Would you Trust an Agent

Nuno Filipe Simões Fialho
nuno.simoes.fialho@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

June 2021

Abstract

Trust is affected by a multitude of factors depending on the setting and interactions between trustor and trustee. This attracts investigators with the purpose of understanding how trust can be developed and what will generate a loss in trust levels. Trust has been studied for a long time but mostly on human-human interaction, with the study of human-machine interaction being less developed. Therefore this thesis proposes a study of the influence of behaviour and agency in Trust using the Trust Game. Trust was measured utilizing both observational measures and questionnaires that allow to measure trust perception. Results revealed an influence of agency, and a combination of both conditions in trust levels. Results also revealed an influence of the behaviour in metrics like cooperation and trust in the questionnaires but not in the game.
Keywords: Trust, Agents, Trust Game, Behaviour, Cooperation

1. Introduction

Nowadays, technology sets the pace of our lives. We have an increasing amount of possible devices and machines to help us in our daily activities. They aim at making our lives easier and hopefully better. The predominance of technology in our homes, work and in society in general brings new challenges and applications never before imagined. One of such technologies are Intelligent Agents either virtual or robotic. An agent is "the term used to denote a hardware or software-based computer system that has autonomy, social ability, reactivity and pro-activeness"[12], because of this it is also important to understand trust in robots, a specific case of agents. Intelligent Agents can take part in a multitude of tasks for example AI assistants, a robot vacuum, traffic applications, air traffic management. In those tasks the aim of Agents is mostly helping and facilitating the user. Since Agents can take a wide variety of roles in our lives it is important to understand how users react to them, and what features improve the relationship between an agent and a user. Furthermore, currently, as we witness an increasing distrust in institutions, media, organisations and people, and because technology is an integral element of these social entities, it is important to consider how humans trust technology (in particular Intelligent Agents) and what factors influence such trust. When designing a system, it is important to know what elements will lead to an increase or decrease of users trust in it in order to explore the full potential of the system. Users that do not trust a particular Agent or system, even if unconsciously, will not use or take advantage of the features it possesses.

Trust has been long studied in human-human interactions for insights about what factors influence trusting one another and what makes one trustworthy. How trust is established and developed in human-agent interactions has been less explored. A good tool to measure trust is the Trust Game, a game that allows measurement of trust in economic decisions. Being a more recent area, knowledge is not as deep, specially the unique characteristics of the interactions between a human and a machine. Even though some parts of the research on Human-Human interaction can be applied the interaction between Humans and technology, it is impossible to claim that the same knowledge can be applied in all cases. Studies reveal that the way a human reacts and bonds with another human is different from the way they do with an Agent or machine[4].

2. Objectives

This work studies what factors influence human-agent trust. Trust is an attitude towards a trustee that involves a multitude of beliefs and expectations. Because trust is influenced by so many factors such as anthropomorphism, morality, transparency, etc... In this study we will focus on only two, behaviour and agency. The study will consist of a version of the Trust Game. The Trust game is an experiment that allows to "study trust and reciprocity in an investment setting"[3]. A version of the Trust game was developed using unity con-

nected to a Server that allows to collect and store the game data. In the study the influence of the behaviour of the agent and it's agency are studied. The study proposes that a more selfish behaviour will lead to less trust. It also proposes that humans will consider Artificial Intelligence and Agents to be more trustworthy then other humans. This is evaluated using both direct measures of trust and questionnaires. It is important to understand how these factors affect trust in order to understand the relations between humans and technology.

3. Background
4. Trust

Because trust has been studied in a variety of situations and interactions, we have a large variety of definitions. Even though those definitions do not generally differ significantly from each other, we still do not have a definition of trust that is widely accepted [8]. However, one thing that is widely agreed upon is that trust is important in a multitude of ways since it can improve cooperation, network relations, reducing risks and conflicts and also improve fast formulation of work groups[9].

