

Effects of Agents' Transparency on Teamwork

Silvia Tulli^(✉), Filipa Correia, Samuel Mascarenhas, Samuel Gomes,
Francisco S. Melo, and Ana Paiva

Department of Computer Science and Engineering,
INESC-ID and Instituto Superior Técnico, Universidade de Lisboa,
2744-016 Porto Salvo, Portugal

{silvia.tulli,samuel.mascarenhas}@gaiips.inesc-id.pt,
{filipacorreia,samuel.gomes}@tecnico.ulisboa.pt,
{francisco.melo,ana.paiva}@inesc-id.pt

Abstract. Transparency in the field of human-machine interaction and artificial intelligence has seen a growth of interest in the past few years. Nonetheless, there are still few experimental studies on how transparency affects teamwork, in particular in collaborative situations where the strategies of others, including agents, may seem obscure.

We explored this problem using a collaborative game scenario with a mixed human-agent team. We investigated the role of transparency in the agents' decisions, by having agents that reveal and tell the strategies they adopt in the game, in a manner that makes their decisions transparent to the other team members. The game embraces a social dilemma where a human player can choose to contribute to the goal of the team (cooperate) or act selfishly in the interest of his or her individual goal (defect). We designed a between-subjects experimental study, with different conditions, manipulating the transparency in a team. The results showed an interaction effect between the agents' strategy and transparency on trust, group identification and human-likeness. Our results suggest that transparency has a positive effect in terms of people's perception of trust, group identification and human likeness when the agents use a tit-for-tat or a more individualistic strategy. In fact, adding transparent behaviour to an unconditional cooperator negatively affects the measured dimensions.

Keywords: Transparency · Autonomous agents · Multi-agent systems · Public goods game · Social dilemma

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT-UID/CEC/50021/2019), and Silvia Tulli acknowledges the European Union's Horizon 2020 research and innovation program for grant agreement No. 765955 ANIMATAS project. Filipa Correia also acknowledges an FCT grant (Ref. SFRH/BD/118031/2016).

1 Introduction

The increase of intelligent autonomous systems capable of complex decision-making processes affects humans' understanding of the motivations behind the system's responses [6]. In this context, evaluating the performance of machine learning algorithms may not be sufficient to prove the trustworthiness and reliability of a system in the wild [25].

Machine learning models appear to be opaque, less intuitive and challenging for the diversified end users. To meet this need, an increasing number of studies has focused on developing transparent systems. However, the definition of transparency is still up for debate. The most commonly used terms are model interpretability, explicability, reliability, and simplicity. Doshi-Velez and Kim define interpretability as the ability to explain or present understandable terms to a human [12]. Instead, Rader et al. explain transparency as providing the non-obvious information that is difficult for an individual to learn or experience directly, such as how and why a system works the way it does and what its outputs mean [26]. The lack of a consensual definition of transparency reflects in a lack of comparable metrics to assess it. Due to this, to understand transparency, it is necessary to manipulate and measure various factors that can influence the perception and behavior of humans. Designing the transparency of a system is therefore not a purely computational problem.

A variety of human challenges demands for effective teamwork [18]. However teamwork has numerous implications: the commitment of all the members to achieve the team goals, the trust among the team members, the mutual predictability for effective coordination, and the capability to adapt to changing situations [19, 22]. Many of the features needed for successful teamwork are well illustrated in video games scenarios [14], and due to this, video games have become a popular object of investigation for social and cultural sciences [23]. When autonomous systems move from being tools to being teammates, an expansion of the model is needed to support the paradigms of teamwork, which require two-way transparency [6]. As in human-human groups, the communication of relevant information can facilitate analysis and decision-making by helping the creation of a shared mental model between the group members. Several studies based on human-agent collaboration suggest that humans benefit from the transparency of the agent, which consequently improves the cooperation between them [26]. Moreover, agents' transparency facilitates the understanding of the responsibilities that different group members might take in collaborative tasks.

Contrary to what we could hypothesize, collaborative games can also encourage anti-collaborative practices that derive from the identification of a single winner and from the fact that players rely on the contribution of others and therefore invest less in their actions (free riding) [4]. For this reason, combining the investigation of the behavioral model of the players in relation to the different strategies of the team members and the transparency of the decision-making process of the artificial players turns out to be useful for the design of systems that aim to facilitate and foster collaboration. The objective of this study is

to investigate the effect of the transparency and strategy of virtual agents on human pro-social behavior in a collaborative game.

