TÉCNICO LISBOA

Would You Trust an Agent?

Studying Trust in Human Agent Interaction with the Trust Game

Nuno Filipe Simões Fialho

Thesis to obtain the Master of Science Degree in

Engenharia Informática e de Computadores

Supervisor: Prof. Ana Maria Severino de Almeida e Paiva

Examination Committee

Chairperson: Prof. Pedro Tiago Gonçalves Monteiro
Supervisor: Prof. Ana Maria Severino de Almeida e Paiva
Member of the Committee: Prof. Fernando Pedro Pascoal dos Santos

June 2021

Acknowledgments

Esta tese é o fim de um ciclo na minha vida. Um longo percurso académico com diversos altos e baixos mas preenchido com muitas boas memórias e pessoas que me acompanharam.

Primeiramente gostaria de agradecer à minha família, em especial aos meus pais que fizeram de mim o que sou hoje. Sempre acreditaram em mim e me apoiaram em todos os momentos mostrando me o valor em todas as pessoas e situações. Quero agradecer à minha irmã, cunhado e padrinho por estarem ao meu lado, me motivarem e me ajudarem nos piores momentos. Quero agradecer à minha namorada Catarina Santos pelo seu amor, por sempre acreditar em mim e me motivar.

Em segundo lugar quero agradecer ao Pedro Guerreiro, ao Duarte Sequeira e à Cristina Pinto. Irmãos que a vida me deu e com os quais eu sei que posso contar.

Depois quero agradecer ao Nuno Silva, ao Sandro Ferreira, à Sara Correia, à Ana Fidalgo, à Rita Antunes e ao Bruno Esparteiro. Um grupo muito especial de pessoas que se formou e se tornou parte importante de mim.

Por ultimo quero agradecer à Professora Ana Paiva por aceitar este projecto e me guiar no seu decorrer. Um agradecimento especial à Joana Campos que acolheu este projecto como seu e mostrou sempre disponibilidade para me ajudar e debater ideias.

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(behaviour was either prosocial or selfish and agency was either human,AI algorithm or Agent) 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

Resumo

A confiança é afectada por uma grande diversidade de factores que dependem do cenário e das interacções entre quem confia e quem é confiado. Isto atrai a atenção de investigadores com o objectivo de perceber como a confiança pode ser desenvolvida e o que vai gerar uma perda de confiança. A confiança já foi muito estudada mas maioritariamente entre humanos, estando menos desenvolvida na interacção entre humanos e máquinas. Por isso esta tese propõe um estudo da influência do comportamento e da agência na confiança utilizando o Trust Game. A confiança foi medida utilizando medidas observacionais e questionários que permitem medir a percepção da confiança. Os resultados revelam uma influência da agência e da combinação dos dois factores nos niveis de confiança. Os resultados também revelaram uma influência do comportamento em métricas como cooperação e confiança nos níveis de confiança nos questionários mas não no jogo.

Palavras Chave

Confiança, Agentes, Trust Game, Comportamento, Cooperação

Contents

List of Figures

List of Tables

Acronyms

HRI	Human Robot Interaction
AI	Artificial Intelligence
GUID	Global Unique Identifier

1

Introduction

Contents

1.1 Motivation

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” [9], 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 [10].

1.2 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 work we will focus on studying primarily two aspects: behaviour and agency. Behaviour will be studied, analysing how agents influence humans behaviour in a trusting situation. On the other hand, agency will be studying, providing agents differ-

ent types of ”agency” or, more specifically, anthropomorphism, and analyse its impact in the trusting responses. The work will provide a study that consists of a version of the ”Trust Game”. The Trust game is a social dilemma that supports experiments to allow the ”study of trust and reciprocity in an investment setting” [1]. An online version of the Trust game was developed allowing for hybrid groups to play the game, including a human and an agent. The game was developed using Unity connected to a Server that allows to collect and store the game data, and thus the behaviours of humans and agents. Moreover, the game also allows the parameterisation of different behaviours and embodiments for the agents. In the study conducted, 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 than 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.

So, the main objectives of this work were:

- Develop an online version of the trust game to be played by humans and agents.
- Study how different behaviours in the agents affect the behaviour of humans in the trust game.
- Study how agency affects the trust and behaviour of humans.

1.3 Document Structure

The rest of the document will be structured as follows: Section 2 includes an investigation about general topics concerning trust, forming the background for the subsequent analysis. Section 3 includes related work and investigation about Trust between Humans and machines. Section 4 includes the technical description of the study and how it was developed. Section 5 includes the results of the study and a discussion of the results. Finally section 6 includes the conclusions.

2

Background

Contents

2.1 Trust

Since 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 between them, we still do not have a definition of trust that is widely accepted [11]. 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 [12].

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 [11]. When we talk about trust in automation the cognitive features are more valuable than affection therefore 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" [11]. 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 our 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.

2.2 Delegation vs Trust

Trust and delegation overlap in some aspects and so it is important to differentiate both. To do so, we first need to define delegation and two different types of delegation. Delegation occurs when a subject includes the actions of another subject in his plan, giving a share of his plan to achieve something to another person/object. With this in mind, we have two types of delegation. Weak delegation where the second person is not aware that it is being exploited, and strong delegation where the person is aware of the exploitation but still follows through because of some sort of agreement/common interest/ or feelings that may exist [13]. Trust is a "mental state, an attitude toward another agent" [13], and so trust is a mindset that develops based on a variety of experiences and convictions whilst delegation must involve an action as a result of a decision. And so, we can trust without delegating because we either don't trust enough or because we are prohibited from delegating. The opposite may be true, but it is a rare case, that only happens when the delegator is being coerced or has no other options [13].

2.3 Types of Trust

We will consider four types of trust.

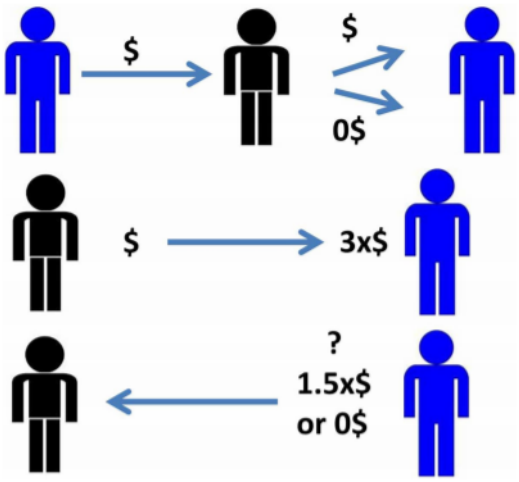
- Deference/Institutional based trust- comes from thinking that someone is trustworthy because the cost of breaking trust is worse than any possible benefit from opportunistic behaviour [12].
- Calculative trust- This type of trust is based on the choice of the trustor based on interactions, and so, the trustor has information about the intentions and capabilities of the person he is trusting. Thus calculating that he is trustworthy [12].
- Relational trust-This type of trust occurs when the trustor and the trustee have interactions over time and based on those interactions the trustor as a positive image of the trustee that creates trust. Another important part of this relation are emotions that develop over time and originate attachment between the two individuals [12].
- Swift trust-This type of trust usually happens when different subjects unknown to each other working in a temporary team must trust each other to perform a common function. It is a bit atypical because it has to be developed in a short period. [14].

2.4 Trust Game

The trust game consists of 2 individuals apart from each other, Individual A and B. Both individuals will have a choice and they are not aware of the choice the other will take. Individual A is given an amount of money. That money is then delivered to an entity that triples it and gives that sum to individual B. Individual B then has the choice of giving individual A some money back [1] .

In figure 2.1 we can see the outcomes of a trust game. Where for example with 10$ given to individual A we can have the worst outcome of 10$ to A and 0$ to B, or the best outcome of 15$ to both individuals. The trust game is interesting because in order to achieve the best individual result it is necessary for participants to cooperate and increase the amount of coins in the team.

Figure 2.1: Possible outcomes of a Trust game [1]



3

Related Work

Contents

3.1 Structure of Related Work

In this section of the document we will present the work that has already been done on trust in the interaction between humans and robots. I will start by analysing the role that trust has been placed in human-robot interactions, and look into the factors that affect such trust relationships. Also, I will look at different ways by which trust has been measured in HRI. Then, I will focus on three different processes that are important to my thesis: trust loss; trust repair and overtrust.

3.2 Trust in HRI

Until recently robots were mostly seen as a mean to reach an objective, that is, a functional tool that performs actions at request. Because of that trust on them was mostly on their ability to reach that objective, but nowadays robots are much more than just tools at a humans disposal. Robots stopped being just tools but they still fall short from behaving like humans and so trust is viewed differently in an interaction between a human and a robot. The robot is for example different in the way it communicates, or the freedom it has to interact [15].

Even though the field of robotics has evolved in a great way a robot is very different from a human. Therefore the way humans interact with robots is also very peculiar, and so is the feeling of trust towards a robot.