Trust can be conceptualized as " a multidimensional psychological attitude involving beliefs and expectations about the trustee's trustworthiness derived from experience and interactions with the trustee in situations involving uncertainty and risk " which is the commonly agreed conceptualization in human-human and human-machine literature[8]. When we talk about trust in automation the cognitive features are more valuable then affection and so another definition in this context can be "an attitude which includes the belief that the collaborator will perform as expected, and can, within the limits of the designer's intentions, be relied on to achieve the design goals"[8]. But for trust to exist we need to have a set of conditions such as Risk which is the high probability of loss perceived by the person making the decision. This is important because if we could have a certainty of the result of our actions we would not need trust. The second condition is interdependence, which happens when we cannot achieve or own goals without reliance on another party. This exists because the decision-maker has a preconceived idea of the way the other person will act, and calculates that acting that way, they can both benefit from cooperation, resulting in a better outcome for both.

5. Measuring Trust

Trust is not something palpable or physical that can easily be measured by a device, and as such, investigators developed strategies and questionnaires in order to be able to quantify it. Without the existence of standardized measurements, many studies were used short idiosyncratically worded questionnaires.

There are two different ways of measuring trust in general reported in the literature. The first is by self-assessment, this is designed to measure the participant's trust perception. Self-assessment has the advantage of easy deployment and assessment of information right from the user, but on the other hand, self-assessment is vulnerable to user bias (for example trying to respond what the user thinks others will like)[1]. Examples of this are:

- The Trust Perception Scale-HRI: a 40 item scale that intends to measure in robots, this scale's main components are, capability, behaviour, task and appearance[8].
- The HRI Trust Scale: 37 item scale based on 5 dimensions, team configuration, team process, context, task and system. This scale is meant to be paired with other scales[8].

The second way of measuring trust is by using observational measurements, where investigators observe the user and quantify/qualify his actions to quantify not the self-reported trust, but the actions took that revealed trust in the robot. Using observational measures eliminates the social acceptance bias for example, but can be biased by the investigator in choosing what actions reveal trust[1].

6. Trust in Humans vs Trust in Machines

S. Shyam Sundar et al. in 2019 [10] realized a study to see if there was a difference in trusting personal information to another human or a machine. The study hypothesised that if the participant realized he was interacting with a machine he would be more likely to reveal personal information. To evaluate this, participants would use an online chat(with a human or a machine) to book a flight, where they would be asked to disclose credit card information. The results show that when Siri(machine condition) asked for the credit card information, users were more likely to disclose it.

E. Bogert et al. in 2021 [4] conducted a series of experiments to understand if participants would rely on algorithms or social influence. In the experiments participants had to say how many people were in a photograph. In experiment 1 participants received a photograph with a labeled advice from either an algorithm trained on 5000 images or the average guess of 5000 persons. All advice received contained the true answer. Results showed that participants revised the answers more when they received advice from an algorithm(11% more) specially when the question was hard. In the second experiment participants would receive a series of 10 pictures, 5 labeled by an algorithm and 5 labeled by the average of other people. The results also showed that participants relied more on advice when it was

labeled as algorithmic. In the third experiment low quality advice was introduced for half of the questions. Results show that participants relied more on high quality advice specially if it came from the algorithm. Indicating that the algorithm condition was penalized more when providing bad advice.

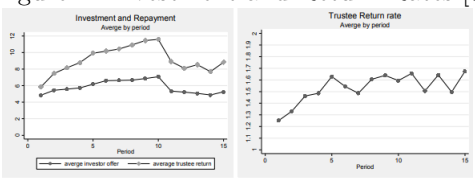
7. Repeated Trust Game

In 2012 Sacha Bourgeois-Gironde et al[5]. did a study in a laboratory where participants would play the trust game in pairs. Participants would be in front of a computer isolated from others, investor and trustee separated physically. The experiment consisted in 3 distinct steps. A software randomly selected the pairs of investors and trustees before each step. The 3 steps were:

- **One-shot anonymous trust game** where the players play a single round of the trust game
- **Repeated Trust Game** consisted of 10 rounds of a regular trust game, but the players were unaware of the number of rounds to be played.
- **Blind phase** were players played five rounds of the trust game, but the counter-offers remained secret until the end of the experience.

The results of this experience demonstrated that the most important feature the player returning the investment needed to have to build trust was predictability and stability in the return rates. Results showed that a stable return was more important than a high unstable return, because it showed that the partner was not opportunistic. Results also show that on average, players offered higher amounts on the second step of the experiment.

