TÉCNICO LISBOA

Trust in Human-Robot Ad Hoc Teamwork

Joana Rocha Raposo

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. José Alberto Rodrigues Pereira Sardinha
Prof. Ana Maria Severino de Almeida e Paiva

Examination Committee

Chairperson: Prof. Alberto Manuel Rodrigues da Silva
Supervisor: Prof. José Alberto Rodrigues Pereira Sardinha
Member of the Committee: Prof. Manuel Fernando Cabido Peres Lopes

November 2023

This work was created using LaTeX typesetting language
in the Overleaf environment (www.overleaf.com).

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

I dedicate this thesis to my beloved friend Márcia Câmara, who influenced my life so positively in her very short time here on Earth. I will always miss you, and I will never forget how much you supported me to achieve my dreams.

Acknowledgments

I would like to thank my mother and father for supporting me in this challenging journey. Developing this project without your encouragement, and care would not have been possible. To my siblings Nuno and Sofia, thank you for your understanding and your guidance through all these years.

I would also like to thank my partner Francisco, who not only helped me through the development of this Thesis but also was my emotional support through the hardest times. Thank you for always believing in me.

I would also like to acknowledge my dissertation supervisors Prof. Alberto Sardinha and Prof. Ana Paiva for their insight, support, and sharing of knowledge that has made this Thesis possible. Thank you for the opportunity to work with you, for being so comprehensive, and for helping me grow during my academic journey.

To all members of GAIPS who received me so well during this period, thank you. In particular, thank you Ana Carrasco and Miguel Faria for assisting me in the development of this project.

I would also like to thank all my friends and colleagues who helped me grow as a person and supported me along this journey. Particularly, to my IST friends Larissa, Diana, Beatriz, Ana, and Rodrigo, thank you for always being there and helping me overcome my fears and always celebrating my conquers. To my Azorean friends Raquel and Marina, thank you for even with an ocean between us always showing your support and never letting me quit.

Last, but not least, I would also like to thank my dog Blackie for being such a good boy. Thank you for your company during the hard times while writing this Thesis.

To each and every one of you who helped me achieve this special moment of my academic journey – Thank you.

Abstract

Ad hoc teamwork addresses the challenge of collaborating with unknown agents without prior coordination methods. When addressing the ad hoc teamwork problem between humans and agents, the current approaches do not account for human trust in the agents. In this work, we present the main research question we attempt to study: how does trust influence human behavior when cooperating with an unknown agent? We conducted a study to prime participants' trust regarding the agent at different levels and analyzed their strategic behavior with the agent. Our results suggest that the way people perceive a robot's competence is correlated with the trust they feel in the robot: the more competent they perceive the robot, the more trust they have in it. However, we could not obtain significant differences regarding the behavior of people who trust more the robot and people who trust it less. With these primary results, we suggest modifications to the experiment to achieve behavior clusters to different trust levels and envision the development of an ad hoc teamwork algorithm tailored for human-robot teams by using mechanisms of trust integration in the decision-making process of a robot.

Keywords

Ad hoc Teamwork; Trust; Human-Robot Interaction; Human-Robot Trust.

Resumo

Ad hoc teamwork aborda o desafio da colaboração com agentes desconhecidos sem métodos de coordenação prévia. No que toca ao problema de ad hoc teamwork entre humanos e agentes, as abordagens atuais não consideram a confiança humana nos agentes. Neste trabalho, apresentamos a principal questão de pesquisa que tentamos estudar: como a confiança influencia o comportamento humano ao cooperar com um agente desconhecido? Para tal, realizou-se um estudo para promover diferentes níveis de confiança dos participantes em relação ao agente e analisamos seu comportamento estratégico com o agente. Os resultados obtidos sugerem que o modo como as pessoas entendem a competência de um robot está relacionado com a confiança que sentem no mesmo: quanto mais competente acharem o robô, mais confiança têm nele. No entanto, não obtivemos diferenças significativas entre o comportamento das pessoas que confiam mais no robô e as pessoas que confiaram menos nele. Com estes resultados iniciais, sugerimos algumas modificações à experiência para alcançar grupos de comportamento relacionados com os diferentes tipos de confiança e perspetivamos o desenvolvimento de um algoritmo de ad hoc teamwork adaptado para equipas humano-robô, utilizando mecanismos de integração de confiança no processo de tomada de decisão de um robô.

Palavras Chave

Ad hoc Teamwork; Confiança; Interação Humano-Robot; Confiança Humano-Robot.

Contents

List of Figures

List of Tables

Acronyms

AHT	Ad Hoc Teamwork
AI	Artificial Intelligence
HP	High Performance
HRI	Human-Robot Interaction
HT	Human Teammate
LP	Low Performance
MDMT	Multi-Dimensional Measure of Trust
POMDP	Partially Observable Markov Decision Problem
P2T	Propensity to Trust
RoSAS	Robot Social Attributes Scale

1

Introduction

Contents

Robots are increasingly becoming part of human daily life. In many risk-involved fields, such as healthcare and search-and-rescue teams, robots are being used to help and collaborate with humans [8–11].

For example, in nursing and healthcare, several robots [8] are being used to aid healthcare professionals in performing tasks such as lifting, transporting, giving assistance, giving bed baths to patients, and even performing medical rounds in the bedrooms of patients at hospitals [9]. In the case of search-and-rescue teams, some robots have been already designed to be part of a team and explore cluttered environments, investigate the dangers or visibility of an area, and search for victims [10, 11]. To rely on robots and have an effective performance in such scenarios, a trust relationship must be established between humans and robots.

In many real-world scenarios, the robot will need to cooperate with unknown teammates without being able to pre-coordinate or even communicate with them. This brings us to the notion of *Ad Hoc Teamwork (AHT)*, first proposed in 2010 by Stone et al. [12]. Since then, many algorithms have been proposed to address this challenge and the applicability of the problem has been growing. For example, the ad hoc teamwork problem has evolved to new domains such as adding the issue of an unknown task to execute [13], the challenge of human-robot teams [14, 15] and dealing with imperfect information and without full observability of the environment [16].

Having human-robot teams requires humans to trust their robot teammates to achieve a successful collaboration [17]. However, there is a gap in the literature about trust in the context of ad hoc teamwork. In this work, we contribute to human behavior research regarding different levels of trust in a human-robot ad hoc teamwork scenario.

1.1 Research Questions

From the above discussion, this research focuses on the following research question: How does trust influence human behavior when cooperating with an unknown agent?

This research question is broken down into two research questions. Below, we formalize them.

Research Question 1: How does the behavior of people change when cooperating with an agent teammate within a scenario that may involve risk?

Research Question 2: How do these different behaviors from Research Question 1 relate to the different trust levels people have in the unknown agent teammate?

1.2 Contributions

In this research, we contribute to understanding of how trust impacts human behavior in a collaborative scenario with an agent. We adapted the environment from the Toxic Waste Domain [5], to create a

collaborative agent and a scenario that allows cooperation between an agent and human participants.

Additionally, we advance to study human behavior in a scenario where there are risks and consequences for each action taken by the human. In Section 2.2, we will discuss how having risk is essential to forming a trust relationship. Each human action has different levels of associated risk, which reflects objectively how much the human teammates trust the agent.

Following these lines, we contribute to encouraging the possibility of developing an AHT algorithm that is tailored to human trust. Following a similar concept as Chen et al. [18], it is possible to adjust the behavior of the AHT agent according to the level of trust of the human teammate.

1.3 Organization of the Document

This thesis is organized as follows: Chapter 2 presents the state of the art of the fields of Human-Robot Interaction (HRI) and AHT. In Chapter 3 we introduce the methods used to develop this work. Chapter 4 shows the obtained results and their respective interpretation and discussion. Chapter 5 summarizes the principal results and future recommendations.

2

Related Work

Contents

In this section, we present and discuss the research work in the field of ad hoc teamwork and human-robot trust, which are fundamental topics for the development of this project.

2.1 Ad Hoc Teamwork

Ad hoc teamwork is described as the problem of developing agents that are capable of collaborating with other unknown agents without prior coordination methods [19]. This research field has relevance since it can have many applications in different areas, such as healthcare systems, team sports, or service robots, where agents need to learn how to cooperate with unknown agents or humans without pre-coordination.

2.1.1 Characteristics and challenges of the ad hoc teamwork problem

The problem of AHT without pre-coordination is introduced by Stone et al. [12], which encouraged the community to develop "an autonomous agent that can efficiently and robustly collaborate with previously unknown teammates on tasks to which they are all individually capable of contributing as team members". These agents must be **flexible** and **adaptive** since they need to adjust their behavior to unknown teammates with whom they will cooperate.