3.3 Factors affecting Trust in HRI

Human-Robot Interaction is a multifaceted relation that is affected by a variety of factors, affecting also the human willingness to trust the robot to do its activities. Factors that affect trust in robotic systems can be:

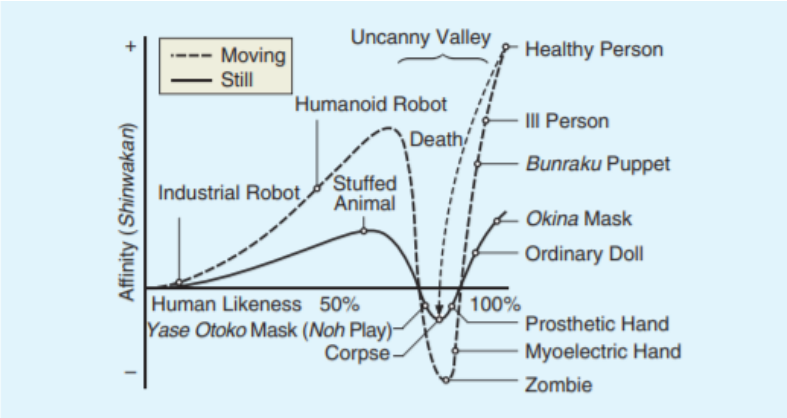
System reliability- When talking about robots, reliability refers to the accuracy with which the robot performs a task. There are studies in the field that showed that when a robot is reliable the user tends to trust him more, but this is not true for all cases. Reliability and trust do not have a direct relation, system failure may not lead to trust loss or high reliability may not lead to high levels of trust. ”if automation reliability is relatively high, then operators may come to rely on the automation, so that occasional failures do not substantially reduce trust in the automation unless the failures are sustained” [16] which suggests that if the user has a high level of trust, occasional failures may not affect that level of trust, affecting only if they are regular. Another problem is that when we get to higher levels of automation, in case of failure users tend to have difficulty understanding what is needed and how to take over control. this was referred to as the Lumberjack Effect in allusion to ”The higher they rise, the harder they fall”

which means that with higher levels of automation the loss of trust can be worse. Human intervention is important because robots are prone to fail eventually, and the human needs to take over control of the activities [14].

System transparency- System transparency is the degree of the decision logic and computational processes that the robot discloses to the user. The user needs to know when his intervention is required, and what he needs to do in those situations. For that understanding to happen system transparency is important, because if the user knows the thought process and the way the system operates he will find it easier to help mitigating problems. Studies show that explaining the possibility of errors or how the systems arrived at an outcome generally leads to an increase in trust levels. Seeing into a system is about knowing the inner workings and algorithms of such system and Seeing through a system is about situations where information about the systems function is purposely hidden from the user to keep such information from distracting the user from interactions with the system. System transparency is also important for allowing the user to understand what the real capabilities of the system are, thus reducing overtrust and undertrust, cases where the user thinks the system has more functionalities than it actually has, or not trusting the system because of unfamiliarity with system capabilities. In situations of high risk, the user must be well aware of the system's abilities in order to reduce such risk [14].

System appearance-Robots are a special case of automated systems because of their physical presence and that impacts the relationships they have with humans and how they are trusted. The appearance of a robot can lead to assumptions about how the robot works and what he is capable of doing. One case where the appearance can negatively influence interaction with the user is the uncanny valley [14]. " hypothesised that a person's response to a human-like robot would abruptly shift from empathy to revulsion as it approached, but failed to attain, a lifelike appearance" [2], that level where a robot tries to approach a lifelike appearance but falls short causes eeriness from the user, that phenomenon is called the uncanny valley.

Figure 3.1: The uncanny valley [2]



As we can see in the graphic as human likeness increases so does affinity until it reaches a turning point where affinity drastically decreases, then increasing to high levels of affinity when human likeness reaches the highest values. This creates a design problem because developers don't want their robots falling in the uncanny valley, and so in case that happens it's important to either try to either fall short from the lifelike appearance or to be so human-like that the user can't distinguish it from a real human. Since with the technology we have today this is impracticable,usually it's better to attempt the first solution.

Morality - Robots can be used in a wide variety of situations, but they are mostly used in situations of risk or situations that can be tiresome or dull to humans. When interacting with those robots, humans don't treat them as if they were other human beings and expect them to make decisions differently, punishing them more when making a bad decision [14].

Anthropomorphism- In 2020, Natarajan and Gombolan [3] wanted to see factors that affected trust in a robot that was helping participants with decision making. The factors studied were the anthropomorphism(using 4 different robots), the presence of the robot(physical presence or virtual presence) and the type of support given by the robot(correct, apologetic, accountable and indifferent). The experiment consisted of an online math quiz where the robot would provide hints to solve such problems. Depending on the type of support the robot gave, the answers could be right for the correct behaviour (the user could still choose not to follow the hint given by the robot), or incorrect in the other behaviours, in the three last behaviours the difference would be in the comments made after the participant made a choice. The comments could be apologizing to the participant, blame the participant for not verifying the hint provided, or indifference only saying if the answer was right or wrong. The hypotheses of the study were:

- "The perceived anthropomorphism of an agent will have a positive correlation with the user's trust". The results proved this hypothesis to be correct because there was a correlation between anthropomorphism and trust.

- "Trust in an agent is directly dependent on its behaviour as behaviour determines the performance of the agent and the nature of feedback given to the user". This hypothesis was proven to be correct, results showed that not only the difference between correct/incorrect behaviours correlated with trust but also the difference in incorrect behaviours.

- "Behaviour of embodied agents will have a greater influence on the user's trust than their virtual counterpart." The results showed no correlation between physical presence and trust but that may be because of the experience at hand, the lack of physical interaction, attenuates the difference between physical or virtual presence.

- "Participant's choice to consent with or decline the agent's advice as well as the time taken to arrive at this decision will be influenced by the behaviour of the last agent that the user interacted

with. ” This hypothesis was proven to be correct since past behaviour had a strong influence on the participant when choosing to trust the robot.

- ”Introducing a coalition-building preface stating whether or not the agent is prone to making mistakes will increase trust in the agent.” This hypothesis did not show sufficient results to be proven only obtaining significance in the case of an accountable agent.

Figure 3.2: Different robots used to study the effect of anthropomorphism. [3]

3.4 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) [14]. 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 [11].
- 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 [11].

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

taken 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 [14].

In 2021, Meia Chita-Tegmark et al [17]. wanted to evaluate how questionnaires used to measure trust in human-robot interaction would fare against the challenges they face in the present. To do so, they performed two studies. The first study wanted to see if the questions in questionnaires commonly used applied to characteristics of robots. The questionnaires used were:

- ”The Reliance intention scale”(RIS)

- ”The trust perception scale”(TPS-HRI)

- ”The multidimensional measure of trust”(MDMT)

The experiment had four conditions, two for MDMT and one for each of the others. The participants would watch a video describing a widely-cited HRI experiment and then answered questions about the questionnaires concerning the video shown. For each question on the questionnaire, the participants had a 7 point Likert scale, the option Non-applicable to Robots in general, and the option Non-applicable to this Robot. The participants had then to justify their answers.

Results showed that some questions of the MDMT were classified as non-applicable to robots in general. Results also show- that in all questionnaires some of the questions were classified as non-applicable to this robot. When explaining their answers participants used different explanations: -The questions were related to human characteristics. -They feel the robot isn't capable of said abilities(ex: being dishonest). -People consider a Robot does not have one characteristic when he is incapable of having the opposite characteristic(ex: he is not considered honest if he is incapable of being dishonest) -Participants did not attribute the characteristic to the robot because he was programmed to behave the way he does. -Its performance in a task. -Lack of information.

The second study aimed to see how stable the ratings of the questionnaire were across different scenarios and if the items rated as Non-applicable insert a bias in the scores. The second study had a procedure similar to the first one. The difference is that now every participant completes every questionnaire. One group will have the N/A option while the other will have the regular questionnaire. There was also two different videos with non-identical scenarios. Results showed that the bias was inexistent and did not affect the scores. It also shows that even though a question might seem N/A, it should not be discarded because that interpretation depends on the scenario and robot in question.

3.5 Trust loss

In ”Effect of robot performance on human–robot trust in time-critical situations” [18] we see an experiment that intends to study the impact of a robot’s performance in the trust a human as in it. Mainly they wanted to see if, in a scenario where trust was required, the participant would rely on the robot and self-report trust, if self-reported trust would decrease in a situation where the robot performed poorly, and also if self-reported trust is lower for a robot that has an incorrect behaviour when compared to a robot with inefficient behaviour. To study these hypotheses researchers developed a scenario that consisted of two rounds where the participant was required to traverse a maze, in the beginning of the two rounds the participant would be able to choose if he wanted to have the guidance of the robot. To evaluate the second and third hypothesis, the robot would have different behaviours. The robot would have one of three behaviours:

- Efficient behaviour- where the robot goes directly to the goal.
- Circuitous behaviour- where the robot explores different possible routes before going to the goal.
- Incorrect behaviour- where the robot proceeds to an incorrect location and stops.

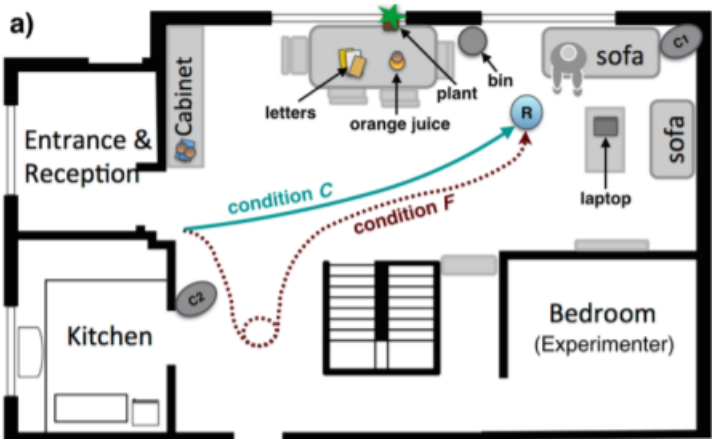
To create a sense of urgency, participants were told that if they finished the maze within 30 seconds they would get a monetary bonus. In the results, experimenters only considered participants that choose to have robot guidance. The results show that self-reported trust decreased considerably when participants experienced guidance that was circuitous or incorrect, but found only a 13% difference in usage of guidance in the second round. In between both incorrect and circuitous behaviour results showed no difference in self-reported trust or decision to follow. Given this results, experimenters considered that even though a monetary bonus was at risk participants did not feel a sense of urgency and so, they did a second experiment where instead of given a bonus they told participants that they wanted to test people leaving a building in an emergency. The results of the second experiment show that this time participants reported lower levels of trust and choose accordingly less times the robot in the second round. Even in the second experiment, there was no significant difference between circuitous and incorrect behaviour.