Figure 1: Investment and Return Rates [5]



In 2004, François Cochard [6] investigated the behaviours shown in a trust game. The experiment consisted of two separate phases. Phase one was a one-shot trust gamer. Phase two consisted of a 7 round trust game. In the second phase, the subjects were in separate rooms and knew about the number of rounds in the game.

-The results showed that in the repeated trust game, both players tend to send a larger percentage than those observed in phase one. This shows that the repeated trust game was advantageous for both players. -During the first six periods the payoff was bigger than in the one-shot game, but in the seventh period, that payoff decreased bellow the average of the first phase. -In the last period of the repeated game, on average, player A sent a similar amount then the average sent on the one-shot trust game, yet the percentages verified for the returned amount were lower. The average payoff ratio starts decreasing at period 5.

The author proposes the reciprocity hypothesis. This hypothesis states that players react to the behaviour of their partner, so when they receive a reward, they return a higher amount. If they are punished by the other

8. Implementation

The main goal of the software developed was to simulate a Trust game in such way that it would be easy for participants to understand the game easily and understand the interface of the game. Instructions needed to be clear in order to allow participants to better understand the game and it's objective.Because of the limitations imposed by the current pandemic situation, the game needed to be played completely online. The game had the objective of studying the influence of Agency and Behaviour on the trust developed by the participant.

The following sections will describe the development of the game.

9. Implementation of the Trust Game

The technology chosen to develop the game was Unity [11]. Unity is a Game Engine that allows the creation of 2D and 3D games programming in C#. This Game Engine allows for a faster and easier development due to a solid Graphical User Interface.

The flow of the game was designed to be the following:

- **Loading Screen**.
- **Instructions** - A small text explaining the rules and functioning of the game.
- **Avatar Selection** - The player chooses between 4 available avatars(Adult Male/Female and young Male/Female) to be displayed on the game.

Figure 2: Game Screen of the Trust Game.

- **Game** - The participant plays the game with the Agent. He has access to the amount of coins he possesses, the amount of coins the Agent possesses and an interface that allows coins to be sent to the Agent.
- **End Screen** - Shows the final amount of coins, and a code to fill in the questionnaire.

When the Participant is playing the game he sees the screen show in Figure 2. He can choose the coins he desires to send(either by using the slider or the buttons). Those coins will then be shown and tripled on the table. Finally he receives a part of the coins he sent to the robot.

At the end of each turn the game stores information in vectors into a static class. The animation of sending coins was divided following the structure presented:

- User selects the amount of coins to send and that is decremented from his total.
- Those coins appear on the table.
- Coins on the table are tripled.
- Coins of the Agent increase and Coins on the table disappear.
- After a waiting time, the agent sends coins to the participant.
- Coins are shown in the table and later added to the participant's total.

A single generic avatar was added in order to be more inclusive, a avatar was also added to the Computer(an avatar similar to the one the participant has in the player and AI condition and an animation of an EMYS robot in the Agent condition).

The Game only proceeds to the End Screen after receiving a Response from the Server. In the end Screen, the game would show the amount of coins each player had at the end of the game, the conversion to the bonus the participant would receive

Figure 3: On the Left the Generic Avatar, and on the Right the Animation of the EMYS.

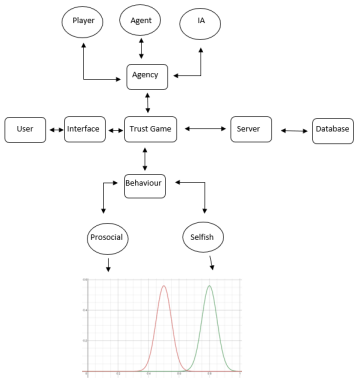
for his performance and the code to paste on the questionnaire.

In order to calculate how much coins the game would return to the participant a normal distribution was used. This normal distribution would return a variable with value between 0 and 1 that would previously be multiplied by the amount of coins the Agent received from the player. For the Prosocial behaviour the Median of the Normal Distribution was 0.8 with a standard deviation of 0.05 and for the selfish behaviour the Median of the Normal Distribution was 0.5 with a standard deviation of 0.05. this means that in the Prosocial behaviour the participant would receive the coins hr invested plus on average half of the profits the team made(Example: Player sent 2 coins, Game received 6 coins and returned 4), in The Selfish behaviour the player would receive the coins he invested and a small part of the profits, meaning the Game would get most of the profit coins(Example: Player sends 3 coins, Game receives 9 coins and returns 4). During the explanation of the game the participant would not be informed of the number of rounds he had to play the game. Participants would be informed that the bonus they would receive in the end would increase with the amount of coins they earned, but they were not informed of the Max amount of coins to get the maximum bonus or how the bonus was calculated. This would incentive participants to risk and try to get the highest amount of coins possible.