In addition, the key assumptions that characterize an AHT agent rely on the following conditions [19]:

- No prior coordination;
- No control over teammates;
- Collaborative behavior.

When developing an AHT agent, Stone et al. [12] address three main technical challenges that the agent must be able to perform:

1. Identify all possible teamwork situations;
2. Search the optimal effective behavior algorithm for each situation;
3. Identify/classify the current teamwork situation the agent is in.

2.1.2 Main subtasks and variations of ad hoc teamwork

According to Mirsky et al. [19], there are four main subtasks that an AHT agent should be able to perform. These subtasks are highly related to the technical challenges mentioned previously in Section 2.1.1.

First of all, there is the need to define the **knowledge representation** of the domain, including several aspects related to the environment, the capabilities of the agent, and the potential teammates.

Moreover, the subtask of **modeling teammates** is particularly important to leverage details about the teammates of the AHT agent, to refine the process of decision-making. These two subtasks address the first technical challenge that Stone et al. [12] identified.

In this way, after the AHT agent has an estimate of the behavior of the other teammates, it needs to perform **action selection**, with the purpose of choosing the most adequate action for the obtained estimation. This corresponds to the second technical challenge: searching for an optimal behavior for the corresponding situation.

Finally, the AHT agent is in a dynamic situation, where it may receive new facts about the environment, its teammates, tasks, or even objectives. Therefore, the AHT agent must be able to **adapt to changes.** This last subtask is correlated with the third technical challenge: the need to identify what is the current teamwork situation, since the situation where the agent is may change over time.

In addition to these main subtasks, [19] also explores additional variations that also define the field of AHT. These variations influence how the previously stated subtasks must be executed.

Starting with **partial observability** instead of full observability, this variation implies a higher complexity level of knowledge representation. Since the agent will not be fully aware of the state of the environment, this introduction of uncertainty impacts how all the subtasks mentioned before are performed.

Another variation is addressed as **open environment**. This happens when the number of teammates is not fixed. This means that the AHT agent will need to adjust how it models its teammates and how it adapts to the environment changes.

Communication is presented as an additional variation. This variation is used to distinguish the cases where there is a communication channel between the agents. When facing such variation, the complexity of the subtasks increases, since the agents may need to learn the communication protocol during task execution.

Moreover, the **adaptive teammates** variation reports to teammates that learn by reacting to the policy of the AHT agent, instead of learning alongside the AHT agent. Therefore, the execution of action selection and the adaption to changes are affected once the methods used by the teammates are not known by the AHT agent.

Lastly, **mixed objectives** describe situations where, even though the AHT agent and its teammates have a common goal, it may be the case that each agent also has individual objectives. In this variation, it is important to remember that these individual objectives can not be purely contradictory to the common shared goal. This assumption will require some changes in the way the AHT models its teammates and how it performs its action selection.

A major insight retrieved from [19] is comparing the AHT setting with the human-agent interaction field. It is brought to attention the fact that, in the case of AHT, the agents need to find a way to coordinate

with previously unknown humans, by using **implicit communication** or **acting in a legible manner**.

In addition, [19] also provides the common solution methods used to solve each of the subtasks defined before.

First of all, knowledge representation can be solved with three distinct approaches: type-based methods, experience replay, and task recognition. The first approach relies on a set of hypothesized types that represent prior experience with agents in the domain at hand. Each hypothesized type models an action selection policy. In this case, an assumption is required: new teammates encountered by the AHT agent must have their behaviors specified by one of the hypothesized types. On the other hand, experience replay methods keep in a buffer transition data. This way, it is possible to compare the current transition with the stored transition and identify the current teammate. In the last approach, task recognition, prior experience or information is encoded in a task library that can be referred to as plays, macro actions, or options.

When it comes to the subtask of modeling the teammates, there are three common solution approaches: task inference, experience recognition, and task recognition. The first approach is in alignment with teammate representation using type-based methods, in which beliefs are inferred over the hypothesized type. When using task inference methods, the prior beliefs about the types of teammates are updated using the history of interactions and their type probability. Instead of inferring types, experience replay is a method where the similarity of the current observations with previous experience is measured more directly. Finally, when task recognition is used to represent previous knowledge, the AHT agent can also seek to infer the current task being executed by the teammate.

To perform action selection, planning methods are often used. In addition, expert policy methods can also be applied, where a policy is chosen and actions are selected according to the picked policy. When considering adaptive teammates, it is used the leading method, where the choice of the AHT agent influences the behaviors of its teammates. Lastly, a common solution relies on metalearning, in which action selection policies are trained to ease the AHT process.

To solve the adaption to changes issue, several approaches can be followed. Starting with belief revision, this method makes use of belief revision protocols to preserve their belief about the identity of the teammates over time. Hypothesis space revision is another solution used for adapting to teammates whose behavior might not be suitably represented in the hypothesized space. Furthermore, metalearning methods are also used, where the action selection policy learns its own adaptation procedures, thus avoiding the necessity to define particular adaptation schemes. Additionally, zero shot coordination techniques are used as an approach. In this setting, the AHT agent is not authorized to have a behavioral adaption during the ad hoc interactions. Hence, in these approaches, the focus is on training agents to robustly coordinate with teammates who were trained using the same algorithm. Finally, the last approach is communication. In this method, the AHT agent can adapt to changes by communicating with

its teammates, either through a query, transfer knowledge or preferences, or providing advice.

2.1.3 Ad hoc teamwork algorithms

In addition to unknown teammates, Melo and Sardinha [13] extended the AHT problem to scenarios where the task is also unknown. Therefore, the AHT agent needs to identify both the task and the strategy being used by the teammates.

In order to solve this problem scenario, they propose the following plan structure: first, the AHT agent must do the **task identification** step, to know which task is being performed by the other agents. Once it knows which is the target task, it must execute the **teammate identification** step, which consists of determining which strategy is being used by the teammates to solve the corresponding task. Finally, when these previous steps are completed, it is time for the AHT agent to take the **planning** step, which relies on acting according to its teammates. This structure is mentioned as the fundamental requirements around which the AHT problem should be formulated.

We now proceed to compare the state-of-the-art algorithms. The work of Ribeiro et al. [14] contributed to the development of the first AHT algorithm that is tailored for human-robot collaboration. Their algorithm, BOPA, enables a robot to cooperate on the fly, without any pre-coordination protocol, with human teammates. Previous AHT algorithms [13, 20, 21] do not take into account specific challenges of human-robot interaction. For example, in a real-world scenario, the robot may not receive environmental signals, nor observe the human teammate's actions. Extending the work of Melo and Sardinha [13] to include sequential tasks under uncertainty, BOPA assumes that there are no visible actions, reward signals are not available, and that the teammates may not always follow an optimal policy. However, the AHT agent, at each time step, is able to observe the current state (full observability), which reduces the applicability of BOPA in real-world scenarios. Despite this disadvantage, the AHT agent was tested with human teammates and does not assume that these teammates always follow an optimal policy. These characteristics are important for our work since we aim to work in similar conditions.

In order to take into account the perceptual limitations of the AHT agent, Ribeiro et al. [15] expanded the decision-making process of [14] to adapt to partial observability, naming their framework as HOTSPOT. As well, in this setting, the human also knows the task, but the robot does not. Therefore, the robot must infer the current task from a set of possible tasks. Additionally, the robot has the ability to communicate with the human teammate. This framework for AHT in human-robot teams was able to infer and complete the target task even in the presence of partial and imperfect information. Although these improvements can provide advantages to real-world human-robot ad hoc teamwork interactions, due to time constraints, it is more likely that we develop a virtual scenario where the AHT agent cooperates with the human teammate, without communicating. Additionally, HOTSPOT was not tested for its robustness to non-optimal teammates.

As well, the work of Ribeiro et al. [16] extended [14] to develop an AHT algorithm under partial observability, ATPO. Besides the assumptions already made in BOPA (teammate's actions not visible, reward signals not available, AHT agent does not know the task to perform *apriori* and the AHT agent has access to a library of possible tasks), ATPO also explicitly assumes that the AHT agent can not communicate with the teammates. Even though ATPO was not tested in a human-robot interaction scenario, it provided efficient results. Since ATPO accounts for partial observability, it broadens its use in real-world situations. Moreover, the main differences between ATPO [16] and HOTSPOT [15] rely on the communication assumption between the AHT agent and its teammate and the type of teammate: virtual agent instead of a human. Despite the fact that ATPO can be more applicable in real-world situations due to partial observability, this algorithm was not tailored for human-robot cooperation and, therefore, it does not take into account the fact that teammates may not always follow the optimal policy.