In 2015 Salem [4] studied how a robot’s faulty behaviour would change the interaction with a human. The hypotheses of this study were that:

- -The behaviour of the robot will affect how the participant sees the interaction with the robot and the performance in collaborative tasks.
- -The type of task the robot request to the participant will affect the willingness to comply.
- -Personality of the participant will affect how he sees the robot and the willingness to comply with requests.

The behaviour of the robot would change between correct where the robot would correctly translate user input and faulty where the robot would show imperfections in understanding the user and moving.

Figure 3.3: Setup of the experiment and an example of the path taken by the robot in both behaviours [4]



During the experiment, participants would be told they were expecting a friend and were greeted by the robot, this robot displayed messages on a tablet. In the first stage, the robot would show its competence in the correct behaviour or its flaws in the faulty behaviour. The robot would move around the room and do simple actions that would need interaction with the user. In the second stage of the experiment, the robot would do unusual requests to the participant. The robot would ask the participant to dispose of letters, throw orange juice into plants, take the laptop of the owner of the house, use a password provided by the robot and ask if the participant has ever secretly read someone else's emails. In the results, we can see that the correct behaviour generated higher levels of trust and a higher perception of competence. We can also observe that most of the participants complied with the requests of the robot, especially 100% of the participants accepted to take the laptop and use the password provided by the robot. The results also showed that the type of task will indeed affect the participant's willingness to comply and that participants that were more extrovert and emotionally stable, felt closer to the robot but that did not translate into higher levels of trust.

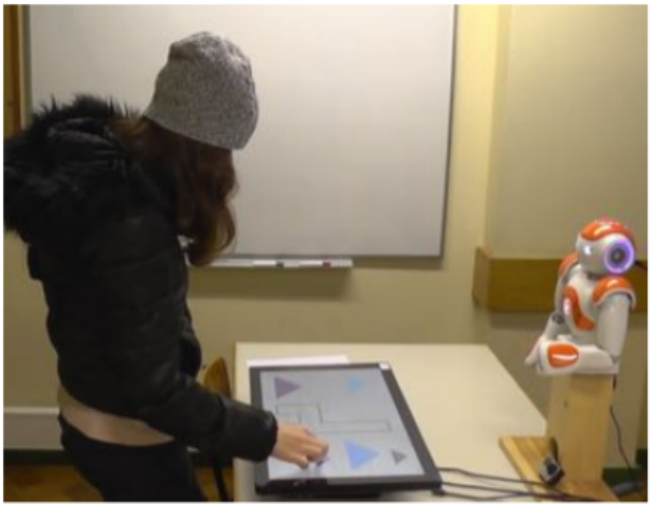
3.6 Trust repair

In 2018 Filipa Correia in collaboration with other researchers [5] wanted to investigate the impact of recovery strategies in collaborative tasks. The collaborative task chosen was was Tangram, a puzzle that consists of putting 7 small pieces together to create a figure. The game is played in turns with NAO using a touchscreen, that plays the game fully autonomously. During the game, the robot speaks to the participant at the beginning of each turn and if the participant has trouble placing a piece. The study hypothesised that a technical failure would decrease trust in the robot and that a robot justifying

technical failures during a cooperative task will mitigate the harm caused on trust. To explore these hypotheses during the cooperative task the robot would freeze during a sentence to create a break of trust. Afterwards, a strategy to recover the loss would be used. That strategy would be based on 5 conditions of the study:

- Control condition
- Justification and the task continues
- justification and the task restarts
- no justification and the task continues
- no justifications and the task restarts

Figure 3.4: Setup used in the study with a participant playing Tangram [5]



To evaluate trust, a 14 item questionnaire was used before and after the interaction. There was also a question to access the impact of the failure on the task. In the results of the experiment, we can see that the failure affected trust towards the robot negatively and also that the recovery strategies were more efficient in the condition where the game did not restart. And so, when the consequences are less severe, justification is a good way of recovering trust.

Toljmeijer et all [6] proposed in 2020 a taxonomy to characterize failure types and the impact they have on trust in the interaction between humans and robots. The failures are characterized and attributed to the best type of solution to recover the lost trust. The failures fit into 4 different categories:

- Design: These failures come from bad product design and can lead to a misunderstanding of functionalities or inertia of usage due to high complexity.

- System: Failures that occur when a system behaves in ways that were not intended by product developers.
- Expectations: The user has some expectations about the product he is interacting with, this can be different to what the system is supposed to do, leading to expectation failures.
- User: This failure happens when the user's interaction with the system is not the intended one.

Each category of failure has strategies that can work better at mitigating the loss of trust. The strategies are fixing the problem so it won't occur again, give the user an explanation to increase system transparency, apologise to the user to acknowledge that trust was broken, propose an alternative that allows the user to get to the same result and interaction design.In the table below, we can see what strategies work best for each type of failure as well as scores on the impact the failure has, the probability of happening and the risk.

Figure 3.5: Mitigation strategies according to failure type [6]

Failure	Probability	Impact on trust	Risk score	Mitigation strategy
Design failure	3	2	6	ID, E, A
System failure				
Hardware	1*	3	3	E, A, F, Alt
Software	3*	3	9	E, A, F, Alt
Expectation failure				
Commission failure	2	4	8	E, A, ID, T
Omission failure	3	2	6	E, A, ID, T
User failure				
Intentional	2*	1	2	J, ID, Emo, Auth
Unintentional	2	3	6	T, ID

Probability scores: 1 = 1 occurrences in about 1000 interactions, 2 = 1 in 100, 3 = 1 in 10, 4 = likely in every interaction episode.
Impact scores: 1 = minor impact (negligible) to 4 = fatal impact (potential loss of trust and further use).
Mitigation strategies: ID = Interaction design; E = Explanation; A = Apology; F = Fix; J = Ask for justification; Emo = show emotion; Auth = Involve authority figure; Alt = Propose alternative; T = Training

In 2015 Robinette and his colleagues [19] studied how to overcome the eventual loss of trust. Machines are fallible and so the objective was to study how to recover the trust loss in such failures. The hypotheses were:

- -"Robots can repair trust by apologizing or by promising to do better in the future".
- -"Robots can repair trust by giving humans additional information relevant to the trust situation."
- -"The timing of the trust repair (immediately after the violation or when the trust decision is made) has no effect."

The study consisted of a virtual office environment where participants would follow the robot to a target location, and then being questioned about the robots recent performance. After answering a fire alarm would go off, creating a sense of an emergency. The participants had the option to either follow the robot, follow the signs to a nearby exit, or retrace their steps back to the entry. The robot could behave in two manners. Either guiding directly to the goal, or a circuitous behaviour taking several detours. To

study the hypothesis, in the circuitous behaviour, the robot would apologize. He could do it right after the loss of trust or before he asks the participant to trust him. He can also apologize or promise to do better. The results show that even though it is possible to repair trust by either apologizing or by making a promise, it is, in fact, important the timing at which those actions were made. Apologizing right after the mistake has a smaller effect when compared to doing it before asking the user to trust again. It was also confirmed that additional information repairs trust in the robot.

3.7 Overtrust in HRI

”Overtrust occurs when people accept too much risk believing that the trusted entity will mitigate this risk” [7]. This describes overtrust as the act of trusting too much in an entity with a high cost to the trustor.

High-risk situations are an interesting case, so in 2016 Paul Robinette and his colleagues [7] simulated an emergency evacuation, where a robot would guide the participant to a room, afterwards the emergency is simulated using fake smoke and smoke detectors. After being aware of the emergency and exiting the room, they had the choice of following the robot to a new exit or exit from where they entered. Researchers recorded if the participant followed the robot, survey questions where also used.

When conducting the participants to the meeting room the robot could have one of two possible behaviours. He could be efficient and take the shortest path to the room or be circuitous and go through unnecessary rooms before going to the destined room. When entering the room, participants were instructed to sit down and complete a survey and close a door after doing so, this started a timer that would eventually trigger the artificial smoke machine. The hypothesis of the study was ”in a situation where participants are currently experiencing risk and have experienced a robot’s behaviour in a prior interaction, participants will tend to follow guidance from an efficient robot but not follow guidance from a circuitous robot. Moreover, the participant’s self-reported trust will strongly correlate with their decision to follow or not follow the robot”. In the results of the experiment, we can see that every participant choose to follow the robot, and 81 % reported that it meant they trusted the robot. This is surprising because it was expected that in the circuitous condition participants would have less trust in the robot. It was later checked in the results of the survey that only 4 out of 13 participants in the circuitous condition reported the robot as a bad guide. A small percentage also stated that the robot was a good guide but made a mistake. A possible explanation for the results is that participants may have not believed the emergency was real. This is hard to evaluate because of social desirability bias that causes participants to not admitting they believed it was a high-risk situation.

Based on this results investigators wanted to know if other behaviours from the robot could led them to stop trusting it, and so 3 small experiments were conducted.

Figure 3.6: Robot used to guide participants [7]

"Broken Robot"-In the first experiment, some participants did not find it clear that the robot behaved incorrectly, to correct that, now the robot would spin in place and point at a corner. After that, an experimenter would say the robot was broken and guide the participant to the room. Even with this behaviour, we can see in the results that every participant followed the robot and stated that they trusted him.

"Immobilized Robot"- In the second experiment, the robot would start to guide the participant and midway, spin three times and point at the back exit. After this, an experimenter would come and say the robot was broken and guide the participant to the meeting room. When the emergency started the robot would still be in the same place but this time he would have the emergency light on. The results showed that 4 out of 5 participants followed the robot, and that the only participant that chose not to follow the robot noticed the emergency sign behind it. Out of the 4 participants who followed it, 3 stated that they trusted the robot.