2 Related Work

The lack of transparency is considered one of the obstacles for humans to establish trust towards autonomous systems [10]. In fact, trust appears as a common measure to assess the effect of transparency and it is related to the level of observability, predictability, adjustability, and controllability, as well as mutual recognition of common objectives of a system. Chen et al. have developed a model for collaboration and mutual awareness between humans and agents [6]. This model is called *Situation Awareness Based Agent Transparency* (SAT) and considers current plans and actions, decision-making and prediction of responses. To sum up, the SAT model describes the type of information that the agent should provide on its decision-making process to facilitate mutual understanding and collaboration between human and agent. The first level of the model includes information related to the actions, plans, and objectives of the agent. This level helps human's perception of the current state of the agent. The second level considers the decision-making process with the constraints and affordances that the agent takes into account when planning its actions. With that, the human can understand the current behavior of the agent. The third level provides information related to the agent's projection towards future states with the relative possible consequences, the probability of success or failure, and any uncertainty associated with the previously mentioned projections. The third level allows the human to understand the future responses of the agent. Our manipulation of the agents' transparency considers the three levels of the SAT informing about the current actions and plans, and including the decision-making process (e.g. "My plan is to always improve the instrument"). The third level results as a projection of the pursued strategy.

Given that, it can be difficult to distinguish in the literature whether transparency refers to the mechanism or the outcome, the cause or the effect [26]. However, in the context of human-machine interaction, transparency means an appropriate mutual understanding and trust that leads to effective collaboration between humans and agents. The act of collaboration and cooperation in group interactions is not only interesting for researchers in the area of human-machine interaction but is also widely studied by social sciences to obtain knowledge on how cooperation can be manipulated. In particular, to understand how individuals in a group can be stimulated to contribute to a public good [13]. Several studies, both theoretically and empirically, shown that transparency has a positive effect on cooperation. For instance, Fudenberg et al. demonstrated that transparency of past choices by the group members is necessary to maintain a sustainable and stable cooperation [15]. Davis et al. shown that transparency allows cooperative players to indicate their cooperative intentions, which may induce others to similar cooperative behaviors [11].

3 Research Design and Methods

We conducted a between-subject user study using the Mechanical Turk and the “For The Record” game [9]. “For the Record” is a public goods game that embraces a social dilemma where a human player can choose to contribute to the goal of the team (cooperate) or act selfishly in the interest of his or her individual goal (defect). In linear public goods environments *maximizers have a dominant strategy to either contribute all of their tokens or none of their tokens to a group activity* [5,28]. In the “For The Record” experimental scenario, three players, one human, and two artificial agents, have the goal of publishing as many albums as possible. The number of albums to be created and produced matches the number of rounds to play, in our case, 5 rounds and if players fail 3 albums they lose the game. During the first round, each player starts playing by choosing the preferred instrument that can be used to create the album. Starting from the second round each player has two possible actions and they concern the possibility of investing in the instrument’s ability (contributing to the success of the album) or in the marketing’s ability (contributing to the individual monetary value, or personal profit, obtained after the album’s success). This investment is translated into the number of dice that the player can use, in the first case to play the instrument and helps to create the album, while in the second case to receive profit. During the creation of the album, each player will contribute equally to the value obtained from the roll of the dice, and the number of die available to the player will depend on the level/value of the skill (marketing or instrument). The score of an album consists of adding up the values achieved by each player during his performance. After creating the album, the band has to release it on the market. The market value is evaluated by rolling 2 dice of 20 faces. If the market value is higher than the album score, than the album is considered a “Fail”. On the other hand, if the market value is less than or equal to the score on the album, that album is considered a “Mega-hit”. From the fourth round on, the band enters the international market, which means that the market value is evaluated by rolling 3 die of 20 faces (instead of the 2 previous dices). This increases the difficulty of getting successful albums. The game has always been manipulated to return a victory.

4 Objective and Hypothesis

The objective of this study was to investigate the effect of the transparency and strategy of virtual agents on human pro-social behavior in a collaborative game. Despite having hypothesized that transparency would affect several measures of teamwork, we have also manipulated the agents’ strategy to confirm if the results would provide similarly when the agents adopted different strategies. In a two by three (2×3) between-subjects design, resulting in six experimental conditions, we manipulated the agents’ transparency and the agents’ strategy, respectively. The two levels of transparency were:

- **Transparent:** The agents explain their strategy;
- **Non-transparent:** The agents do not explain their strategy.

The three possible strategies for the agents were:

- **Cooperative:** The agents always cooperate;
- **Individualistic:** The agents cooperate only if the last round has been lost;
- **Tit for Tat:** The agents cooperate only if the player cooperates.

We expected that the transparency of the agents will positively affect teamwork and make the agents' strategy easily to interpret. We also expected transparency to increase trust and facilitate collaboration due to mutual understanding and shared responsibilities. Therefore we have the following hypotheses:

- H1: The agents' transparency increases the number of cooperative choices of the human player;
- H2: The agents' transparency results in greater trust and group identification;
- H3: The agents' transparency increases the likeability and human likeness of the artificial player;

The hypothesis that the transparency increases the number of cooperative choices is based on the fact that transparency about choices tends to lead to an increase in contributions and collusion [13]. The hypothesis that positive effect of transparency on trust and group identification relies on the evidence that transparency have the (perhaps counter-intuitive) quality of improving operators' trust in less reliable autonomy. Revealing situations where the agent has high levels of uncertainty develops trust in the ability of the agent to know its limitations [7,8,16,24]. The hypothesis that the agents' transparency results in greater likeability and perceived human likeness of the artificial player refers to the experimental evidence of Herlocker et al. showing that explanations can improve the acceptance of automated collaborative filtering (ACF) systems [17].