9.1. Server

In order to save variables of the game session each participant plays, emerged the need to develop a server. The server was developed using **NodeJS** and it was connected to a database located on **MongoDB**. The Database stores information about each Game Session. It stores the number of coins the participant sent, received and the total owned at the end of each round, in addition the Database stores which condition of the study was being played. When the server receives a HTTP POST request with information about a game session, the server checks the validity of that data and then generates a GUID that allows each session to be identified and connect the answers on the ques-

Figure 4: Communication between Game and Server



tionnaire to the respective game session.

In figure 4 we can see the different behaviours of the game and the different types of agency. The user interacts with the interface while he plays the trust game.

10. Studying Trust with the Trust game

10.1. Hypothesis

The hypothesis of this study are:

- **Hypothesis 1** The behaviour of the agent will influence the player's trust in him. A more punishing behaviour from the agent will lead to less trust by the player.
- **Hypothesis 2** The Autonomous Agent will have lower levels of trust than the AI Algorithm.
- **Hypothesis 3** The Human partner will have lower levels of trust than the AI Algorithm and the Autonomous Agent

10.2. Conditions

10.2.1 Number of Rounds

As stated previously, in [5] we can see results from the Investments made in a Repeated Trust Game (Figure 1). The results show that the investment made increased until around turn five of the Trust Game where it would stabilize until the end. In [6], the author states that in round seven the average investment is similar to the one observed in the one shot trust game.

Based on this information the game will have six rounds.

10.2.2 Behaviour

As stated in the Related Work section, in [6] it is proposed a reciprocity hypothesis that states that players will react to the behaviour of the person they are cooperating with, if they receive a reward they will trust and therefore send higher amounts of coins. And when punished they will trust less and by that send less coins. Based on this, in the study we will have two different behaviours:

- **Selfish** - In the Selfish behaviour the game will on average return what the player invested and in some cases a low percentage of the profit. The objective of this behaviour is to mimic a "punishing" behaviour that does not cooperate
- **Prosocial** - In the prosocial behaviour the game will on average return what the player invested plus an high percentage of the profits, making investment more beneficial to the player. The objective of this behaviour is to mimic a rewarding behaviour that cooperates.

10.2.3 Agency

Trust in Human-Human interaction and in Human-Machine interaction are both studied in a multitude of scenarios but there is still a lot of research needed to truly understand the differences between them. An example of a study that researched this difference is [10], where we can see that in fact there is a difference to the user when he is trusting a machine or another human. In [4] we see results that show differences when relying on algorithms or in other humans. As both studies were described in the Related Work it is easier to understand the influence of trusting a an human or a machine. In this study the aim is at understanding if the perception of the partner in the trust game will change the way the participant acquires trust. In order to do that, the study will have 3 different partners to play the Trust Game:

- **Autonomous Agent** - In this condition the participant is told he will play with an Agent. That agent will have an animation.
- **Artificial Intelligence Algorithm** - In this condition the participant is told he will play with an AI algorithm. The AI algorithm will have an generic Avatar.
- **Human** - In this condition the participant will be told he is playing with another human being. The Avatar will be a generic one, similar to the one the participant has. The human condition is controlled by the same AI as the other conditions, and so it has the same behaviour.

10.2.4 Conditions of the Study

The conditions of the study were labeled as shown in the table 1.

Condition	Behaviour	Partner
1	Prosocial	Player
2	Selfish	Player
3	Prosocial	Agent
4	Selfish	Agent
5	Prosocial	AI
6	Selfish	AI

Table 1: Conditions of the study.