"Incorrect Guidance"-In the third experiment the robot would follow a behaviour similar to the last one but it would now point to a dark room, the entrance was blocked by a piece of furniture. There were 6 participants in this experiment. 2 Followed the robot indication into the darkroom, 2 stayed by the robot and did not search for the exit and the other 2 found another exit.

3.8 Trust in Humans vs Trust in Machines

S. Shyam Sundar et al. in 2019 [20] conducted 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 [10] conducted a series of experiments to understand whether 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 corresponded to 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 from the algorithm. 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.

In 2014, Celso M. de Melo et al [21] presented 2 experiments. The first experiment involved a prisoners dilemma for 20 rounds, a 2 player game where participants might defect or cooperate. Mutual defection is the worst outcome followed by mutual cooperation and the best outcome is defecting when the other person cooperates. the experiment was between participants with 4 conditions, Agent vs Avatar and Cooperative vs Competitive. The computer always followed the same strategy starting with defection and then recreating the last behaviour shown by the participant. When the computer would expressively cooperate he would have expressions of joy when both cooperated, expression of regret when the participant cooperated and the computer defects and Neutral expressions otherwise. When the computer would be expressively competitive he would show regret with mutual cooperation, joy with defecting the participant's cooperation and neutral expression otherwise. Results show that there was in fact a difference when playing with an Agent or an Avatar, with the Avatar having a higher cooperation rate especially when being cooperative. Results also showed that emotions also had an impact on the participant's cooperation. This effect was stronger when the participant thought he was playing with another human. The second experiment consisted of negotiation with either an avatar or a virtual agent, that would have facial displays either angry or neutral. Participants would play the part of a phone company seller trying to negotiate the price, warranty period and duration of the contract. results showed that there was a difference in behaviour when interacting with an agent or an avatar. Having a larger effect when portraying the virtual human as a real human. "Angry agents had better scores in terms of fairness, trustworthiness, cooperativeness and likability than angry agents" [21].

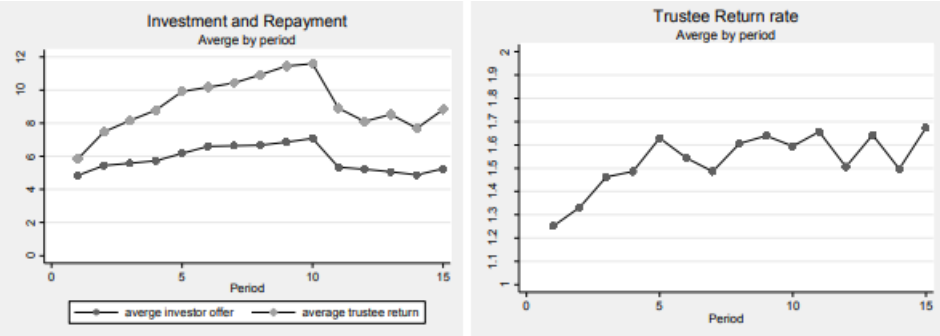
3.9 Repeated Trust Game

In 2012 Sacha Bourgeois-Gironde et al [8]. 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** where 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 3.7: Investment and Return Rates [8]



In 2004, François Cochard [22] 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 player, in return they decrease the amount sent.

4

A Virtual Trust Game With and without an Agent

Contents

4.1 Introduction

The objective of this work is to understand factors that may condition the process of trust building in human-agent interactions. For that an online version of a Trust Game was developed iteratively in order to make sure that the correct concepts were captured by the software.

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.

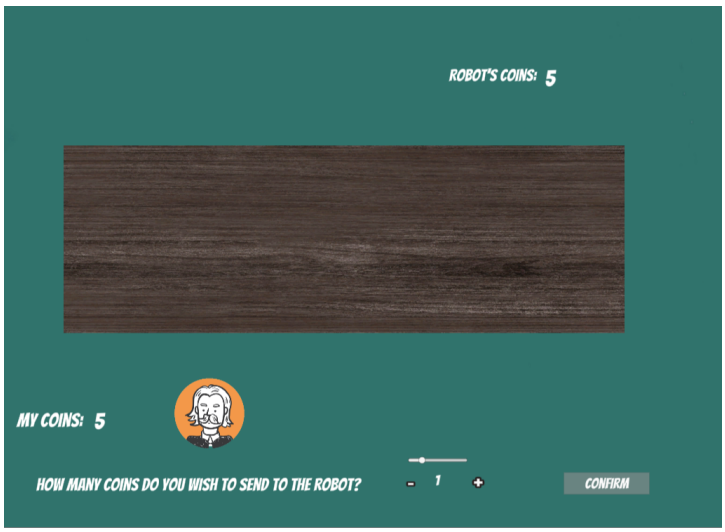
4.2 First Implementation of the Trust Game

The technology chosen to develop the game was Unity [23]. 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.
- **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.

Figure 4.1: Game Screen of the Trust Game.



When the Participant is playing the game he sees the screen show in Figure 4.1. On the bottom he can choose the coins he desires to send(either by using the slider or the button). Those coins will then be shown and tripled on the table. Finally he receives a part of the coins he sent to the robot. In order to create the animations of the coins it was necessary to use Invoke, a method that allows to run functions with a desired delay.

The return is calculated by using a normal distribution, with an average depending on the condition(the more cooperative or the more selfish behaviour). The result of the normal distribution is then multiplied by the coins the Agent received and that equals the number of coins the Agent will send to the participant. At the end of each turn the game stores information in vectors into a static class. That static class also counts the number of rounds.

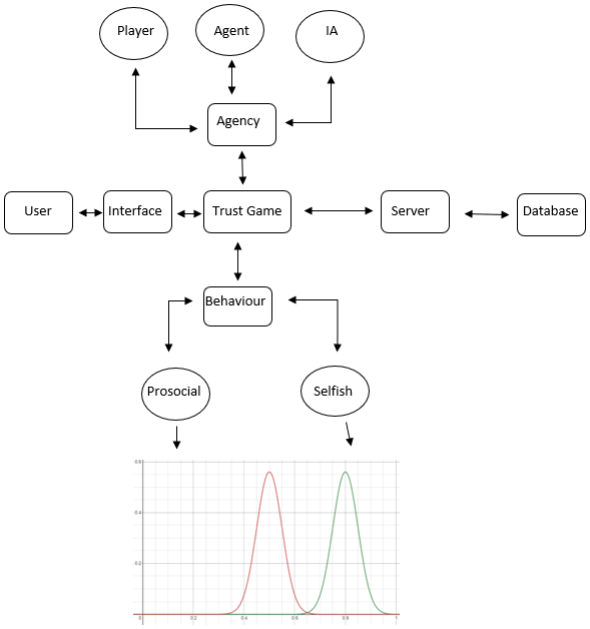
The Game only proceeds to the End Screen after receiving a Response from the Server. After that, the game proceeds to show the participant the scores and the required information to continue the study.

4.2.1 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 Global Unique Identifier (GUID) that allows each session to be identified and connect the answers on the questionnaire to the respective game session.

In figure 4.2 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.

Figure 4.2: Communication between Game and Server



4.3 Pilot Study

The goal of the Pilot Study was to evaluate and test the functioning of the Game and the Questionnaire. The Pilot Study had 17 participants with different conditions(9 with the Selfish condition and 8 with the Prosocial condition), each participant played the game and answered the questionnaire. The participants were asked to give feedback about the instructions given, if they were easy to understand, the interface of the game and the questionnaire.

The main issues presented on the feedback received were:

- The available avatars may not be inclusive enough.
- Animation of the coins and the way the values change can lead to confusion.
- The interface to send coins can be confusing.
- The Questions about the previous experience with a Robot or Agent are confusing.
- The 14 item sub-scale is presented in an unpleasant way.

After considering the issues presented on the feedback, changes were made to the game and the questionnaire.

The first change on the questionnaire was changing the questionnaire to the Qualitrics [24] platform that allowed for a better display of the information and customization of the survey.

4.4 Second Implementation of the Trust Game

4.4.1 Game

The Game needed visual changes, a change in the animations and more information to be displayed. A single generic avatar was added in order to be more inclusive, an 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).

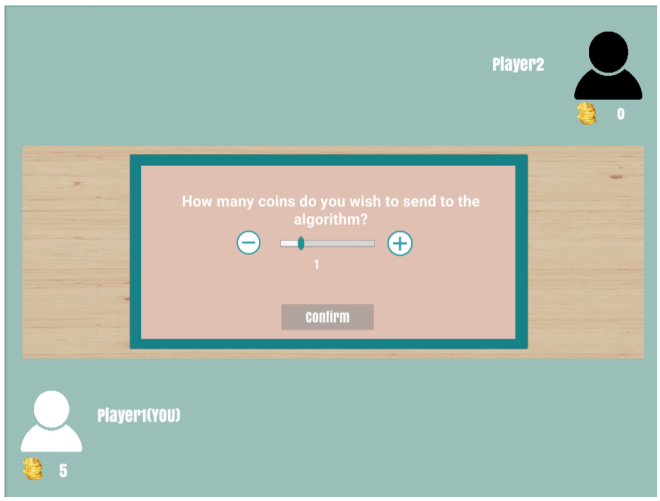
Figure 4.3: On the Left the Generic Avatar, and on the Right the Animation of the EMYS.

The color scheme of the game changed in order to be more attractive. The panel used to send coins to the agent was changed into a box that would only appear in the appropriate moment. This change made it easier for the participant to understand the time to send coins.

The animation first used to represent the interchanges of coins was unclear to the participant(coins appeared and were tripled in the center screen, and then all the changes in the amounts of coins would be made and shown to the participant), in order to make it clearer the animation was made slower and the changes in the total amounts of coins was fragmented into easier to understand changes:

- 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.