4.1 Materials and Methods

Agents' Transparency Manipulation. The interactive agents commented some game events through text in speech bubbles, e.g., *That was very lucky!* or *Lets record a new album.*

The duration of such stimuli depend on the number of words shown, according to the average reading speed of 200–250 words per minute. However, the speech bubbles containing the manipulation of each experimental condition lasted twice as much to make sure the participants would read them (Fig. 1).

Table 1 shows the explanation given by the artificial agents while they are choosing the main action of adding a point to either the instrument or the marketing in the transparent and non-transparent conditions:

In the non-transparent conditions the agents explain what they are doing for that current round, in the transparent conditions they explicitly refer to their plans and intentions.



Fig. 1. Example of a speech bubble with the explanation of the agents' strategy

Table 1. Manipulation of transparent and non-transparent behaviour for each agents' strategy

Strategy	Transparency	
	Transparent	Non-transparent
Cooperative	<ol style="list-style-type: none"> 1. <i>"My strategy is to always improve the instrument."</i> 2. <i>"My plan is to always improve the instrument."</i> 	<ol style="list-style-type: none"> 1. <i>"I am going to improve the [instrument/marketing]."</i> 2. <i>"I will put one more point on my [instrument/marketing]."</i>
Individualistic	<ol style="list-style-type: none"> 1. <i>"My plan is to improve my marketing skill only when the album succeeds."</i> 2. <i>"My plan is to improve my instrument skill only when the album fails."</i> 	
Tit for tat	<ol style="list-style-type: none"> 1. <i>"My strategy is to improve my instrument skill only when you also improve your instrument."</i> 2. <i>"My strategy is to improve my marketing skill only when you also improve your marketing."</i> 	

4.2 Metrics and Data Collection

To test our hypotheses and, therefore, analyse the effects of the strategy and transparency adopted by the agents, we used different metrics and items from standardized questionnaires. The self-assessed questionnaire included some demographic questions (e.g., age, gender and ethnicity), a single-item on their

self-perceived competitiveness level, two items regarding the naturalness and human-likeness of the agents' strategies, and two validation questions to evaluate the understanding on the rules of the game. The remaining measures are detailed as follows.

Cooperation Rate. The cooperation rate was an objective measure assessed during the game-play. In the beginning of each round, each player has to choose between to cooperate with the team (i.e., by upgrading the instrument skill) or to defect for individual profit (i.e., by upgrading the marketing skill). This measure sums up the total number of times the human player opted to cooperate and can range, in discrete numbers, from zero to four. It represents the degree of pro-sociality that the human participant expressed while teaming with the agents.

Group Trust. We chose the Trust items by Allen et al. in [1], which were explicitly designed for virtual collaboration to assess the trust through the agents. Trust is described as a key element of collaboration and is divided into seven items with a 7 points likert-scale from totally disagree to totally agree.

Multi-component Group Identification. Leach et al. identified a set of items for the assessment of the Group-Level Self-Definition and Self-Investment in [21]. The idea behind this scale is that individuals' membership in groups has relevant impact on humans behavior. Specifically designed items represents the five dimensions evaluated: individual self-stereotyping, in-group homogeneity, solidarity, satisfaction, and centrality. These items were presented with a Likert-type response scale that ranged from 1 (strongly disagree) to 7 (strongly agree). We decided to use the dimensions of homogeneity, solidarity and satisfaction as relevant metrics for our study.

Godspeed. The Godspeed scale was designed for evaluating the perception of key attributes in Human-Robot Interaction [3]. More precisely, the scale is meant to measure the level of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. Each dimension has five or six items with semantic differentials couples that respondents are asked to evaluate in a 5 points Likert scale. We used the dimensions of the likeability (Dislike/Like, Unfriendly/Friendly, Unkind/Kind, Unpleasant/Pleasant, Awful/Nice) and perceived intelligence (Incompetent/Competent, Ignorant/Knowledgeable, Irresponsible/Responsable, Unintelligent/Intelligent, Foolish/Sensible).

4.3 Procedure

Participants were asked to complete the task in around 40 min. The experiment was divided in three phases. The first phase consisted of the game tutorial, and lasted around 15 min. The second phase consists in playing a session of

“For the Record” with the two artificial agents, which lasted around 15 min. The last phase was represented by the questionnaire and took round 10 min. We informed the participants about the confidentiality of the data, voluntary participation and the authorization for sharing the results with the purpose of analysis, research and dissemination. We specified that we were interested in how people perceive teamwork and the game strategies of the two artificial players they were going to play with. After finishing the experiment and providing their judgments, we thanked the participants for their participation giving them 4\$. We collected the data for the non transparent and the transparent condition separately, ensuring that none of the participants repeat the experiment twice.