11. Measures

In this study trust will be measured using objective measures in the game and questionnaires. In the game trust is measured by calculating what percentage of the coins owned by the participant he sends to the game. This is a metric of trust. In the questionnaire trust is measured in the questionnaire using sentences in a likert-scale to understand how the user feels about them. The questionnaire also measures if the participant feels cooperation existed, the anthropomorphism and likeability of the Agent. The number of rounds is fixed(6 rounds) and the participant is not informed of this number.

12. Participants

The study was carried out using the amazon Mechanical Turk platform using a between subjects design. In the Mechanical Turk platform settings were defined to request participants. Participants needed to have more then 5000 participation's in studies, an approval rate in those studied higher than 95% and residence in Australia, United States of America, Canada or United Kingdom. For participating in this study every person would get 1.8 dollars for correctly completing the game and a reward for the success in the game. The formula to calculate the bonus would be different depending on the behaviour they would playing with. The bonus would be calculated by multiplying the amount of coins he had at the end of the game by 0,01666(Selfish condition) or 0,00166(Prosocial behaviour) with a maximum of 1 dollar. On average participants achieved a bonus of 0.53 dollars

13. Procedure

In the beginning of the study every participant received instructions about the actions required, the payment and links to the game and to the questionnaire. At the end of the game the participant received a code to insert in the questionnaire and at the end of the questionnaire a code to insert in the Mechanical Turk platform. The study was carried out for 7 days(from 23 of April 2021 to 30 of April 2021), every one of the 6 conditions had 30 participants(in batches of 5 at a time to make sure everything was working correctly) with 6 extra participants to cover for participation's rejected due to not completing the questionnaire. In a total of 186 tests, 76.3% were valid tests. The other tests were rejected due to failed attention checks. Of the 142 valid tests 70,4% were Male, with an average age of 39.58 with a range of 20-73.

14. Results

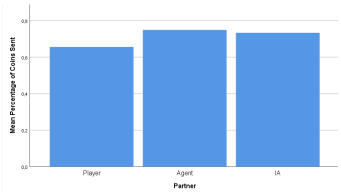
14.1. First Round

Item	F-Value	p-Value
Partner	1.078	0.343

Table 2: First Round trust Score One way Anova test.

There was no significant effect of partner in the level of trust shown in the first round($p>0.05$).

Figure 5: Trust results for first round grouped by Partner.



The first round is solely influenced by the partner, and it is the first impression the participant has of its partner since its the only characteristic that the user interacts with until that point. Because of that it is important to test the influence of the partner on the results of the first round.

One would expect the agency to have an influence on the first round of the game . But as it can be observed in table 2 it is not possible to confirm it. One possible reason of this result is that the majority of participants reported that they already had interacted with a virtual agent before this experiment. This means that this round was not the first impression they had of a virtual Agent, based on previous interactions and perhaps similar studies on the mechanical turk platform the user already had an image of virtual Agents. This means that the way they played this first round might have been influenced by the image they had from interactions with other Virtual agents or AI algorithms.

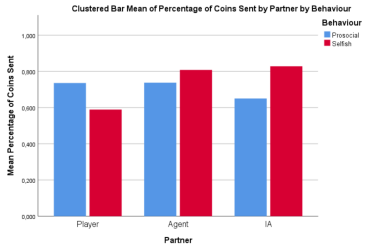
14.2. Second Round

To study the influence of the Behaviour and Partner we used a Univariate Analysis of Variance test.

As observed in Table 3, we can confirm that the combination of both conditions had an influence in the trust scores in the second round(p-Value $<$ 0.05). Neither of the conditions alone seemed to have an effect strong enough to allow a declaration of statistically significant difference.

When playing the game, only after the first round

Figure 6: Trust results for second round grouped.



Item	F-Value	p-Value
Behaviour	0.398	0.529
Partner	1.532	0.220
Partner *Behaviour	3.214	0.043

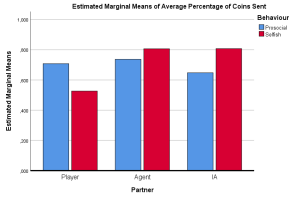
Table 3: Average Trust Score Univariate Analysis of Variance test statistics.

does the participant get a first impression of the behaviour of the partner collaborating with him.This round is important to check the reaction of the participant to said behaviour. This can be influenced by the behaviour but also by the partner or the image the participant has of it. The results show that the behaviour did not have the influence expected on the trust levels in the second round. Possibly because of the influence of Agency and the differences of a behaviour coming from a human or a machine. In order to achieve a stronger reaction, a even more selfish behaviour should be used.