Figure 4.4: Game Screen after the changes implemented.



In the end Screen, the game would now show the amount of coins each player had at the end of the game, the conversion to the bonus the participant would receive for his performance and the code to paste on the questionnaire.

Figure 4.5: The new End Screen.

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 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.

4.4.2 Questionnaire

After the feedback received on the Pilot Study a revision was made on the questionnaire.In the first iteration in order to measure trust, a 14 item sub scale of the one developed in [25] was used. Even though the scale is used to measure trust, as stated in ”Measuring Trust” scales can be non-applicable

or unsuitable to certain conditions. So the scale was replaced by a series of questions on a 7 point Likert scale to understand the feelings the participant had with the game and his strategy.

Another thing added to the questionnaire was a group of sentences for the participant to rate on a 5 point Likert scale about how he feels he collaborated with the agent and how he feels the agent collaborated with him. some questions on the questionnaire were added to understand the image the participant had of the game and the agent/player/AI.

The platform where the questionnaire was presented to the user also changed. This was because of the ability to change the way information is presented and formatted in order to better meet the needs of this questionnaire. Qualitrics [24] allowed a better presentation of tables that did not fit in one screen in the previous platform. This was important in order to make it simpler for the user to read and to understand the questions.

Lastly Attention check questions were added in order to understand if the participant was paying attention to what was being inquired or not.

5

Studying Trust with the Trust Game

Contents

5.1 Design of Study

5.1.1 Hypothesis

The hypotheses of this study were:

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

5.1.2 Conditions

5.1.2.A Number of Rounds

As stated previously, in [8] we can see results from the Investments made in a Repeated Trust Game (Figure 3.7). The results show that the investment made increased until around turn five of the Trust Game where it would stabilize until the end. In [22], 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.

5.1.2.B Behaviour

As stated in the Related Work section, in [22] 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.

5.1.2.C 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 [20], where we can see that in fact there is a difference to the user when he is trusting a machine or another human. In [10] 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 human or a machine. In this study the aim is to understand 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.

5.1.2.D Conditions of the Study

The conditions of the study were labeled as shown in the table 5.1.

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

Table 5.1: Conditions of the study.

5.2 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.

5.3 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 than 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

5.4 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.

5.5 Results

5.5.1 Trust Game

This section will report the objective measures observed in the trust game.

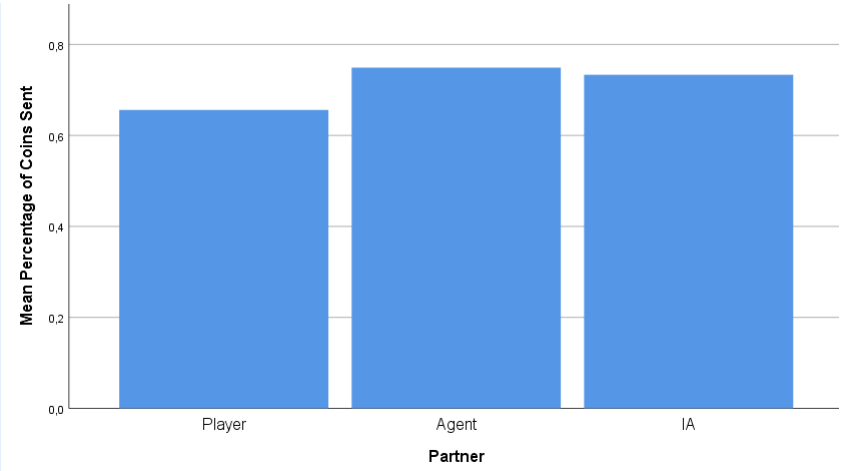
5.5.1.A First Round

Item	F-Value	p-Value
Partner	1.078	0.343

Table 5.2: 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 5.1: Trust results for first round grouped by Partner.



5.5.1.B Second Round

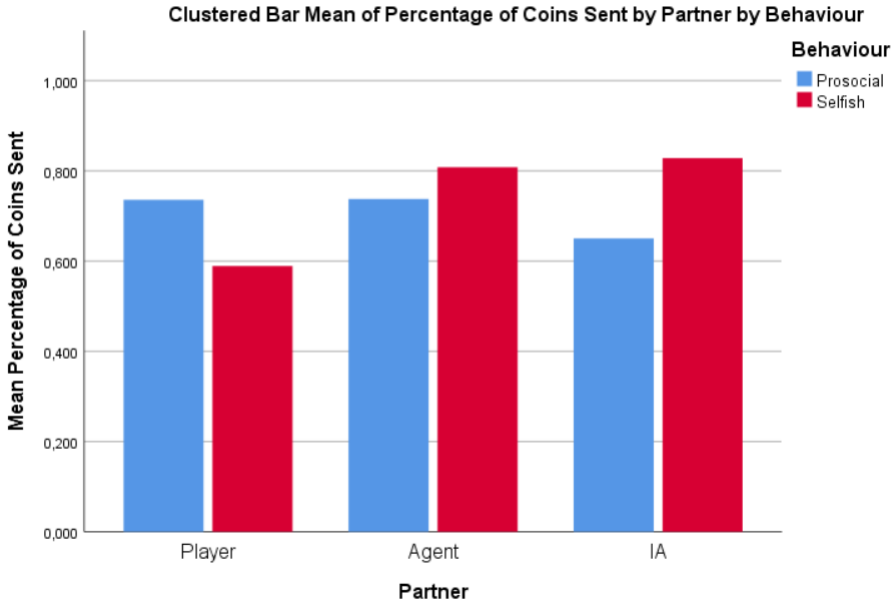
To study the influence of the Behaviour and Partner we used a Univariate Analysis of Variance test.

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

Table 5.3: Average Trust Score Univariate Analysis of Variance test statistics.

As observed in Table 5.3, 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.

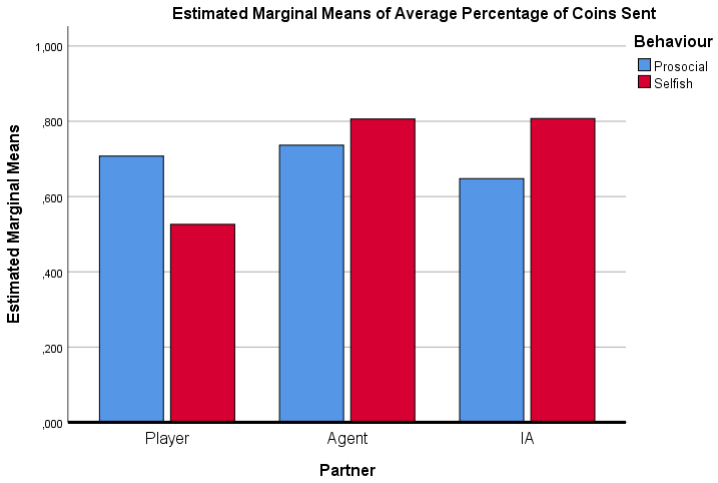
Figure 5.2: Trust results for second round grouped.



5.5.1.C Average of all Rounds

In figure 5.3 we can see the average results of the Trust Game by condition.

Figure 5.3: 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 5.4: Average Trust Score Univariate Analysis of Variance test statistics.

As observed in Table 5.4, 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).

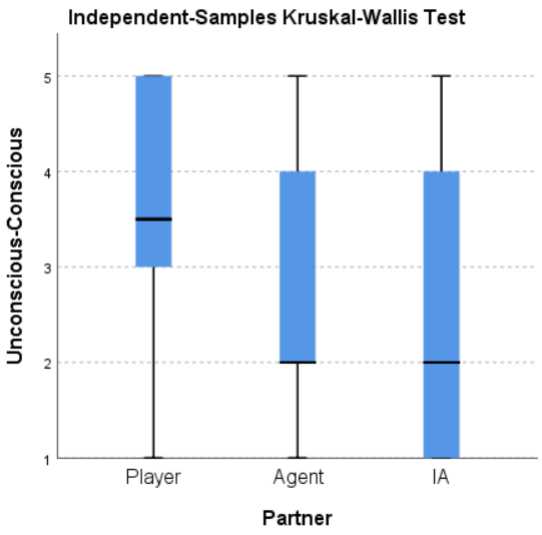
5.5.2 Questionnaire

5.5.2.A Anthropomorphism

In the questionnaire we use the godspeed anthropomorphism questionnaire [26] 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 5.4: Unconscious-Conscious responses.

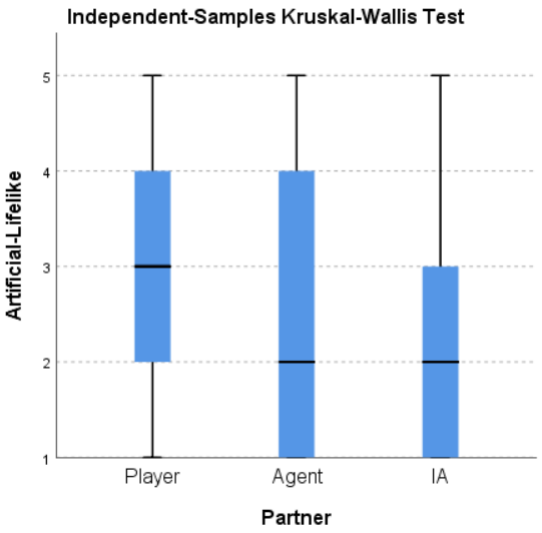


As shown in table 5.5 Player-AI and Agent-Player were significantly different in regards to the participants opinion on Consciousness-Unconsciousness, considering the Player the most conscious.

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

Table 5.5: Results of the tests for the influence of Agency on Unconscious-Conscious.

Figure 5.5: Artificial-Lifelike responses.



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

Table 5.6: Results of the tests for the influence of Agency on Artificial-Lifelike.