5 User Study

The main goal of our study was to explore the role transparent behaviors have on the perception of intelligent agents during human-agent teamwork. In particular, to analyze if transparency can enhance the perception of the team and the display of pro-social behaviors by humans.

5.1 Participants

The participants involved in the study were 120, 20 participants per each experimental condition (Cooperative, Individualistic and Tit for Tat). Considering the study was done in MTurk and the fact that the experiment took more time than the turkers are used to, we introduced some attention and verification questions in order to ensure the quality of the data. The criteria to exclude participants were: not having completed the entire experiment; having reported an incorrect score of the game; and having provided wrong answers to the questions related to the game rules (e.g., *How many dices are rolled for the international market?*). Consequently, we run the data analysis on a sample of a sample of 80, 28 in the non-transparency conditions and 52 in the transparency conditions. The average age of the sample was 37 years ($\text{min} = 22$, $\text{max} = 63$, $\text{stdev} = 8.78$) and was composed of 52 males and 27 females and one other. The participants were randomly assigned to one of three condition of the strategy: 19 in the cooperative condition (13 in the transparency condition and 6 in the non-transparency condition), 30 in the individualistic condition (17 for the transparency condition and 13 in the non-transparency condition), 18 for the tit-for-tat condition (9 for the transparency condition and 11 in the non-transparency condition).

5.2 Data Analysis

We analyzed the effects of our independent variables - transparency (binary categorical variable *Transparent* and *Non-Transparent*) and strategy (three categories: *Cooperative*, *individualistic* and *Tit for Tat*) - on the dependent variables.

The reliability analysis for the dimensions of the Trust scale, the Group Identification scale, the Godspeed scale as well as the Human likeness and Naturalness revealed excellent internal consistency among items of the same dimensions (Trust: $\alpha = 0.912$; Group Identification: $\alpha = 0.972$; Group Solidarity: $\alpha = 0.953$; Group Satisfaction: $\alpha = 0.969$; Group Homogeneity: $\alpha = 0.923$; Perceived Intelligence: $\alpha = 0.962$; Likeability: $\alpha = 0.978$; Human-likeness and Naturalness: $\alpha = 0.938$).

Cooperative Rate. The analysis of the number of defects, revealed that the main effect of transparency was not significant ($F(1, 73) = 0.320, p = 0.573$), and the main effect of strategy was not significant ($F(3, 73) = 2.425, p = 0.072$). The interaction effect between the two factors was not significant ($F(2, 73) = 0.003, p = 0.997$). The specific values per each strategy were: Cooperative ($M = 1.11, SE = 0.201, SD = 0.875$), Individualistic ($M = 1.70, SE = 0.153, SD = 0.837$), Tit for Tat ($M = 1.06, SE = 0.249, SD = 1.056$).

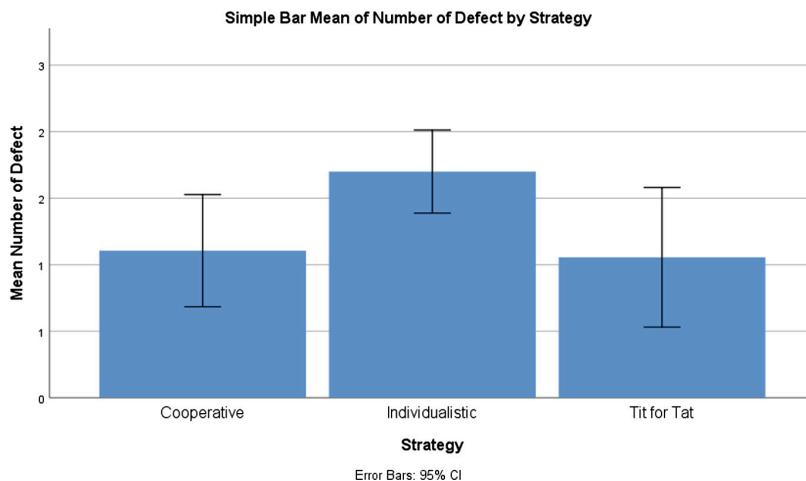


Fig. 2. Number of defects by strategy

Group Trust. The Analysis of Variance in Trust, showed that the main effect of transparency was not significant ($F(1, 73) = 0.337, p = 0.563$), and the main effect of strategy was significant ($F(3, 73) = 8.117, p < 0.001$). The specific values for each strategy were: Cooperative ($M = 5.25, SE = 0.265, SD = 1.154$), Individualistic ($M = 4.42, SE = 0.230, SD = 1.261$), Tit for Tat ($M = 5.22, SE = 0.221, SD = 0.938$).

The interaction effect between the two factors was significant ($F(2, 73) = 3.833, p = 0.026$).