Additionally, with the aim of evaluating human behavior in ad hoc teamwork, research done by Suriadinata et al. [22] assessed, in terms of optimality and legibility, the participants' behavior under three distinct conditions of instructions given prior to the AHT interaction. It was concluded that people do not always act in an optimal or legible way, which can influence both the AHT agent and team performance. Once again, this work emphasizes the importance of, in real-world scenarios, it should not be assumed that the behavior of the teammates is always the optimal one. Thus, they encourage having different representations of human behaviors in ad hoc teamwork scenarios. This conclusion enhances the preference for the use of BOPA over ATPO, in which is considered that the human teammate may not always follow the optimal policy.

Along those lines, Hanina et al. [23] also underlined the importance of building a model that encloses different human behaviors. To address this challenge of cooperation with people, [23] explored if it was feasible to quantify human rationality. In addition, they introduced a basic algorithm in which it is possible to fit the parameters of a bounded rationality model to learn human behavior. Hence, this algorithm can provide a prediction about the optimality of human participants, and can further be implemented in the decision-making of an AHT agent to improve team performance. Our main goal is to develop an algorithm capable of adapting to different humans with different levels of trust. Additionally, we primarily intend to study if a trustworthy AHT algorithm improves team performance. For this reason, we decided not to include this approach in the design of our algorithm.

After analyzing these recent works, we suggest the extension of BOPA [14] since our long-term goal is integrating trust in human-robot AHT. Given the prior arguments, having an algorithm tailored for human teammates and robust to non-optimal teammates are the features that best fit the conditions we intend to have. In our proposed setting, there will only be one type of task that the AHT agent and the human must cooperate to complete, instead of different possible execution tasks as in BOPA. The challenge of our proposed algorithm will be to identify the current level of human trust in the agent. Therefore,

we can model the different levels of trust as being the different possible tasks to execute. Similarly, each trust level type will determine what behavior the agent should follow. This identification of trust through behavior is what we intend to research through this project. Following the same assumptions as BOPA, the AHT agent will not be able to see the human's actions. To determine which level of trust the AHT agent is facing, it must rely on the current state, which includes the human's position.

Table 2.1 presents the main features of each ad hoc teamwork algorithm mentioned before. Although a new AHT algorithm is the long-term goal of this research, it is not the focus of this work. However, our initial proposal is listed as Trustworthy AHT to allow a comparison with the other algorithms presented here. The first seven features are possible assumptions made in the algorithms: whether the AHT agent can see the teammate's actions or not, whether the reward signals are available for the AHT agent or not, how the tasks to execute are represented, whether the teammate follows always an optimal policy or not, if the ad hoc agent knows beforehand what task to execute as a team or if it has to find it first, whether if the AHT agent has access to a library of the possible tasks it that may execute, and lastly if the agent is in a setting with partial observability. The remainder last six features are the possible outcomes that were obtained in the algorithms. The first two regard whether the AHT is capable of identifying the correct task that needs to be solved and if it is capable of solving it in an efficient time. Scalable addresses if the approach used is able to adjust to different problem sizes. Furthermore, the robustness of an approach indicates if the AHT agent used is able to adapt to non-optimal teammates. An algorithm is tailored for human-robot teams if it was developed (and tested) for human-robot ad hoc teamwork. Lastly, the last feature indicates if an AHT algorithm is capable of predicting how optimal human decision-making is.

2.2 Trust in Human-Robot Interaction

Human trust in robots is a highly discussed topic, since without a proper trust calibration, human-robot interactions may have undesirable consequences, such as robot inefficiency or even dangerous outcomes [24].

2.2.1 Challenges associated with trust

First of all, Lee and See [25] state several problems associated with a lack of trust in automation. Without trust, people tend to misuse and even disuse automation. Nowadays, automation is a key element in work environments and people need to be able to rely on it with confidence.

The importance of trust is also enhanced by Schaefer [26], which states that the lack of a solidly established trust in robots leads to their disuse and, therefore, impedes opportunities for developing

Table 2.1: Comparison of the characteristics of the ad hoc teamwork algorithms

	BOPA [14]	**HOTSPOT** [15]	**BRM** [23][1]	**ATPO** [16]	**Trustworthy AHT**[2]
Visible Teammate's actions	✗	✗	-	✗	✗
Available reward signals	✗	✗	-	✗	✗
Task description	MMDP	MMDP	-	MMDP[3]	MMDP
Teammates may not always follow optimal policy	✓	✓	-	-	✓
Ad hoc agent knows the task *apriori*	✗	✗	-	✗	✗[4]
The ad hoc agent has access to a library of possible tasks	✓	✓	-	✓	✓[4]
Accounts perceptual limitations of the ad hoc agent[5]	✗	✓	-	✓	✗
Effectively identifies the correct target task	✓	✓[6]	-	✓	**?**
Solves the correct task efficiently	✓[7]	✓[6]	-	✓[6]	**?**
Scalable	✓	-	-	✓	-
Robust to non-optimal teammates	✓	-	✓	-[8]	-
Tailored for human-robot teams	✓	✓	✓	✗	✓
Predicts the optimality of human decision making	✗	✗	✓	-	✗

✓: The algorithm has the mentioned feature.
✗: The algorithm does not have the mentioned feature.
- : The algorithm does not address the mentioned feature.
? : Results not known yet.
MMDP : Multi-agent Markov Decision Process.
[1] Bounded Rationality Model.
[2] Algorithm proposed in this project.
[3] From the AHT agent perspective, each task defines a Partially Observable Markov Decision Problem (POMDP).
[4] In our case, the unknown "task" is the trust level that must be inferred. The collaborative task will always be the same.
[5] Perceptual limitations: partial observability
[6] Near-optimal time.
[7] Optimal and near-optimal times.
[8] But robust to a noisy environment.

trust between humans and robots. In contrast, when people have too much trust in a robot, they may place themselves at risk, since they assume that the robot knows more than they do [27].

Currently, there is no agreed-upon definition of trust [28, 29]. The work of Lee and See [25] explores that trust may have several definitions but ends up emphasizing trust as being described by **uncertainty** and **vulnerability**: "Trust can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Alternatively, Salem et al. [29] state that the promoting factors of trust are **reliability** and **predictability**, while Wagner et al. [27] and Groom and Nass [30] propose a trust definition based on **risk**, since without it we may not be truly facing a trust relationship where the entity that trusts have something to lose when trusting another.

Not having a consensual definition impacts how trust is measured and how well-evaluated it is. To attempt to solve this problem, Law and Scheutz [28] proposed a formalization for the definition of trust: "A *truster* is the person doing the trusting; a *trustee* is the person or system that is being trusted; **entrusting** with is to assign the responsibility of doing something (to someone)".

In this work, we will follow the definition provided by Lee and See [25] since we can associate vulnerability with risk, an essential factor in developing a trust relationship. In our proposed setting, people will face uncertainty when it comes to deciding whether or not to trust the robot and will face risks depending on their decision. A more detailed explanation of the setting will be provided in Chapter 3.

2.2.2 Factors that affect trust

There are several causes that impact trust. As stated by Lee and See [25], automation faults cause a decline in trust and lead to the disuse of automation. One trust-specific finding is the importance of the design of the robot for the perceived trust since it is related to the physical form and functional capability of the robot [26].

Several system properties can affect trust [31]. Beginning with system predictability, this feature is important because early in a human-robot relationship, trust is based on how predictable the robot is. Moreover, the system's intelligibility and transparency have a huge impact on the perceived trust, since systems that can describe their reasoning will be more easily understood by their users. Finally, the level of automation is a factor to take into consideration: higher levels of automation are more complex to understand how it performs, therefore are less transparent to users and may generate less trust.

Moreover, the capability of a robot to express and explain its decision-making process was also proven to impact trust. As identified by Wang et al. [32], human trust in a robot increases when the robot is capable of explaining its decision-making process.

Furthermore, risk highly influences trust. Bridgwater et al. [33] investigated how the approach to risk of a robot affects human trust in that designated robot. Their results demonstrated that a robot with a risk-seeking profile, meaning that it only takes into account the best-case outcome and therefore only chooses the riskier actions, is less trusted and, therefore, should be avoided by engineers.

In order to find how humans transfer or generalize trust in robot skills across tasks, Shu et. al [34] discovered that human trust generalization is affected by the following characteristics: task similarity, task difficulty, and robot performance. Similarly, Hancock et al. [35] also found that the most important characteristic of a robot that influences trust evolution is the robot's performance.