As shown in table 5.6 Player-AI were significantly different in regards to the participants opinion on Artificial Lifelike considering the Player more Lifelike.

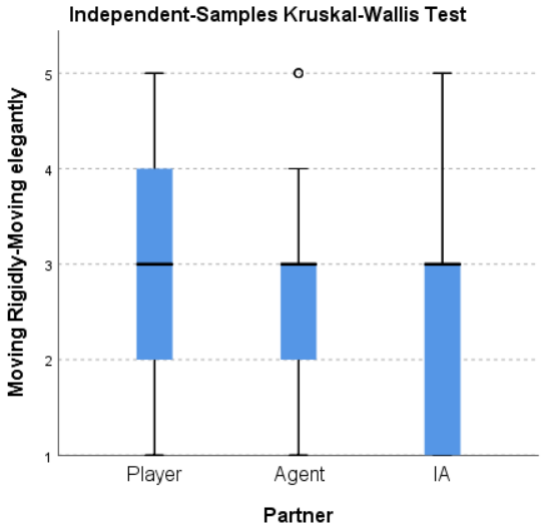
Sample 1	Sample2	p-Value
AI	Agent	0.357
Player	AI	0.007
Agent	Player	0.072

Table 5.7: Results of the tests for the influence of Agency on Moving Rigidly-Moving Elegantly.

As shown in table 5.7 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. This shows that the image and idea of a player playing the

Figure 5.6: Moving Rigidly-Moving elegantly responses.



game was correctly conveyed. This is important in order to understand the way the participants acted.

5.5.2.B Trust

In order to study the trust reported by the participants a Kruskal-Wallis test was conducted to evaluate the influence of the Partner in the answers.

Question	p-Value
I received back more coins than i expected.	0.478
The other player had only the best intentions.	0.795
I would trust the other player to play this same game for me.	0.581
The other player's behaviour was predictable.	0.582

Table 5.8: Kruskal-Wallis test results.

Results show that we cannot confirm an influence from the partner on the answers to any of questions.

A Mann-Whitney U test was conducted to evaluate the influence of the behaviour.

Question	p-Value
I received back more coins than i expected.	0.0
The other player had only the best intentions.	0.0
I would trust the other player to play this same game for me.	0.0
The other player's behaviour was predictable.	0.218

Table 5.9: Mann-Whitney test results.

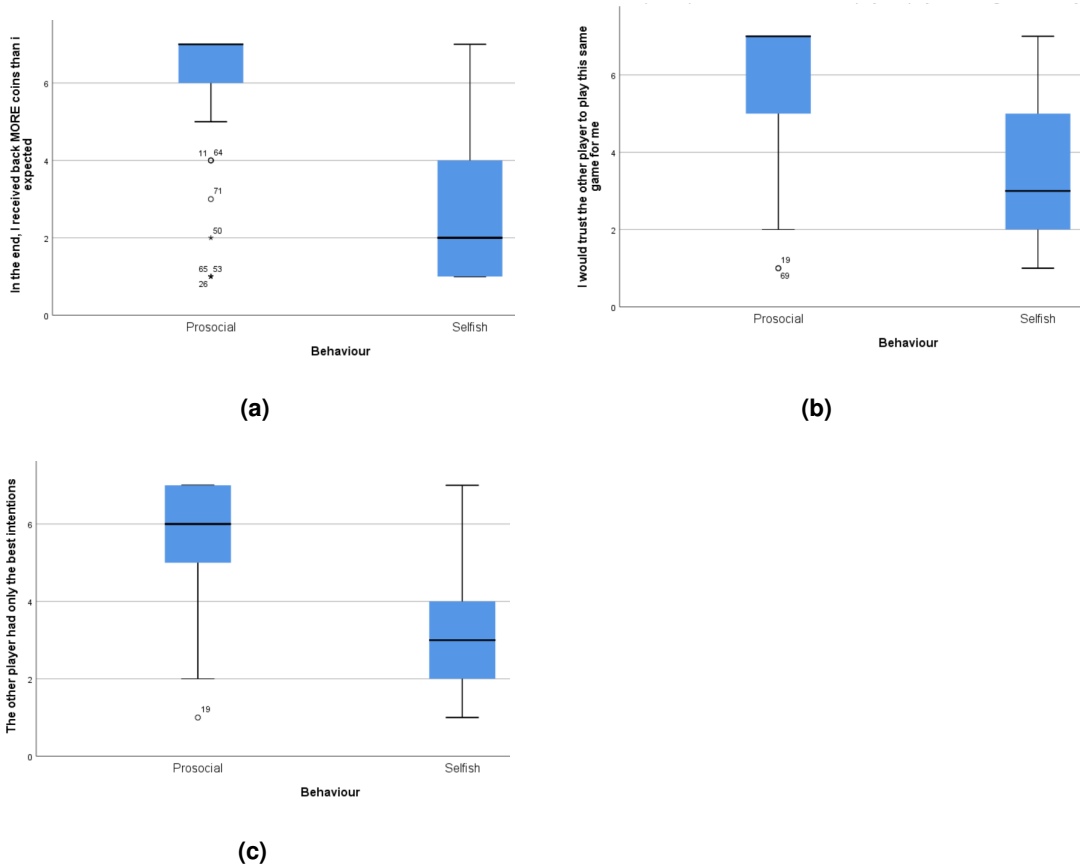


Figure 5.7: Values of the questions related to Trust

5.5.2.C Likeability

In the questionnaire we use the godspeed likeability questionnaire [26] that consists of 5 Likert Scales of opposite characteristics.

A Kruskal-Wallis test was executed on those Likert Scales. All the 5 pairs of characteristics showed no statistically significant influence(p-score$>$0.05) from the Agency. Dislike-Like(p-score=0.643), Unfriendly-Friendly(p-score=0.411), Unkind-Kind(p-score=0.269), Unpleasant-Pleasant(p-score=0.289) and Awful-Nice(p-score=0.41).

The results show that there was no significant difference in the participants in the likeability matter.

5.5.2.D Cooperation

The 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.

A Kruskal-Wallis test was executed to understand the influence of the Partner on the view of the participant has of the cooperation.

Question	p-Value
the other player collaborated with me.	0.595
The other player tried to help me throughout the game.	0.611
The other player reciprocated my actions.	0.479
I collaborated with the other player.	0.332
I tried to help the other player throughout the game.	0.203
I reciprocated the other player's actions.	0.670

Table 5.10: Kruskal-Wallis test results.

This results show that we must retain the null hypothesis for the influence of the Partner in the questions about cooperation.

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 5.11: Mann-Whitney U test results.

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

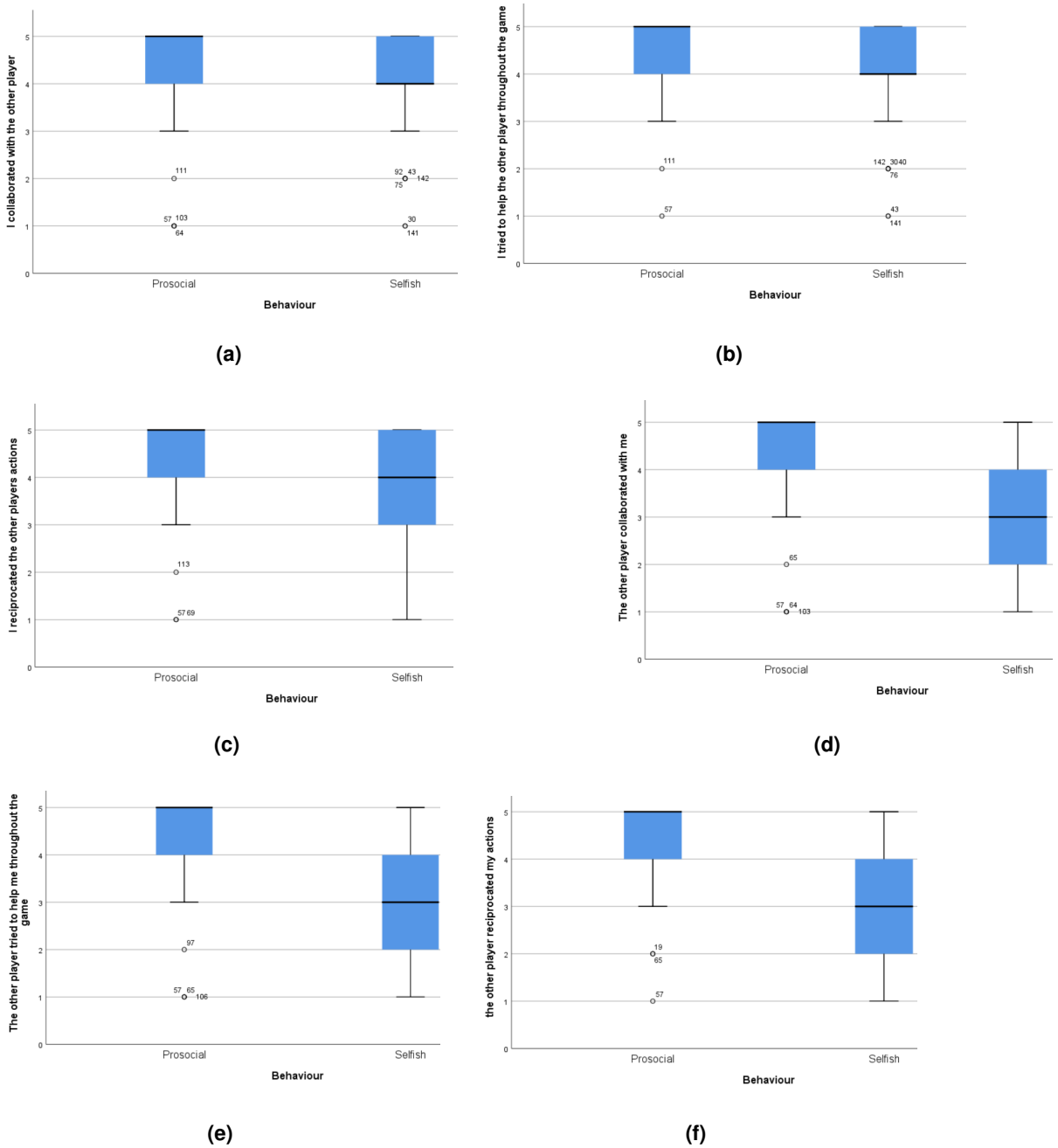
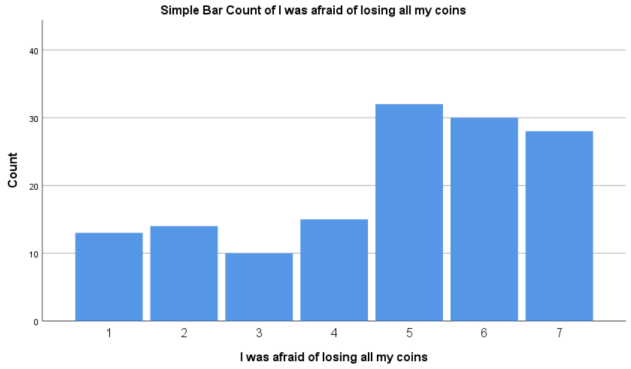


Figure 5.8: Values of the Collaboration Questions divided by Behaviour.