14.3. Average of all Rounds

In figure 7 we can see the average results of the Trust Game by condition.

Figure 7: Trust results for the average of all rounds.



To study the influence of the Behaviour and Partner we used a Univariate Analysis of Variance test.

Item	F-Value	p-Value
Behaviour	0.121	0.728
Partner	4.140	0.018
Partner *Behaviour	5.037	0.008

Table 4: Average Trust Score Univariate Analysis of Variance test statistics.

As observed in Table 4, we can confirm that the partner condition and both conditions combined had an influence in the average scores of trust(p-Value < 0.05). With the partner condition having average trust scores of: Agent(0.772), $>$ AI(0.733) and $>$ Player (0.624).

Since trust is built over several interactions it is important to measure the influence of the conditions on the totality of the rounds. The results shows that the Partner playing the game influences how the participants play. Participants in general Considered the Agent to be more trustworthy, then the AI algorithm and lastly the human Player. This is important to see that even without the influence of the behaviour, the image and conception of the partner cooperating changes the trust in it.

14.4. Combination of Behaviour and Partner

In the book "How Humans Judge Machines" [7] the author states that the judgment made when interacting with humans and with machines is different. The author also proposes two principles and a specific effect:

- **Principle 1 - "People judge humans by their intentions and machines by their outcomes."[7]**
- **Principle 2 - "People assign extreme intentions to humans and narrow intentions to machines[7]"**
- **Effect - "People tend to judge humans more harshly in scenarios involving a lack of fairness.[7]"**

This principles helps us understand that humans judge other humans by the intention they seem to have, and that humans are more willing to excuse humans in accidental scenarios and machines in intentional scenarios. Humans also judge other humans more when lack of fairness is involved, this is due to the intention that can be attributed to the intention behind the actions of humans and machines.

In this study we can see both in Round 2 and in the Average of all rounds that there was difference when combining the Behaviour and Partner conditions. This can be attributed to:

- The player condition would have a intention of wining and a machine would not. This results

in a scenario with lack of fairness(The other player would not share the profits fairly with the participant) and as stated before in a scenario like this humans tend to judge more another human.

- Because people judge less machines in scenarios of perceived intentional. Humans are more punished for not cooperating in this scenario.

- These factors create a difference in the way participants react to a selfish behaviour depending on the partner they are playing with.

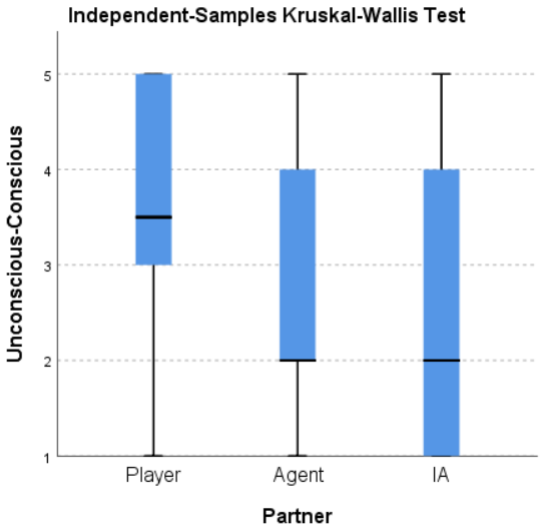
14.5. Questionnaire

14.5.1 Anthropomorphism

In the questionnaire we use the godspeed anthropomorphism questionnaire[2] that consists of 5 Likert Scales of opposite characteristics.

A Kruskal-Wallis test was executed on those Likert Scales. Fake-Natural(p-score=0.296) and Machinelike-Humanlike(p-score=0.123) showed no statistically significant difference, while Unconscious-Conscious(p-score=0.008), Artificial-Lifelike(p-score=0.03) and Moving Rigidly-Moving elegantly(p-score=0.021) showed statistically significant difference in the scores between different players.

Figure 8: Unconscious-Conscious responses.