Not only the lack of trust bring consequences in human-robot interaction. Overtrust can lead to unwanted situations, where the human blindly trusts automation [27] and this leads to its misuse and reduction of the detection of automation failures [25]. Overtrust occurs when people take up too much risk, thinking that the trusted system will mitigate this risk [24]. As explained by Lee and See [25], researchers should aim to build **appropriate trust**, since both lack or excessive trust may lead to unwanted consequences.

To generate appropriate trust and reliance by the users, Lee and See [25] propose that the capabilities of the automation need to be well conveyed to the user, which can be done with simpler algorithms or with algorithms that reveal their operation in a more clear way. This helps to communicate the **purpose** of the automation.

However, trustworthy automation is described as "automation that performs efficiently and reliably" [25]. Occasionally, reaching such performance requires complex algorithms that are hard to understand. If the system performance depends on appropriate trust, then having a complex system may compro-

mise the trustability of automation. This raises an important research issue: the trade-off between trustworthy and trustable automation.

2.2.3 Trust measurement

Furthermore, there is a clear limitation in the field of trust in human-robot interaction: its accurate measurement. Trust is mostly measured through subjective assessment [26] and there is no standard way of measuring trust [4]. According to Hancock et al. [35], the majority of human-robot trust reviews only measure a momentary state of trust, rather than the process of trust growth.

Following this issue, the work of Schaefer [26] proposes a trust scale that was proven to assess the construct of trust. Since trust between humans is dynamic and changes through time [31], Schaefer [26] also introduces a new notion of pre-post interaction measurement of trust. This means that trust should be measured multiple times especially before and after the interaction between the human and robot, in order to evaluate how trust evolves. Pre-interaction trust measurement is used to recognize the initially perceived trust, influenced by human traits, robot features, perceived robot capabilities, and perception of the environment. With a post-interaction trust measurement, it is possible to identify the changes in trust related to human trust perceptions.

In the same way, Law and Scheutz [28] also point to the problem of the lack of objective means to measure trust. This work divides trust into two categories: performance-based trust and relation-based trust. **Performance-based trust** is described as "the robot being trusted to be reliable, capable, competent at its task or tasks, without needing to be monitored by a human supervisor" while **relation-based trust** is reported as meaning that "a person trusts the robot to be part of society in some way, not just off in a factory doing a job". At a performance-based level, it is relevant to mention the following conclusions: failures decrease trust and the greatest cause of trust growth was success obtained through teamwork.

They also divided objective measures into four categories:

1. Task intervention;
2. Task delegation;
3. Behavioral change;
4. Following advice.

The first two measures are more appropriate to assess performance-based trust while the last two are appropriate to assess both performance-based and relation-based trust or even a mix of the two types. The measure *following advice* is particularly relevant in the domain of human-robot teams, meaning that this measure could be useful for our study case.

The survey [28] highlights the importance of knowing and explicitly defining what type of trust we want to measure. Besides that, it is also important to choose correctly how to evaluate the type of trust chosen, to avoid measuring, by mistake, the incorrect type of trust.

Additionally, in a study performed by Chita-Tegmark et al. [4], they demonstrated that the robotic trust measure can be expanded and addressed with more nuance to capture the convictions of the participants in a better way. The study shows that, when given a N/A (not applicable) option in the rate items of several trust questionnaires, people did rate these items as N/A. The participants had available two types of N/A options: the "N/A to robots in general" option, to use when they felt that the statement did not apply to robots at all, and the "N/A to this particular robot" option, for when they felt that the statement did not apply to that particular robot at the specific scenario presented, but that might apply to other robots in other scenarios. They conclude that "N/A to robots in general" ratings designate a perceived category mistake. In Figure 2.1, we can see how they recommend the rate items to be formulated, where they also added a "not enough information" option, to avoid possible interpretation errors, since their exploratory analyses pointed out ambiguities in how trust measures are formulated.

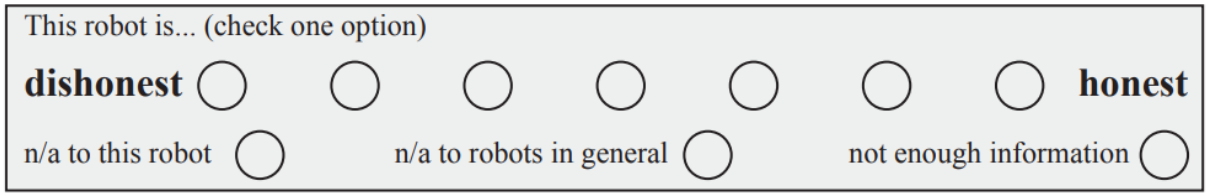


Figure 2.1: Recommended items formulation by Chita-Tegmark et al. [4] (2021)

In this project, we intend to measure performance-based trust relying on task intervention behaviors. To assess the overall trust, we intend to use and adapt some trust measure questionnaires that are often used and accepted as reliable [1–3]. Further information about the task intervention measure and the questionnaires will be discussed in Chapter 3.

2.2.4 Trust into robot decision making

Another essential point that motivated this project, is the work of Chen et al. [18], with the development of a computational model (trust-POMDP) that integrates trust into robot decision-making. With this model, the robot infers the trust of a human teammate through interaction, reasons about the results of its own actions on human reactions, and chooses the actions that maximize team performance over the long term. Some of the obtained results included:

- Team performance was improved;
- Maximizing trust in the robot by itself might not improve team performance;

- Human trust evolves based on the performance of the robot;
- Human trust affects human behaviors.

In this study, the robot had to pick up different objects - three water bottles, one fish can, and one wine glass - and, at each time, the human would either let it grab, or block it if the human did not trust the robot to grab such object.

They concluded that the robot was able to make good decisions: either pick up a low-risk object (such as a water bottle) to increase trust or to grab directly a high-risk object (such as the wine glass) when trust is high enough. With this, the robot is capable of inferring and influencing human trust, which is a feature that we should replicate in the end-goal trustworthy AHT algorithm. However, in the new suggested algorithm, the AHT agent will not reason about the results of its own actions on human reactions. Instead, the AHT agent will try to identify patterns in human behavior, infer the level of trust, and adequate its actions and behavior according to the level of trust of the human teammate.

2.2.5 Importance of trust in ad hoc teamwork

It is important to study trust in the context of human-robot teamwork, due to the required collaboration. In order to accomplish this collaboration, it is imperative that humans trust the robots to complete a certain task [17]. When people trust their robot teammate to execute the tasks that are more suitable to be performed by a robot than a human, the team performance increases [32, 36]. In the same way, Lee et al. [37] highlighted that trust may increase a robot's capacity to be accepted as a cooperative partner by humans.

Research done by Herse et al. [17] suggested that participants are more willing to comply with a collaborative agent when placed in a position of difficulty or uncertainty, which are situations that require humans to trust the robot. This led to the conclusion that the more people trust the robot, the more team performance increases. Similarly, Groom and Nass [30] enhanced the importance of a collaborative system to forge trusting relationships with users since its success highly depends on the user's trust.

Although not focused on ad hoc teamwork, Brezeal et al. [38] addressed the problem of how to design communication strategies that can contribute to efficient human-robot teamwork. They found that nonverbal communication has an essential role in coordinating the actions of the teammates and that people actually prefer to have these non-verbal cues from their robot teammates. Likewise, according to Habibian and Losey [39], humans prefer to collaborate with legible and fair teams of robots, instead of robots who are only optimized for efficiency.

Moreover, an agent must calibrate a human's trust in it, by dynamically modeling each human individually [36]. Then, the agent is able to adapt its behavior to the human teammate and achieve better performance.

In human-robot ad hoc teamwork scenarios, the goal of the AHT agent is to collaborate with unknown human teammates. These human teammates will feel different levels of trust when collaborating with the unknown robot. Since trust has an important role in collaboration, the AHT agent must infer the level of trust of its teammates. Then, it can adapt its behavior accordingly, to achieve the best team performance possible.

In conclusion, we address in Table 2.2 a summary of the obtained results of the research done in the area of trust and human-robot collaboration. After analyzing Table 2.2, we can see how trust and collaboration are highly intertwined: to achieve effective collaboration, trust must be established; on the other hand, human trust in the agent is highly influenced by the agent's characteristics, such as the agent performance.