Figure 3 shows that only in the cooperative condition the transparency negatively influenced the level of trust towards the agents. The specific values per each strategy in the transparent and non-transparent conditions were: Transparent - Cooperative ($M = 4.90, SE = 0.334, SD = 1.204$), individualistic ($M = 4.89,$

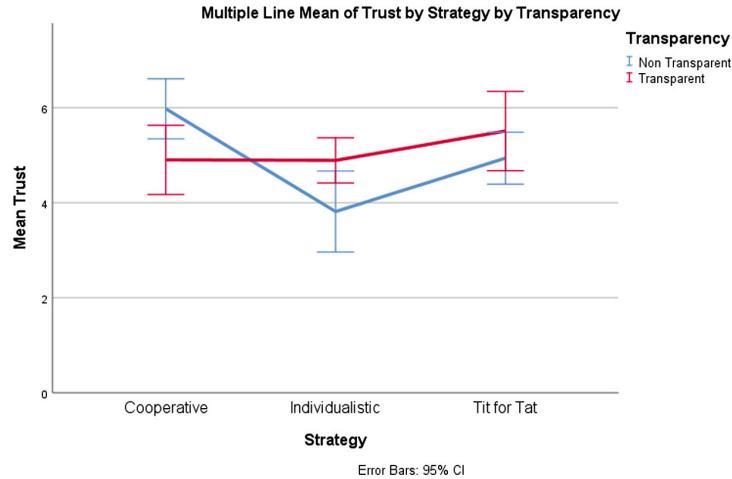


Fig. 3. Interaction effect between strategy and transparency in trust

$SE = 0.224$, $SD = 0.925$), Tit for Tat ($M = 5.51$, $SE = 0.362$, $SD = 1.086$) Non-Transparent - Cooperative ($M = 5.98$, $SE = 0.246$, $SD = 0.602$), Individualistic ($M = 3.81$, $SE = 0.291$, $SD = 1.411$), Tit for Tat ($M = 4.95$, $SE = 0.239$, $SD = 0.711$).

Multi-component Group Identification. The Group Identification, did not reveal main effect of single factors of transparency and strategy ($F(1, 73) = 2.674$; $F(3, 73) = 2.360$, $p = 0.106$, $p = 0.078$). However, the interaction between the two factors was significant ($F(2, 73) = 4.320$, $p = 0.017$). The specific values per each strategy in the transparent and non-transparent conditions: Transparent - Cooperative ($M = 4.15$, $SE = 0.500$, $SD = 1.801$), Individualistic ($M = 5.06$, $SE = 0.336$, $SD = 1.387$), Tit for Tat ($M = 5.27$, $SE = 0.427$, $SD = 1.282$). Non-Transparent - Cooperative ($M = 5.19$, $SE = 0.394$, $SD = 0.965$), Individualistic ($M = 3.32$, $SE = 0.359$, $SD = 1.292$), Tit for Tat ($M = 3.98$, $SE = 0.559$, $SD = 1.676$).

As we can notice from the Fig. 4, transparency and strategy influenced the perception of Group Identification in the opposite direction among the agents' strategies. In the transparency condition, the agents foster less group identification when they acts cooperatively. However, transparency had a positive influence in the group identification in the Individualistic and Tit for Tat condition. The One-way ANOVA in Group Identification reveals that the effect of transparency in Cooperative condition was not significant ($F(1, 17) = 1.732$, $p = 0.206$), the effect of transparency in Individualistic condition was significant ($F(1, 28) = 12.178$, $p = 0.002$) and the effect of transparency in Tit for Tat condition was not significant ($F(1, 16) = 3.398$, $p = 0.084$).

Goodspeed. The Likeability did not reveal a main effect of transparency ($F(1, 73) = 0.001$, $p = 0.973$) but informed a main effect of the strategy on the likeability ($F(3, 73) = 3.279$, $p = 0.026$) Fig. 5. The interaction between the

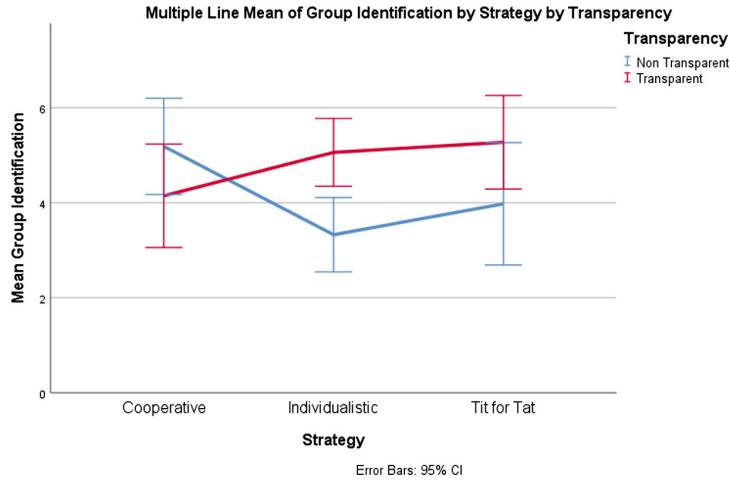


Fig. 4. Interaction effect between strategy and transparency in Group Identification

transparency and strategy was not significant ($F(2, 73) = 0.855, p = 0.429$). Again in this case, the strategy affected the perception of likeability, and no interaction was found regardless of whether or not the agents employ transparent behaviors.