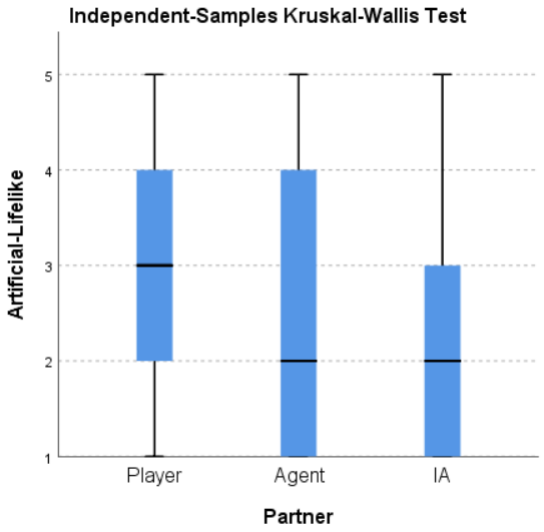


As shown in table 5 Player-AI and Agent-Player were significantly different in regards to the participants opinion on Consciousness-Unconsciousness, considering the Player the most conscious.

Sample 1	Sample2	p-Value
AI	Agent	0.382
Player	AI	0.003
Agent	Player	0.032

Table 5: Results of the tests for the influence of Agency on Unconscious-Conscious.

Figure 9: Artificial-Lifelike responses.



Sample 1	Sample2	p-Value
AI	Agent	0.228
Player	AI	0.008
Agent	Player	0.152

Table 6: Results of the tests for the influence of Agency on Artificial-Lifelike.

As shown in table 6 Player-AI were significantly different in regards to the participants opinion on Artificial Lifelike considering the Player more Lifelike.

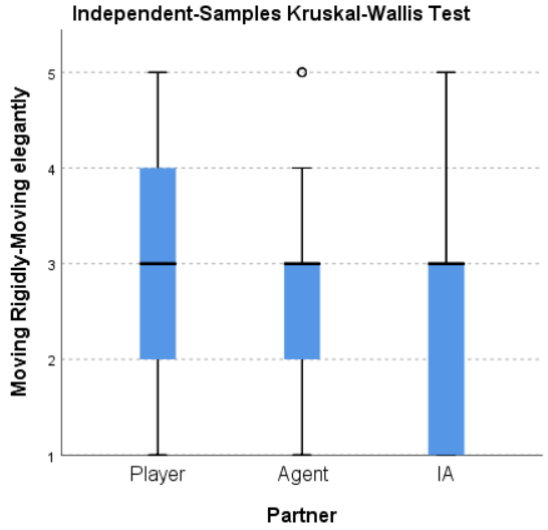
Sample 1	Sample2	p-Value
AI	Agent	0.357
Player	AI	0.007
Agent	Player	0.072

Table 7: Results of the tests for the influence of Agency on Moving Rigidly-Moving Elegantly.

As shown in table 7 Player-AI were significantly different in regards to the participants opinion on Moving Rigidly-Moving Elegantly considering the Player more Elegant Moving.

This results show that participants attribute human characteristics to the condition where they would supposedly be playing with another person.

Figure 10: Moving Rigidly-Moving elegantly responses.



This shows that the image and idea of a player playing the game was correctly conveyed. This is important in order to understand the way the participants acted.

14.5.2 Cooperation

he first part of the questionnaire aimed at accessing how the participant felt the agent cooperated with them, and how they felt they collaborated with the agent. This is important to understand how the participant views the performance of the partner playing the game.

In order to understand the influence of the Behaviour a Mann-Whitney U test was executed on the same questions.

Question	p-Value
the other player collaborated with me.	0.0
The other player tried to help me throughout the game.	0.0
The other player reciprocated my actions.	0.0
I collaborated with the other player.	0.002
I tried to help the other player throughout the game.	0.012
I reciprocated the other player's actions.	0.0

Table 8: Mann-Whitney U test results.

The results show that the Behaviour had an influence on the answers of every question.

In the results we can see that the behaviour had an influence in the collaboration reported by the participants. Participants reported they felt more collaboration from the Prosocial Partner. This is in line with the design of the study. A Partner that has a prosocial behaviour will be perceived has more collaborative. The behaviour also influenced the participant's collaboration since players that played with the Prosocial behaviour report more collaboration than those who played with the Selfish Behaviour. In the graphs it is also possible to see that participants feel the Partner did not reciprocate their actions in the Selfish behaviour.