Table 2.2: Results and comparison between studies in the field of Human-Robot Interaction

	Characteristics that enhance efficiency in collaboration	**Characteristics that influence human trust**
Lee and See (2004) [25]	-	Automation faults
Breazeal et al. (2005) [38]	Implicit / Explicit non-verbal cues	-
Groom and Nass (2007) [30]	Trust	-
Hancock et al. (2011) [35]	-	Robot performance
Lee et al. (2013) [37]	Trust	-
Schaefer (2016) [26]	-	Design of the robot
Chen et al. (2018) [18]	Integrating trust into robot decision-making	-
Lewis et al. (2018) [31]	-	Predictability, intelligibility, transparency and level of automation of the system.
Shu et al. (2018) [34]	-	Perceived task similarity, task difficulty and robot performance
Wang et al. (2018) [32]	Trust	Robot explaining its decision-making process
Pynadath et al. (2019) [36]	Trust	Dynamically model the heterogeneous human teammates as individuals
Bridgwater et al. (2020) [33]	-	Risk
Theresa and Scheutz (2020) [28]	-	Failure; Success through teamwork.
Herse et al. (2021) [17]	Trust	Uncertainty, difficulty
Habibian and Losey (2022) [39]	Teams of robots that act in a legible and fair way	-

- : The study did not explore this feature.

3

Methodology

Contents

In this chapter, we provide a description of the approach and methods used to study the research questions proposed in Section 1.1. In addition, we present the simulation environment in which the experiment was developed.

3.1 Simulation Environment

The Toxic Waste Domain [5] is a simulated environment where an agent and a human must cooperate to achieve a common task: gathering a set of toxic objects placed on some counters around the room and storing them in a container. The container is a part of the agent that can receive the toxic waste. However, the robot is not able to pick up toxic waste. Hence, the human must pick up and drop the toxic objects inside the container.

This experiment is essentially a game, where the goal of the human is to finish the game with the best score possible. The game finishes when all toxic objects in the room are placed in the container. The longer the human-agent team takes to clear all toxic objects from the room, the worse their score will be. In addition, the longer a human carries toxic waste objects, the more penalties the team will also receive. Since the agent may be slower than the human, the human must decide whether to wait for the robot to arrive at the location of the objects, or to carry the objects and take them to the robot.

In Figure 3.1 we can see an example of the game interaction scenario: the yellow agent is controlled by the human, the blue agent is the corresponding agent teammate that has the container, and the dark grey zones with stripes are the counters. Above the counters, there are some green toxic wastes that the human and the agent must clean. The light grey cells are regular floors, where the agent and the human can walk normally; the green cells represent interference zones, where the human can walk normally as on the regular floor but the agent may slip due to interference and is not able to walk correctly. At the bottom, we can see the total time that has passed since the game began and the total time the human has carried the toxic objects. Both these timers are used to calculate the final score of the game. The final formula to compute the final score is:

$$score = 100 - GT - \lfloor TWT \rceil, \tag{3.1}$$

where GT stands for Game Time, and TWT stands for Toxic Waste Time. GT denotes the total time the user takes to finish the game and TWT denotes the total time the human carries the toxic waste objects. To achieve the best score possible, the human participant must trust the robot. By trusting that the agent is capable of arriving at the location of the objects, variable TWT can be severely reduced. On the downside, there is always the risk of GT being overly increased. This risk allows the formation of a trusting relationship between the human and the agent.

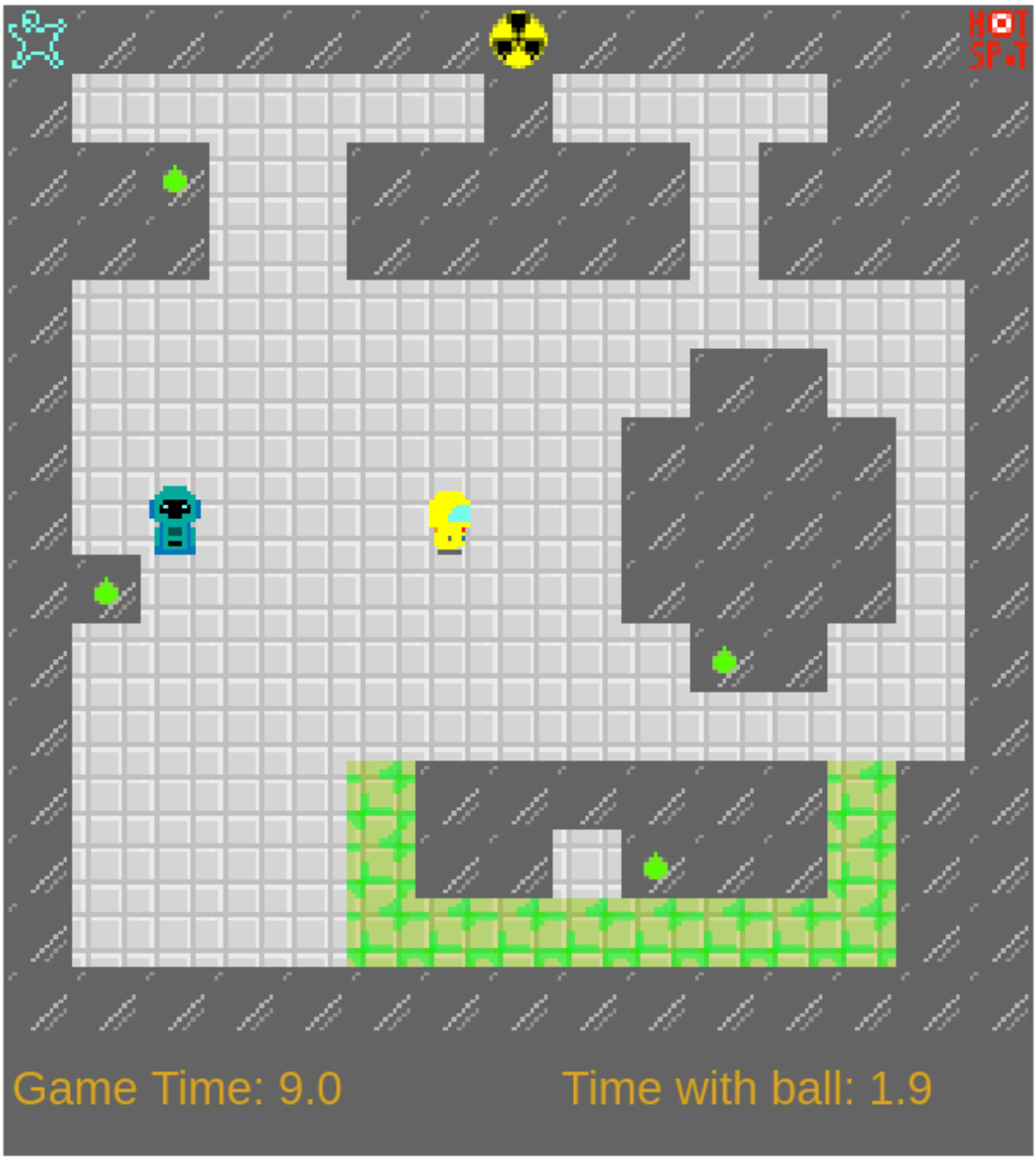


Figure 3.1: Adaptation of Toxic Waste simulation environment [5]

3.2 Solution Approach

First of all, we modified the Toxic Waste Domain [5] which was an offline game to be an online game, so we could conduct an experiment with several participants playing the Toxic Waste game. After the participants played the game, their trust in the agent teammate was assessed, with the adaption of the

questions and items from some trust questionnaires [1–3] (see Section 3.3 for further details).

Having this information, we hypothesized that would be possible to form clusters of behaviors that depend on the level of trust. The hypotheses that conducted this trial are the following:

- **Hypothesis 1:** When people have a **low trust** level in their agent teammate, they will carry the toxic waste objects for longer periods of time, they will not wait for the agent and will go toward the agent's position while handling the toxic object.
- **Hypothesis 2:** When people have a **high trust** level in their agent teammate, they will wait for the agent to arrive closer to the objects and only then they will pick up the toxic object and drop in the agent's container.

After the respective analysis of this study, we are able to infer what actions people take with low and high levels of trust. The results from this experiment will serve as a way of measuring trust for the suggested Trustworthy Ad Hoc Teamwork algorithm. This method of measuring trust is a task intervention measure considering that the human must decide whether to wait for the robot or intervene by carrying the toxic waste to the robot's container. More details on how this information can be used within the algorithm will be provided in Section 5.2.

3.3 Questionnaires

In this study, we intended to evaluate the participant's trust in the collaborative agent at the Toxic Waste Game, by providing, after the interaction, an adaption of the following questionnaires to the participants:

- **Self-assessment scales about Artificial Intelligence**, based on:
  - The Robot Social Attributes Scale (RoSAS) by Carpinella et al. [1]. With this standard scale, we can measure the social perception of the agent developed.
  - The Multi-Dimensional Measure of Trust (MDMT) scale by Ullman and Malle [2]. This questionnaire is relevant for measuring trust as a multidimensional concept.
- **Propensity to trust questionnaire**, based on the Propensity to Trust (P2T) scale by Frazier et al. [3]. This questionnaire is pertinent to use since an individual's propensity to trust is associated with their willingness to be vulnerable to another and, therefore, trust another entity [3].