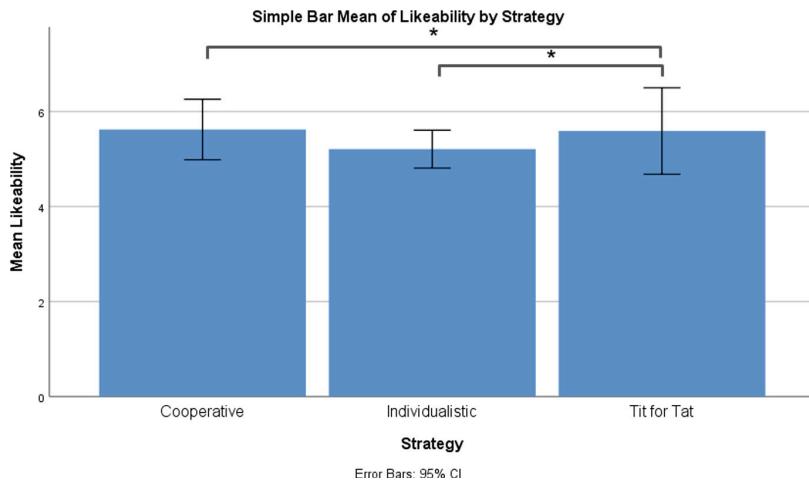


Fig. 5. Main effect of the strategy on likeability

For the human-likeness dimension, there was no main effect of transparency ($F(1, 73) = 0.145, p = 0.704$) and no main effects of the strategy ($F(3, 73) = 2.181, p = 0.098$). However, there was a significant interaction effect between transparency and strategy for the Human-likeness attributed to the agents ($F(2, 73) = 3.585, p = 0.033$).

In Fig. 6 we confirmed the trend of a different effect of transparency in the cooperative condition in respect to the strategy. For the Tit for Tat condition we can notice that both strategy and transparency positively affect the perceived human likeness of the agents.

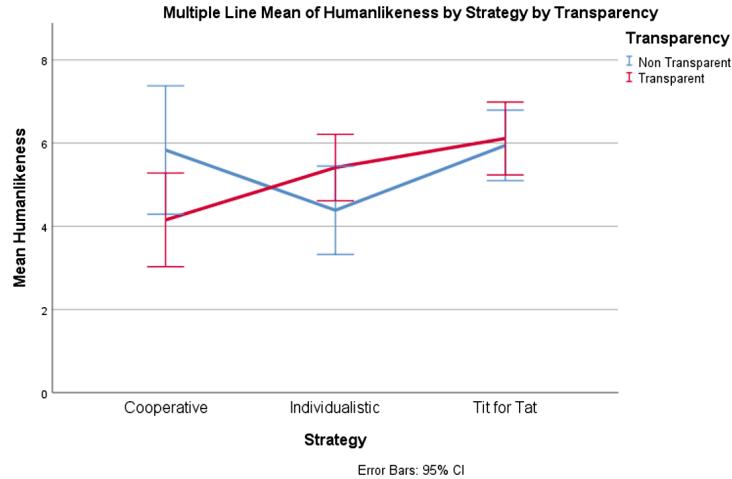


Fig. 6. Interaction effect between strategy and transparency in humanlikeness

The Univariate Analysis of Variance of the transparency and strategy for the Perceived Intelligence informed that the main effect of transparency was not significant ($F(1, 73) = 0.652, p = 0.422$), but the main effect of strategy was significant ($F(3, 73) = 5.297, p = 0.002$) Fig. 7. The interaction effect between the two fixed factors was not significant ($(2, 73) = 3.632, p = 0.179$). In other words, only the strategy of the agents, regardless of whether or not the agents employ transparent behaviors, affects the perceived intelligence of the agents, in particular for the Tit for Tat strategy as confirmed by several studies about game theory [2, 27]. The specific values per each strategy were: Cooperative ($M = 5.39, SE = 0.348, SD = 1.518$), Individualistic ($M = 5.23, SE = 0.227, SD = 1.244$), Tit for Tat ($M = 6.11, SE = 0.249, SD = 1.054$).