14.6. Hypothesis

Hypothesis 1 - The behaviour of the agent will influence the player's trust in him. A more punishing behaviour from the agent will lead to less trust by the player.

In the trust game, the results can not confirm the influence of the behaviour of the agent on the player's trust. However the responses of the questionnaire showed an influence of the behaviour on the participant's perception of the collaboration. With the Prosocial behaviour being viewed as more collaborative. Responses also showed an influence of the behaviour on the perception of the intentions and if participant's would trust the other player to play this same game. The results are not enough to allow a confirmation of the hypothesis, but based on the literature it is possible to speculate that a behaviour that would be more punishing, for example instead of taking most of the profit, a behaviour that would also take coins from the participant would have a greater influence on trust (The reciprocity hypothesis that says players react to the behaviour punishing when punished[6].)

As observed in the results of the Trust Game The Autonomous Agent had the highest levels of trust(0.772 on a scale from 0 to 1), with the AI Algorithm having a lower score (0.733). The player condition had the lowest trust scores(0.624) The Partner condition also revealed to have a statistical significant difference.

Hypothesis 2 - The Autonomous Agent will have lower levels of trust than the AI Algorithm.

Hypothesis 3 - The Human partner will have lower levels of trust than the AI Algorithm and the Autonomous Agent.

The results allow us to confirm Hypothesis 3 but we cannot confirm Hypothesis 2. Results showed that The Agent had higher levels of trust than the AI Algorithm which is the opposite of the Hypothesis(Agent had a trust score of 0.772 in the Average of all rounds and AI had 0.733 in the Average of all rounds).

15. Conclusions

Technology evolves a lot every day taking up a different variety of roles in our lives, and the interactions between humans and technology are getting more diverse and complex. It is important to understand those interactions and how to profit the most from them. Trust is an important part of interactions that allows for cooperation.

In this this thesis we present a background on

Trust and Trust in interactions between Humans and Machines. A Trust Game was developed with an Agent as one of the players. study on the influence of Agency and Behaviour on Trust in Human-Agent interaction. How a Prosocial behaviour and a Selfish behaviour would influence Trust, and how cooperating with either another Human, an AI Algorithm or an Agent. Results showed that in this study it was impossible to confirm the influence of the behaviour in the trust game even though the questionnaires demonstrated a difference in the way the participant viewed the Agent. Agency was significant in the Trust Game with the Agent having higher levels of Trust and the Human partner with the lowest levels of Trust. This is relevant to understand how a person Trusts an Agent and how the Behaviour and Agency influence it.

16. Future Work

There is still a wide variety of factors that affect Trust that need to be studied and explored. One of those factors is how the user reacts when the Agent is in a situation of disadvantage for example starting with less coins or losing all his coins mid game. Another interesting thing would be to understand the difference of letting the participant be aware of the number of rounds or not, and understand if the user takes advantage of that knowledge to take advantage of the Agent. There is still a big variety of situations where Humans and Agents can interact, that have different conditions that can have a different influence on the interaction.

References

[1] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(4):1–30, 2018.

[2] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[3] J. Berg, J. Dickhaut, and K. McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.

[4] E. Bogert, A. Schecter, and R. T. Watson. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, 11(1):1–9, 2021.

[5] S. Bourgeois-Gironde and A. Corcos. Discriminating strategic reciprocity and acquired trust in the repeated trust-game. *Economics Bulletin*, 31(1):177–188, 2011.

[6] F. Cochard, P. N. Van, and M. Willinger. Trusting behavior in a repeated investment game. *Journal of Economic Behavior & Organization*, 55(1):31–44, 2004.

[7] C. A. Hidalgo, D. Orghiain, J. A. Canals, F. De Almeida, and N. Martín. *How Humans Judge Machines*. MIT Press, 2021.

[8] M. Lewis, K. Sycara, and P. Walker. The role of trust in human-robot interaction. In *Foundations of trusted autonomy*, pages 135–159. Springer, Cham, 2018.

[9] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404, 1998.

[10] S. S. Sundar and J. Kim. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*, pages 1–9, 2019.

[11] U. Technologies, 2021.

[12] M. J. Wooldridge and N. R. Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.