RoSAS can be divided into three subscales: Competence, Warmth, and Discomfort [7]. We measured only the competence subscale, and the respective items measured can be seen in Table 3.1. The six attributes Capable, Responsive, Interactive, Reliable, Competent, and Knowledgeable are correlated with how competent people perceive a robot.

Table 3.1: RoSAS [1] Competence subscale

Competence Attributes
Capable
Responsive
Interactive
Reliable
Competent
Knowledgeable

Additionally, the second version of the MDMT scale [6] is divided into Performance and Moral scales. For the purpose of this research, we only assessed the performance scale. As well, the Performance scale can also be split into two subscales: Reliable and Competent. In Table 3.2 we can see the items that compose each subscale in detail. The mean of the reliable subscale items evaluates how much people feel they can rely on and count on the robot. As well, the mean of the competent subscale measures how much trust people feel in the robot to complete a certain task. The mean of the items of both subscales compose the Performance trust, which describes how much the robot is trusted to be reliable and competent to perform its tasks on its own.

Table 3.2: Performance Trust of MDMT scale (version 2) [6]

Performance Trust	
Reliable Subscale	Competent subscale
Reliable	Competent
Predictable	Skilled
Dependable	Capable
Consistent	Meticulous

Finally, we present in Table 3.3 the sentences that compose the P2T scale proposed by Frazier et al. [3]. Each of these sentences is rated by people regarding their general experience and is independent of the experiment. The mean score of these items provides the propensity to trust of a participant.

Table 3.3: Propensity to Trust Scale [3]

Propensity to Trust
It is easy for me to trust others.
Even if I am uncertain, I will generally give others the benefit of the doubt.
I generally believe that others can be counted on to do what they say they will do.
I usually trust people until they give me a reason not to trust them.
I tend to trust others even if I have little knowledge of them.
I generally give people the benefit of the doubt when I first meet them.
Trusting another person is not difficult for me.
My typical approach is to trust new acquaintances until they prove I should not trust them.
I am seldom wary of others.
I don't mind giving up control to others over matters which are essential to my future plans.
I believe that people usually keep their promises.
My tendency to trust others is high.

3.4 Pilot Study

In order to evaluate if the scenario was adequate for our research, we developed a pilot study with some volunteer participants from Instituto Superior Técnico to gather some information on how to improve our main experiment.

3.4.1 Participants

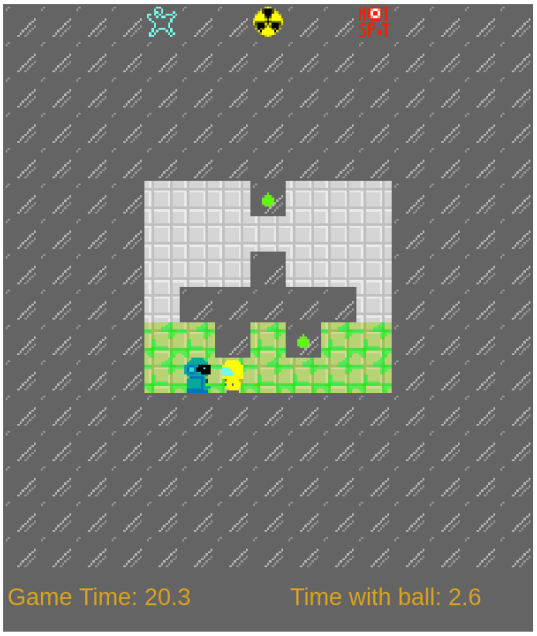
We recruited a total of 30 participants from Instituto Superior Técnico to play the Toxic Waste game. Of the 30 participants (7 female, 23 male), 86.7% were between 18-25 years old; 6.7% were between 26-35 years old; 3.3% were between 36-45 years old; and 3.3% were between 46-55 years old. Additionally, 23 participants had the highest education degree in high school or lower, 3 participants had a Bachelor's degree, 3 participants had a Master's degree, and 1 participant had a Doctoral Degree. Regarding studying or working with Artificial Intelligence (AI) subjects, 18 participants responded negatively, while 12 participants said to work or study with AI subjects.

3.4.2 Procedure

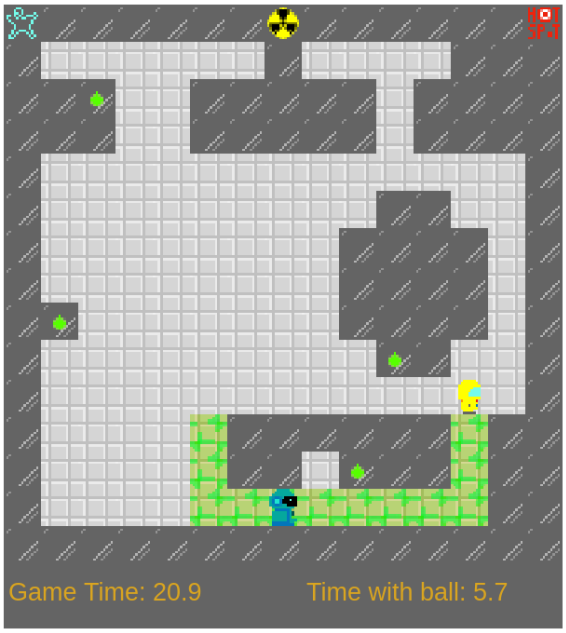
The study consisted of human players playing the Toxic Waste Game which has one tutorial part (see Figure 3.2(a)) and two game levels (see Figures 3.2(b) and 3.2(c)).

This experiment involved doing a priming of trust in two groups: the Low Performance (LP) group and the High Performance (HP) group. Of the 30 participants, 15 participated in the LP condition, and the other 15 were in the HP condition. The LP group interacted in the tutorial with an agent that had a higher probability of showing malfunction problems (due to the radioactive areas) than the HP group (LP = 65%, HP = 15%). The introductory speech in the LP condition was also different from the HP condition. As a way to induce low trust in the LP group, the participants were told some sentences like "This robot has a simple programming. . . " while in the HP group, the agent would be enhanced as "This robot is equipped with an advanced AI algorithm, . . . ". However, the game levels were equal in both conditions.

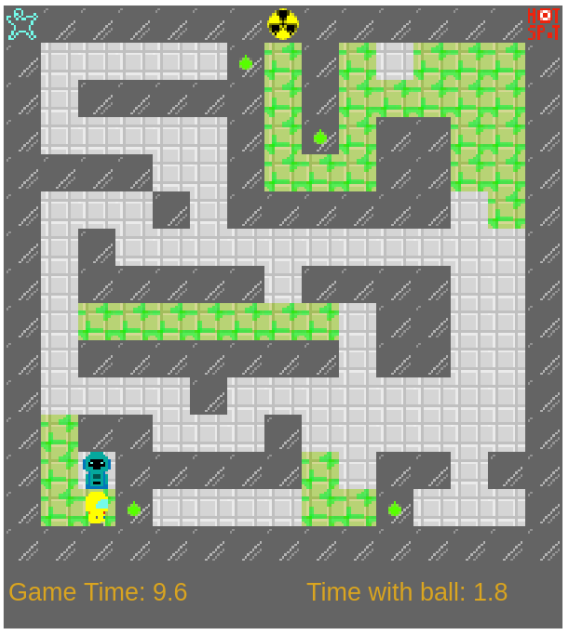
At the end of the game, participants answered three questionnaires: RoSAS [1], MDMT [2], and P2T scale by Frazier et al. [3].



(a) Tutorial level



(b) Level 1



(c) Level 2

Figure 3.2: All levels from the Toxic Waste Game

3.5 Main Study

For this research, we developed a study with several participants from Prolific [40]. This platform enables the worldwide collection of participants' data. In order to accomplish data quality, we filtered the participants to have a fluent English level. Each participant received a payment of 3 GBP. Participants took a median of 14.72 minutes to complete the study, for a median hourly rate of 12.23 GBP.

3.5.1 Participants

We recruited a total of 121 participants from Prolific [40]. Of the 121 participants (28 female, 92 male, 1 prefered not to say), 44.45% were between 18-25 years old; 33.88% were between 26-35 years old; 11.57% were between 36-45 years old; 4.96% were between 46-55 years old; and 3.31% were above 55 years old. Additionally, 37 participants had the highest education degree in high school or lower, 60 participants had a Bachelor's degree, 22 participants had a Master's degree, and 2 participants had a Doctoral Degree. Regarding studying or working with AI subjects, 82 participants responded negatively, while 43 participants said to work or study with AI subjects.