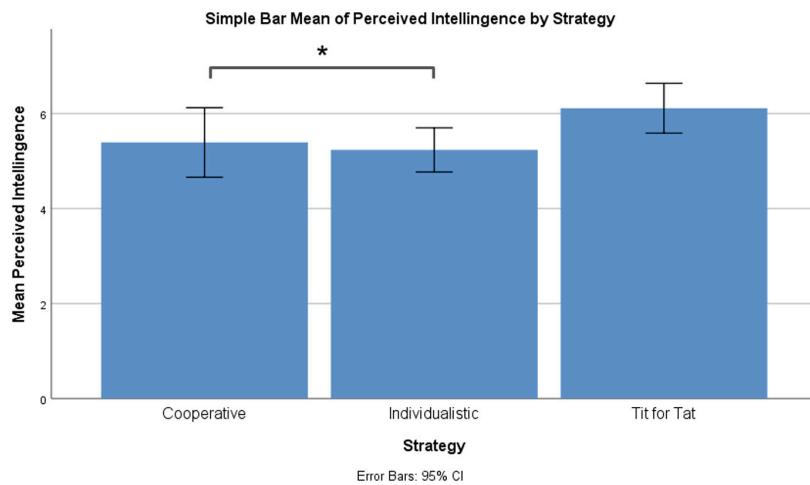


Fig. 7. Main effect of the strategy on perceived intelligence

6 Discussion

This paper explores group interactions involving mixed groups of humans and virtual agents in collaborative game settings. In particular, it is focused on how agents' transparency affects teamwork and the perception of autonomous teammates. Although we have hypothesized that transparency would positively influence several measures of teamwork, we have also manipulated the strategy of the agents to ascertain if the results would hold similarly when the agents adopted different Strategies.

According to **H1**, we expected that the agents' transparency would increase the number of cooperative choices of the human player, which was not confirmed. In fact, we only found a partially significant main effect of the strategy on the number of cooperative choices, which suggests people cooperated differently according to which strategy the agents adopted. In the post hoc analysis, cooperation towards the individualistic agents was lower than towards cooperative and tit-for-tat agents. Additionally, we analyzed the cooperation rate of the agents and we found the individualistic strategy led the agents to cooperate less compared to the other two Strategies, which suggests people might have reciprocated the autonomous agents to a certain extent Fig. 2. In our experiment, we could not find evidence that transparency affects people's behaviour.

Regarding **H2**, we have hypothesized that trust and group identification would be positively affected by transparent behaviour. On both measures, we found a significant interaction effect of transparency and strategy, which reveals the effect of transparency on trust and group identification was different across the three Strategies. In terms of the trust, the post-hoc analysis did not reveal a significant effect of transparency in any of the Strategies. However, the trends that are visible in Fig. 3 suggest this effect was negative for the cooperative agents and was positive for both the individualistic and tit-for-tat agents. In the post-hoc analysis for the group identification, we found a significant positive effect of transparency for the individualistic agents. For the remaining Strategies, similar trends are visible in Fig. 4 suggesting a negative effect for cooperative agents and a positive effect for tit-for-tat agents. Our hypothesis was only partially validated due to the fact that both group measures showed a positive effect only for two Strategies, the individualistic and tit-for-tat. Later in this section, we discuss the negative effect on the cooperative strategy.

In **H3**, we have predicted that transparent behaviours would positively affect the likeability and human-likeness of the agents. We only found a significant interaction effect between transparency and strategy on the perceived human-likeness. In other words, the effect of transparency on the perception of human-likeness was different across the three Strategies. Although the post hoc analysis did not reveal a significant effect of transparency in any of the Strategies, the trends suggest a negative effect on the cooperative agents, a positive effect on the individualistic agents and no effect is suggested for the tit-for-tat agents. In terms of likeability, we found a significant main effect of the strategy with the individualistic agents being significantly rated as less likeable compared to the

cooperative and tit-for-tat agents Fig. 5. This hypothesis was validated in terms of human-likeness for the agents that use a individualistic strategy.

Our results suggest that adding transparent behaviour to an unconditional cooperator negatively affects the perceptions people have in terms of trust, group identification and human likeness. Although these differences were not statistically significant, the trends are congruent in the same direction. Further investigation is needed to support this claim. In terms of human-likeness, our intuition is that the unconditional cooperator might have revealed to the participants a non-optimal strategy, which a human would probably not do. However, the result for the group measures are counter-intuitive because the non-optimality of this strategy is related to the individual gains and it is not clear why the unconditional cooperator negatively affected then perception of the group.

7 Conclusions

Research in the field of artificial intelligence requires the design of system transparency able to improve the collaboration in human-agents and human-robot scenarios. This research discusses how strategy and transparency of artificial agents can influence human behavior in teamwork. Within the limits of the results found, we can state that transparency has significant effects on the trust, group identification and human likeness. This aspect turns out to be interesting in the context of public goods games and the design of relational and social capabilities in intelligent systems. Further research should consider the use of the Social Value Orientation [20] to randomize the sample between the condition before running the study. In addition, other type of transparency exploitation should be explored, as well as other game scenario and a more selected sample based on specific objectives, such as education or ecological sustainability. To conclude, a more comprehensive investigation of the methods to evaluate and implement the system transparency considering the effect of agents' strategy should be considered and tested in the wild.