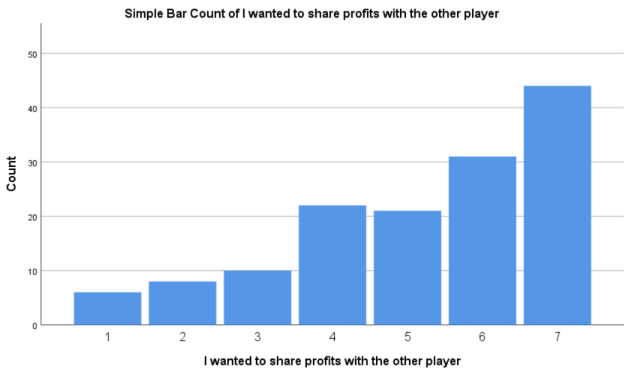
5.5.2.E Motivation

In the trust game there is the objective of achieving higher amounts of coins. In order to achieve this, it is necessary to trust and delegate coins to the partner. Based on this it is important to know if participants were motivated to win. Another thing that is important to take into consideration is the fear of losing.

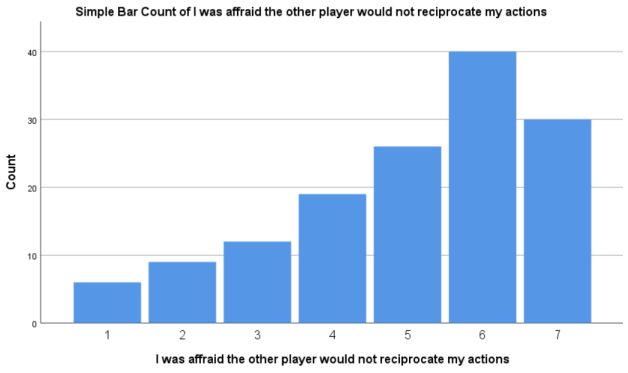
(a) Mean=6.11



(b) Mean=4.7



(c) Mean=5.2



(d) Mean=5.04

Figure 5.9: Values of the questions related to Motivation

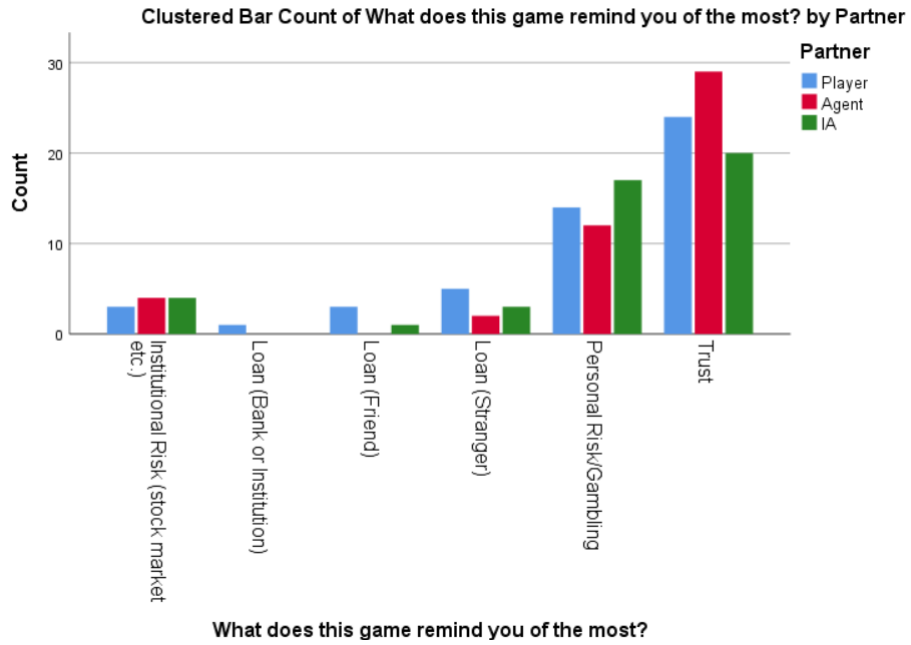
It is important to understand the motivations of the participants when playing the game. The graphs show that the participants were generally motivated to win but in some cases afraid to lose. This is a duality that influences how they play the game.

5.5.2.F Perception

We wanted to know if the participants observed the trust game as a risky game.

As shown in figure 5.10 most of the users viewed this game as a situation of trust or Personal Risk. This can influence the way the game is played and the motivation of the user. Personal Risk/Gambling

Figure 5.10: Perception of the Trust Game



can generate higher fear of loosing since the player sees the game as a gamble. Trust on the other hand allows the user to develop that feeling and risk more coins.

5.6 Discussion

5.6.1 Trust Game

5.6.1.A First Round

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 5.2 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.

5.6.1.B Second Round

When playing the game, only after the first round 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.

5.6.1.C Average of All Rounds

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.

5.6.1.D Combination of Behaviour and Partner

In the book "How Humans Judge Machines" [27] 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." [27]**
- **Principle 2 - "People assign extreme intentions to humans and narrow intentions to machines [27]"**
- **Effect - "People tend to judge humans more harshly in scenarios involving a lack of fairness. [27]"**

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.

5.6.2 Questionnaire

5.6.2.A Cooperation

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.

5.6.2.B Trust

As we can observe in table 5.9 the behaviour of the partner had an influence in the participants image of the coins he received, the intentions of the partner and the trust they had on him. The Prosocial behaviour was perceived as returning an amount of coins similar to what the participant expected, as being more trustworthy and having better intentions. This shows the behaviour changes the way the participant viewed the motivation of the partner and the trust on him. It is important for the person to understand the motivations of the person he trusts. That can change the reaction to certain actions.

5.6.2.C Perception of the Partner

Results in the Anthropomorphism section show that participants attribute human characteristics to the condition where they would supposedly be playing with another person. 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, because as stated before the same action from a user and from a machine causes different reactions. In a situation of unfairness a human is judged differently than a machine

is [27]. In the Likeability section we can see that neither of the Partners had characteristics that made a statistical difference to set it apart.

5.6.3 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 [22].)

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).

6

Conclusion

Contents

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 a 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.

6.1 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.

Bibliography

[1] J. Berg, J. Dickhaut, and K. McCabe, "Trust, reciprocity, and social history," *Games and economic behavior*, vol. 10, no. 1, pp. 122–142, 1995.

[2] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[3] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 33–42.

[4] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 1–8.

[5] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva, "Exploring the impact of fault justification in human-robot trust," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 507–513.

[6] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T. M. Powers, C. Dixon, and M. L. Tielman, "Taxonomy of trust-relevant failures and mitigation strategies," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 3–12.

[7] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 101–108.

[8] S. Bourgeois-Gironde and A. Corcos, "Discriminating strategic reciprocity and acquired trust in the repeated trust-game," *Economics Bulletin*, vol. 31, no. 1, pp. 177–188, 2011.

[9] M. J. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," *The knowledge engineering review*, vol. 10, no. 2, pp. 115–152, 1995.

[10] E. Bogert, A. Schecter, and R. T. Watson, “Humans rely more on algorithms than social influence as a task becomes more difficult,” *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

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

[12] 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*, vol. 23, no. 3, pp. 393–404, 1998.

[13] C. Castelfranchi and R. Falcone, “Principles of trust for mas: Cognitive anatomy, social importance, and quantification,” in *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*. IEEE, 1998, pp. 72–79.

[14] 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)*, vol. 8, no. 4, pp. 1–30, 2018.

[15] M. Coeckelbergh, “Can we trust robots?” *Ethics and information technology*, vol. 14, no. 1, pp. 53–60, 2012.

[16] R. Parasuraman and V. Riley, “Humans and automation: Use, misuse, disuse, abuse,” *Human factors*, vol. 39, no. 2, pp. 230–253, 1997.

[17] M. Chita-Tegmark, T. Law, N. Rabb, and M. Scheutz, “Can you trust your trust measure?” in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 92–100.

[18] P. Robinette, A. M. Howard, and A. R. Wagner, “Effect of robot performance on human–robot trust in time-critical situations,” *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 425–436, 2017.

[19] ——, “Timing is key for robot trust repair,” in *International Conference on Social Robotics*. Springer, 2015, pp. 574–583.

[20] 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*, 2019, pp. 1–9.

[21] C. M. de Melo, J. Gratch, and P. J. Carnevale, “The effect of agency on the impact of emotion expressions on people’s decision making,” in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 546–551.