3.5.2 Procedure

In this study, we separated the participants along the following three conditions:

1. High-performance trust with the belief of the robot being autonomous group - HP group;
2. Low-performance trust with the belief of the robot being autonomous group - LP group;
3. No priming of trust with the belief of the robot being controlled by a human partner group - Human Teammate (HT) group.

The first two conditions were equal to the pilot study (see Section 3.4.2). The major difference in the Main Study is the creation of a new condition, the HT group. In this condition, the intention is to pretend the participants will play with an agent controlled by a human agent, instead of an autonomous one. This was done by modifying the game instructions in the initial speech. At the tutorial level, the probability of the robot showing malfunction problems was 15%. Before levels 1 and 2, a message of "Please wait for the other player to join" was shown for a few seconds (a random number between 5 to 10 seconds), to make it more credible. Despite that, levels 1 and 2 were equal in all conditions. From the 121 participants, 41 participated in the HP condition, 40 in the LP condition, and 40 in the HT condition.

As in the pilot study, the participants were distributed by the different conditions and played the tutorial and the two levels of the Toxic Waste Game (see Figure 3.2). At the tutorial level, the participants could re-play the tutorial as many times as they wanted, to ensure they could understand the game mechanics.

After each level, the participants had to fill in a formulary the attention check code to ensure they had completed the level successfully.

Once finished the three levels, the participants were asked to answer the following questionnaires: RoSAS [1], MDMT [2] and P2T [3]. After the study was completed, the participants received the accorded payment.

4

Results

Contents

In this chapter, we present how both the studies mentioned in Chapter 3 were analyzed and their respective results. We recall the hypotheses that were formulated in Section 3.2:

Hypothesis 1: When people have a **low trust** level in their agent teammate, they will carry the toxic waste objects for **longer periods**.

Hypothesis 2: When people have a **high trust** level in their agent teammate, they will **avoid** carrying the toxic waste objects, leading to a **lower period** of holding the objects.

We also present the explanation of the results obtained in the pilot study and how they guided the improvements made to the main study. Additionally, we discuss the results captured from the main study, highlight their correlation with the state-of-the-art, and finally suggest possible applications and recommendations of the results.

4.1 Pilot Study Results

In this study, we used both subjective and objective measures. Regarding the subjective measures, they were:

- The **social attributes** of the robot, using RoSAS [1],
- **Trust** in the robot, using MDMT [2],
- **Propensity for trust** of each participant using P2T [3].

On the other hand, we analyzed trust objectively by examining the number of timesteps the participants took holding the ball for each game level. Additionally, we investigated how many timesteps the participants took to finish the two levels of the game.

4.1.1 RoSAS [1] questionnaire

We compared each attribute of the Competence Attributes Scale [7] from RoSAS [1] in both conditions. The results indicated that there was no significant difference between all attributes of the LP group and the HP group.

In Tables 4.1 and 4.2 we show the details of the results of the analysis performed for each attribute and for each condition group. The attributes were scored on a scale from 1 to 9. For instance, when looking at Table 4.1 the LP group had a mean score of 7.07 for the attribute Capable with a standard error (SE) of 2.02, while the HP group had a mean score of 7.27 for the attribute Capable with a standard error of 0.8. In the attribute Capable, the Mann–Whitney U value is 105.00, and the corresponding p-value is .775. In Table 4.2 t is the two-sample t-test value, df denominates the degrees of freedom, and p is the p-value.

Table 4.1: Mann-Whitney U Test results of some of the attributes of the RoSAS competence subscale [7]

Attribute	Mean LP	SE LP	Mean HP	SE HP	U	p
Capable	7.07	2.02	7.27	0.80	105.00	.775
Responsive	6.60	1.80	7.27	1.79	82.00	.217
Interactive	6.47	2.20	6.00	2.42	97.50	.539
Competent	6.67	1.72	7.07	1.49	98.50	.567
Knowledgeable	6.53	2.42	7.07	1.58	107.50	.838

Table 4.2: Two-sample t-test results of the attribute of Reliable of the RoSAS competence subscale [7]

Attribute	Mean LP	SE LP	Mean HP	SE HP	t	df	p
Reliable	6.00	2.07	6.93	1.87	-1.36	28	.185

4.1.2 MDMT [2] questionnaire

We compared the performance-based trust of both conditions by comparing the average result of each item that composes the performance trust scale [6]. The results indicated that there was no significant difference between the performance-based trust of the LP group and the HP group.

Furthermore, we compared the reliable and competent trust subscales between the two conditions, comparing the average result of the items composing each of these subscales (see Table 3.2). The results showed that there was no significant difference between the reliable trust of the two conditions. Additionally, there was also no significant difference between the competence trust of the LP group and the HP group.

Moreover, we compared each item of the questionnaire in both conditions. The results indicated group LP had a significantly lower score on item **predictable** than group HP. For the remainder items, there was no significant difference between the scores of the two groups.

In Table 4.3 we show the details of the results of the analysis performed for each scale and item, for each condition group. Each item was scored on a scale from 0 to 7, or with the option “Does not fit” whenever a participant believed that characteristic could not be applied to the agent. For example, the LP group had a mean score of 4.72 for the Performance Trust Scale with a standard error (SE) of 1.85, while the HP group had a mean score of 5.29 for the Performance Trust Scale with a standard error of 1.15. In the Performance Trust Scale, the Mann–Whitney U value is 127.00, and the corresponding p-value is .567.

Table 4.3: Mann-Whitney U Test results of the Performance Trust Scale, the Reliable and Competent subscales, and their individual items [2].

Scale / Item	Mean LP	SE LP	Mean HP	SE HP	U	p
Performance Trust Scale	4.72	1.85	5.29	1.15	127.00	.567
Reliable Subscale	4.45	1.72	5.33	1.17	134.50	.201
Competent Subscale	5.02	2.07	5.24	1.30	98.50	.780
Reliable Item	4.80	2.08	5.29	1.49	116.50	.621
Competent Item	5.00	2.45	5.27	1.62	109.00	.902
Predictable Item	3.93	2.09	6.13	1.06	194.50	$<$ .001
Skilled Item	4.46	2.37	5.40	1.59	118.50	.339
Dependable Item	4.57	2.24	4.53	2.00	101.50	.880
Capable Item	5.53	2.03	5.40	1.68	101.50	.653
Consistent Item	4.15	2.54	5.43	1.34	115.00	.259
Meticulous Item	4.55	2.07	4.92	1.44	75.50	.820

4.1.3 Timesteps and Game Transitions analysis

We compared for both levels of the game the average number of timesteps that participants held the ball of toxic waste in their hands and the number of total timesteps they took to finish the game. Additionally, we also recorded the total number of different states where participants held the ball.

4.1.3.A Level 1 Results

The results indicated that there was no significant difference between the average timesteps that the participants held the toxic waste of the LP group and the HP group at level 1 of the game. The same conclusion was obtained regarding the total timesteps participants took to finish the game.

In Tables 4.4 and 4.5 we present the details of the results of the analysis performed, for each condition. For instance, when looking at Table 4.4 the LP group had a mean number of timesteps holding the toxic waste of 104.4 with a standard error (SE) of 100.22, while the HP group had a mean number of 99.07 with a standard error of 18.51. Regarding the variable Timesteps holding toxic waste, the Mann–Whitney U value is 72.50, and the corresponding p-value is .098. In Table 4.5 t is the two-sample t-test value, df denominates the degrees of freedom, and p is the p-value. In addition to these results, we obtained a total of 52 different states where participants from the LP group held the ball, while in the HP group, the total of states was 30.

Table 4.4: Mann-Whitney U Test results of the total timesteps in which participants held the toxic waste object at level 1 of the game

Variable	**Mean LP**	SE **LP**	**Mean HP**	SE **HP**	U	p
Timesteps holding toxic waste	104.4	100.22	99.07	18.51	72.50	.098

Table 4.5: Two-sample t-test results of the total timesteps participants took to finish level 1 of the game

Variable	Mean LP	SE LP	Mean HP	SE HP	t	df	p
Total timesteps to finish game	465.33	133.43	546.93	78.56	-2.04	28	.051

4.1.3.B Level 2 Results

The results indicated that there was no significant difference between the average timesteps that the participants held the toxic waste of the LP group and the HP group at level 2 of the game. The same conclusion was obtained regarding the total timesteps participants took to finish the game.

