and ending the game, or by moving right and giving  $x$  écus from her initial endowment to her partner. The partner gives back nothing. The participant decides to punish her partner, by subtracting  $y$  écus from her remaining amount of money.



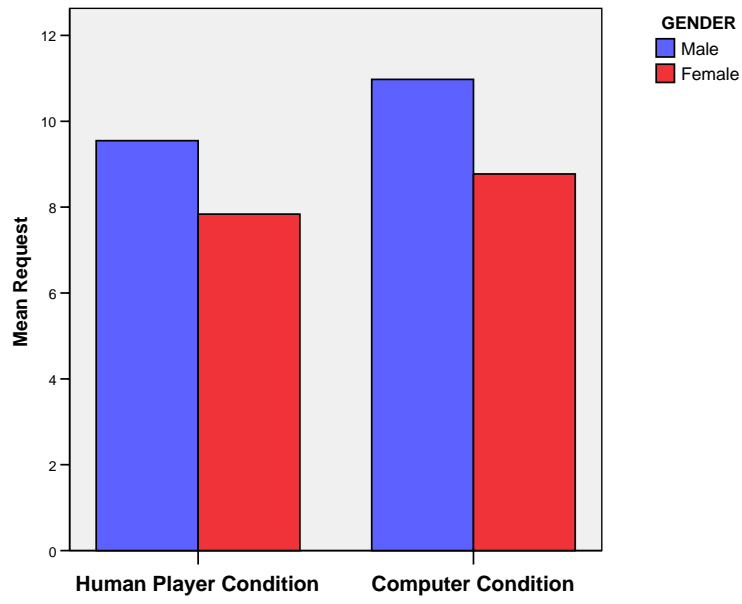
**Figure 2:** Mean offer in écus by gender and condition



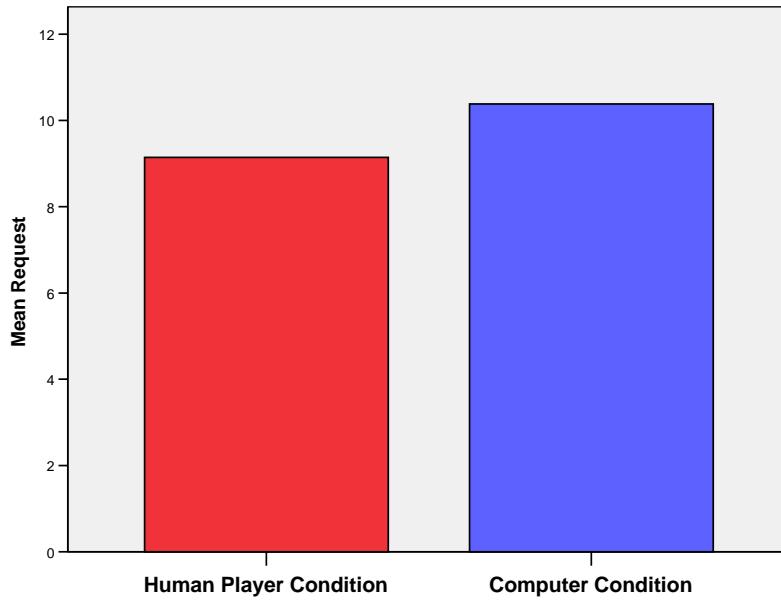
**Figure 3:** Mean offer in écus in the human player condition and in the computer condition



**Figure 4:** Mean request in écus by gender and condition



**Figure 5:** Mean request in écus in the human player condition and in the computer condition



**Figure 6:** Mean relative amount invested in punishment in the human player condition and in the computer condition (in percentage)