References

- Allen, K., Bergin, R.: Exploring trust, group satisfaction, and performance in geographically dispersed and co-located university technology commercialization teams. In: Proceedings of the NCIIA 8th Annual Meeting: Education that Works, pp. 18–20 (2004)
- Axelrod, R.: On six advances in cooperation theory. *Anal. Kritik* **22**, 130–151 (2000). <https://doi.org/10.1515/auk-2000-0107>
- Bartneck, C., Kulic, D., Croft, E., Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **1**, 71–81 (2008). <https://doi.org/10.1007/s12369-008-0001-3>
- Bornstein, G., Nagel, R., Gneezy, U., Nagel, R.: The effect of intergroup competition on group coordination: an experimental study. *Games Econ. Behav.* **41**, 1–25 (2002). <https://doi.org/10.2139/ssrn.189434>

5. Burton-Chellew, M.N., Mouden, C.E., West, S.A.: Conditional cooperation and confusion in public-goods experiments. *Proc. Nat. Acad. Sci. U.S.A.* **113**(5), 1291–6 (2016)
6. Chen, J.Y.C., Lakhmani, S.G., Stowers, K., Selkowitz, A.R., Wright, J.L., Barnes, M.: Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theor. Issues Ergon. Sci.* **19**(3), 259–282 (2018). <https://doi.org/10.1080/1463922X.2017.1315750>
7. Chen, J.Y., Barnes, M.J.: Human-agent teaming for multirobot control: a review of human factors issues. *IEEE Trans. Hum.-Mach. Syst.* **44**(1), 13–29 (2014)
8. Chen, J.Y., Barnes, M.J.: Agent transparency for human-agent teaming effectiveness. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics, pp. 1381–1385. IEEE (2015)
9. Correia, F., et al.: Exploring prosociality in human-robot teams. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 143–151. IEEE (2019)
10. DARPA: Explainable artificial intelligence (XAI) program (2016). www.darpa.mil/program/explainable-artificial-intelligence,fullsolicitationatwww.darpa.mil/attachments/DARPA-BAA-16-53.pdf
11. Davis, D., Korenok, O., Reilly, R.: Cooperation without coordination: signaling, types and tacit collusion in laboratory oligopolies. *Exp. Econ.* **13**(1), 45–65 (2010)
12. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
13. Fiala, L., Suetens, S.: Transparency and cooperation in repeated dilemma games: a meta study. *Exp. Econ.* **20**(4), 755–771 (2017)
14. Fredrickson, J.E.: Prosocial behavior and teamwork in online computer games (2013)
15. Fudenberg, D., Maskin, E.: The Folk theorem in repeated games with discounting or with incomplete. *Information* (2009). https://doi.org/10.1142/9789812818478_0011
16. Helldin, T.: Transparency for future semi-automated systems, Ph.D. dissertation. Orebro University (2014)
17. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 241–250. ACM (2000)
18. Sedano, C.I., Carvalho, M., Secco, N., Longstreet, C.: Collaborative and cooperative games: facts and assumptions, pp. 370–376, May 2013. <https://doi.org/10.1109/CTS.2013.6567257>
19. Klien, G., Woods, D.D., Bradshaw, J.M., Hoffman, R.R., Feltovich, P.J.: Ten challenges for making automation a “team player” in joint human-agent activity. *IEEE Intell. Syst.* **19**(6), 91–95 (2004). <https://doi.org/10.1109/MIS.2004.74>
20. Lange, P., Otten, W., De Bruin, E.M.N., Joireman, J.: Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. *J. Pers. Soc. Psychol.* **73**, 733–46 (1997). <https://doi.org/10.1037/0022-3514.73.4.733>
21. Leach, C.W., et al.: Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. *J. Pers. Soc. Psychol.* **95**(1), 144 (2008)
22. Lee, C.C., Chang, J.W.: Does trust promote more teamwork? Modeling online game players’ teamwork using team experience as a moderator. *Cyberpsychology Behav. Soc. Netw.* **16**(11), 813–819 (2013). <https://doi.org/10.1089/cyber.2012.0461>. pMID: 23848999

23. McEwan, G., Gutwin, C., Mandryk, R.L., Nacke, L.: “i’m just here to play games”: social dynamics and sociality in an online game site. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW 2012, pp. 549–558. ACM, New York (2012). <https://doi.org/10.1145/2145204.2145289>
24. Mercado, J.E., Rupp, M.A., Chen, J.Y., Barnes, M.J., Barber, D., Procci, K.: Intelligent agent transparency in human-agent teaming for multi-uxv management. *Hum. Factors* **58**(3), 401–415 (2016)
25. Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H.M.: Manipulating and measuring model interpretability. CoRR [arXiv:abs/1802.07810](https://arxiv.org/abs/1802.07810) (2018)
26. Rader, E., Cotter, K., Cho, J.: Explanations as mechanisms for supporting algorithmic transparency, pp. 103:1–103:13 (2018). <https://doi.org/10.1145/3173574.3173677>
27. Segal, U., Sobel, J.: Tit for tat: foundations of preferences for reciprocity in strategic settings. *J. Econ. Theory* **136**(1), 197–216 (2007). <https://EconPapers.repec.org/RePEc:eee:jetheo:v:136:y:2007:i:1:p:197-216>
28. Zelmer, J.: Linear public goods experiments: a meta-analysis. Quantitative studies in economics and population research reports 361. McMaster University, June 2001. <https://ideas.repec.org/p/mcm/qseprr/361.html>