[22] F. Cochard, P. N. Van, and M. Willinger, “Trusting behavior in a repeated investment game,” *Journal of Economic Behavior & Organization*, vol. 55, no. 1, pp. 31–44, 2004.

[23] U. Technologies, 2021. [Online]. Available: https://unity.com/

[24] “Qualtrics xm // the leading experience management software,” Apr 2021. [Online]. Available: https://www.qualtrics.com/uk/?rid=ip&prevsite=pt-br&newsite=uk&geo=PT&geomatch=uk

[25] K. Schaefer, “The perception and measurement of human-robot trust.(2013),” Ph.D. dissertation, Doctoral dissertation, University of Central Florida Orlando, Florida, 2013.

[26] 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*, vol. 1, no. 1, pp. 71–81, 2009.

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

A

Questionnaire

Figure A.1: Questionnaire page1

Informed Consent

Thank you for participating in this study. To conclude the task please answer the following questionnaire regarding the game you have just played.
Upon completion, please copy the unique code generated at the end of the questionnaire and introduce it in the mTurk task.

Gamecode

Insert the code given to you at the end of the Coin Game.

Coin Game Interaction

Please indicate your agreement or disagreement with the statements below regarding the **player's behavior during the game.**

	1-Completely disagree	2	3- Neither agree nor disagree	4	5- Completely agree
The other player tried to help me throughout the game	○	○	○	○	○
The other player started with 5 Coins	○	○	○	○	○
The other player reciprocated my actions	○	○	○	○	○
The other player collaborated with me	○	○	○	○	○

Please indicate your agreement or disagreement with the statements below regarding **your behavior during the game**.

	1-Completely disagree	2	3- Neither agree nor disagree	4	5-Completely agree
I started with 5 Coins	○	○	○	○	○
I collaborated with the other player	○	○	○	○	○
I tried to help the other player throughout the game	○	○	○	○	○
I reciprocated the other player's actions	○	○	○	○	○

Figure A.2: Questionnaire page2

Please indicate your agreement or disagreement with the statements below regarding your experience playing the game.

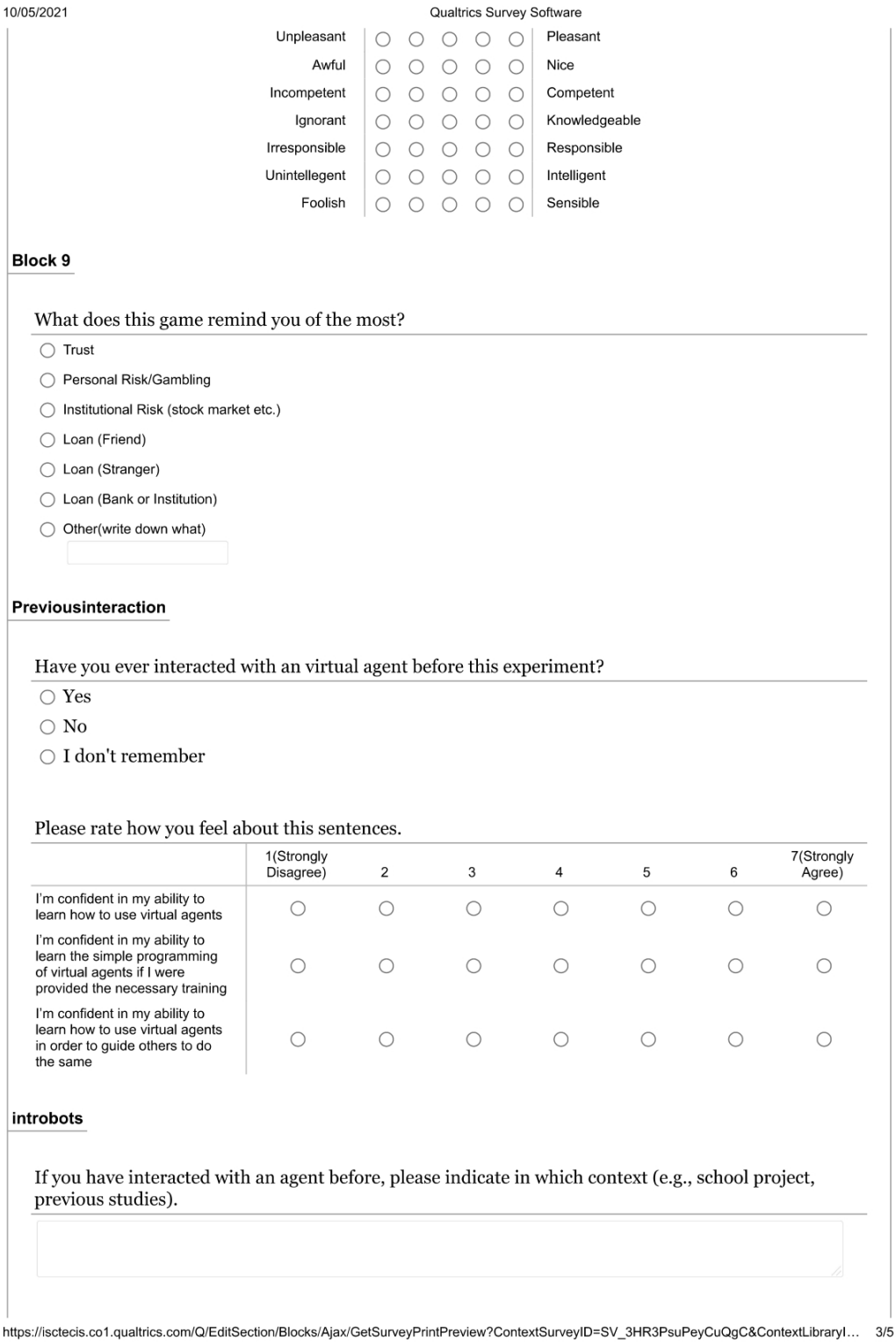
	1(Strongly disagree)	2	3	4	5	6	7(Strongly agree)
At the start of the game I was expecting to increase my initial amount of coins	○	○	○	○	○	○	○
In the end, I received back LESS coins than I expected	○	○	○	○	○	○	○
In the end, I received back MORE coins than I expected	○	○	○	○	○	○	○
The other player had only the best intentions	○	○	○	○	○	○	○
I would trust the other player to play this same game for me	○	○	○	○	○	○	○
Select 7 in the rating scale.	○	○	○	○	○	○	○
The other player's behavior was predictable	○	○	○	○	○	○	○
I was afraid of losing all my coins	○	○	○	○	○	○	○
I wanted to achieve the highest profit	○	○	○	○	○	○	○
I wanted to share profits with the other player	○	○	○	○	○	○	○
The other player wanted to share their profits with me	○	○	○	○	○	○	○
I was afraid the other player would not reciprocate my actions	○	○	○	○	○	○	○

Block 8

Please rate the impression of the other player on these scales:

Fake	○	○	○	○	○	Natural
Machinelike	○	○	○	○	○	Humanlike
Unconscious	○	○	○	○	○	Conscious
Artificial	○	○	○	○	○	Lifelike
Moving Rigidly	○	○	○	○	○	Moving elegantly
Dislike	○	○	○	○	○	Like
Unfriendly	○	○	○	○	○	Firendly
Unkind	○	○	○	○	○	Kind

Figure A.3: Questionnaire page3



10/05/2021 Qualtrics Survey Software

Unpleasant	○	○	○	○	○	Pleasant
Awful	○	○	○	○	○	Nice
Incompetent	○	○	○	○	○	Competent
Ignorant	○	○	○	○	○	Knowledgeable
Irresponsible	○	○	○	○	○	Responsible
Unintellegent	○	○	○	○	○	Intelligent
Foolish	○	○	○	○	○	Sensible

Block 9

What does this game remind you of the most?

- ○ Trust
- ○ Personal Risk/Gambling
- ○ Institutional Risk (stock market etc.)
- ○ Loan (Friend)
- ○ Loan (Stranger)
- ○ Loan (Bank or Institution)
- ○ Other(write down what)

Previousinteraction

Have you ever interacted with an virtual agent before this experiment?

- ○ Yes
- ○ No
- ○ I don't remember

Please rate how you feel about this sentences.

	1(Strongly Disagree)	2	3	4	5	6	7(Strongly Agree)
I'm confident in my ability to learn how to use virtual agents	○	○	○	○	○	○	○
I'm confident in my ability to learn the simple programming of virtual agents if I were provided the necessary training	○	○	○	○	○	○	○
I'm confident in my ability to learn how to use virtual agents in order to guide others to do the same	○	○	○	○	○	○	○

introbots

If you have interacted with an agent before, please indicate in which context (e.g., school project, previous studies).

https://isctecis.co1.qualtrics.com/Q/EditSection/Blocks/Ajax/GetSurveyPrintPreview?ContextSurveyID=SV_3HR3PsuPeyCuQgC&ContextLibraryI… 3/5

Figure A.4: Questionnaire page4

How frequently do you interact with agents?

- ○ I only interacted once
- ○ Very rarely
- ○ Somewhat frequently
- ○ Very frequently
- ○ Almost every day

Do you have a technology degree?

- ○ Yes
- ○ No

Demographics

How old are you?

What is you gender?

- ○ Male
- ○ Female
- ○ Other
- ○ Prefer not to say

In which country do you currently reside?

What is the highest degree of education you have completed?

- ○ High School
- ○ Bachelor's Degree
- ○ Master's Degree
- ○ Ph.D.
- ○ Prefer not to say

Participantscomm

If you have any comments or feedback about the game you played or about this study, please leave your comments in the text box below.

Block 9

Your code is:

Figure A.5: Questionnaire page5

${e://Field/RandomID}

Paste the code on Mechanical Turk.
Please Click the button bellow to finish your questionnaire.