In Tables 4.6 and 4.7 we present the details of the results of the analysis performed, for each condition. For example, when looking at Table 4.6 the LP group had a mean number of timesteps holding the toxic waste of 147.67 with a standard error (SE) of 52.70, while the HP group had a mean number of 137.00 with a standard error of 40.38. Regarding the variable Timesteps holding toxic waste, the Mann–Whitney U value is 110.00, and the corresponding p-value is .064. In Table 4.7 t is the two-sample t-test value, df denominates the degrees of freedom, and p is the p-value. In addition to these results, we obtained a total of 55 different states where participants from the LP group held the ball, while in the HP group, the total of states was 35.

Table 4.6: Mann-Whitney U Test results of the total timesteps in which participants held the toxic waste object at level 2 of the game

Variable	Mean LP	SE LP	Mean HP	SE HP	U	p
Timesteps holding toxic waste	147.67	52.70	137.00	40.38	110.00	.064

Table 4.7: Two-sample t-test results of the total timesteps participants took to finish level 2 of the game

Variable	Mean LP	SE LP	Mean HP	SE HP	t	df	p
Total timesteps to finish game	694.27	105.99	834.87	255.17	-1.97	18.69	.064

4.1.4 P2T [3] questionnaire

We compared the propensity to trust of the participants of the two conditions, by comparing the average result of each item that composed the P2T scale [3]. The results indicated that there was no significant difference between the propensity to trust of the LP group and the HP group.

In Table 4.8 we show the details of the results of the analysis performed, for each condition. Each sentence on the scale was scored on a scale from 1 to 5. The LP group had a mean P2T score of 3.53 with a standard error (SE) of .615, while the HP group had a mean P2T score of 3.36 with a standard error of .564. The two-sample t-test value is .830, the degrees of freedom (df) are 28, and the corresponding p-value is .414.

Table 4.8: Two-sample t-test results of the P2T scale [3]

Scale	Mean LP	SE LP	Mean HP	SE HP	t	df	p
P2T	3.53	.615	3.36	.564	.830	28	.414

Since there was no representative sample of participants with a low propensity for trust, we did not compare participants with high vs. low propensity to trust in the pilot study.

4.1.5 Pilot Study Discussion

First of all, the results showed that the participants found that the RoSAS competence attributes [1] were correlated with the agent, more or less at the same amount, in both conditions LP and HP. Moreover, in terms of performance trust, people did trust the robot in a similar amount in both groups. Only when considering the predictable item, we can affirm that the HP group perceived the agent as more predictable than the LP group, as we intended to be. Lastly, in both conditions, the participants had a high propensity to trust, in a similar range value. These primary results mean we failed, at least subjectively, to obtain trust differences between the groups.

Looking at the objective results, that consider the number of timesteps the participants held the ball, although their difference is not significant, an interesting pattern that we found is that, on average, the LP group held the ball for more time. Additionally, on average, the LP took less time to finish the game than the HP group. Also, the LP participants would walk with the toxic ball along more states than the HP group. These results suggest that the HP participants had a tendency to wait more for the robot and avoided walking with the ball.

Along these lines, although the pilot study was done with few participants, we understood that the experiment needed some improvements before testing with a higher number of participants. During the experiment, we noticed that the participants were trying to finish the game very fast to score as high as possible. This led to a very quick interaction with the robot, especially at the tutorial level, designed to prime the level of trust in participants. To fix this, we decided to add an option of replaying the tutorial level as many times as the participant wished, so they could feel comfortable with the robot before proceeding.

As well, we observed that the behavior of the agent was not always the optimal one. So we decided to change the policy of the agent. One significant aspect of training this new policy of the agent is that we extracted a new policy of human behavior using the actual behavior that participants used in the pilot study. Despite the effort to introduce this change, this idea was not viable due to the fact that people in the study tended to rely on the agent to decide which action to take, and using that policy to train an agent that relies on the human to learn how to act introduces a paradox that leads to no learning from the agent side.

Finally, we noticed that we should make an effort to enhance the difference of timesteps that participants hold the ball between groups. Once the participant's beliefs about AI versus human performance on a given task influence whether they follow or not AI clues [41], we decided to add a new condition group to our study. For that matter, we decided that we would also create a human-player group, versus robot-player group conditions. This means, we would tell the human-player group that the robot would be controlled by a human, when in fact, it was the same robot as in the other condition. This led us to our final study with three conditions:

1. High-performance trust with the belief of the robot being controlled by AI group;
2. Low-performance trust with the belief of the robot being controlled by AI group;
3. No priming of trust with the belief of the robot being controlled by a human partner group.

4.2 Main Study Results

In the same way as the pilot study, we analyzed the following subjective measures: social attributes of the robot, participant's trust in the robot, and propensity to trust of each participant. We used, respectively, RoSAS [1], MDMT [2] and P2T [3] questionnaires.

Regarding objective measures, we took into consideration for each game level the number of total timesteps the participants took holding the ball, the total of time steps they took to finish the game, and also the number of transitions between different positions of the environment where they held the ball.

4.2.1 RoSAS [1] questionnaire

We compared each attribute of the Competence Attributes Scale [7] from RoSAS [1] in the three conditions. The results indicated that, for all attributes, the HP group had a significantly higher score than the HT group. Additionally, for all items except Reliable, the HP group also had a significantly higher score than the LP group. There was no significant difference between the LP group and the HT group.

In Table 4.9 we show the details of the results of the analysis performed for each attribute and for each condition group. The attributes were scored on a scale from 1 to 9. For instance, the LP group had a mean score of 5.80 for the attribute Capable with a standard error (SE) of 2.37, while the HP group had a mean score of 6.98 for the attribute Capable with a standard error of 1.60, and the HT group had a mean score of 5.78 with a standard error of 1.89. In the attribute Capable, the Kruskal-Wallis H value is 13.31, the degrees of freedom (df) are 2, and the corresponding p-value is .001. In addition, we show in Table 4.10 the pairwise comparison results that allowed us to conclude the previous results. For example, for the Capable attribute, the p-value between the groups LP and HT is .726, the p-value between the groups LP and HP is $<$.001 and the p-value between the groups HT and HP is .003.

Table 4.9: Kruskal-Wallis test results of the attributes of the RoSAS competence subscale [7]

Attribute	Mean LP	SE LP	Mean HP	SE HP	Mean HT	SE HT	H	df	p
Capable	5.80	2.37	6.98	1.60	5.78	1.89	13.31	2	.001
Responsive	6.08	1.73	6.85	2.09	6.18	1.60	8.75	2	.013
Interactive	5.70	1.70	7.29	1.37	5.98	1.61	21.28	2	$<$ .001
Reliable	5.88	1.74	6.59	2.01	5.69	1.79	6.24	2	.044
Competent	5.68	1.66	6.73	1.91	5.70	1.74	11.13	2	.004
Knowledgeable	5.40	2.07	6.61	2.19	5.38	1.74	11.88	2	.003

Table 4.10: Pairwise Comparison results of the attributes of the RoSAS competence subscale [7]

Attribute / Pairs	p LP-HT	p LP-HP	p HT-HP
Capable	.726	$<$ .001	.003
Responsive	.986	.011	.010
Interactive	.427	$<$ .001	$<$ .001
Reliable	.703	.052	.020
Competent	.800	.003	.006
Knowledgeable	.771	.005	.002

4.2.2 MDMT [2] questionnaire

We compared the performance-based trust of both conditions by comparing the average result of each item that composed the performance trust scale [6]. The results indicated that the HP group had a significantly higher trust score than the LP group. There is no significant difference between HP-HT groups and the LP-HT groups.

We also compared the reliable and competent trust subscales between the three conditions, comparing the average result of the item composing each of these subscales (see Table 3.2). The results showed that there was no significant difference between the reliable trust of the three conditions. Regarding competent trust, the HP group had a significantly higher score than the LP group. There is no significant difference between HP-HT groups and the LP-HT groups.

In Table 4.11 we show the details of the results of the analysis performed for each scale and subscales, for each condition group. Each item was scored on a scale from 0 to 7, or with the option "Does not fit" whenever a participant believed that characteristic could not be applied to the agent. For instance, the LP group had a mean score of 4.12 for the Performance Scale with a standard error (SE) of 1.37, while the HP group had a mean score of 4.89 for the Performance Scale with a standard error of 1.41, and the HT group had a mean score of 4.51 with a standard error of 1.14. In the Performance Trust Scale, the Kruskal-Wallis H value is 7.47, the degrees of freedom (df) are 2, and the corresponding p-value is .024. Additionally, we present in Table 4.12 the pairwise comparison results that allowed us to conclude the previous results. For example, for the Performance Scale, the p-value between the groups LP and HT is .255, the p-value between the groups LP and HP is .006 and the p-value between the groups HT and HP is .119. Lastly, in Figure 4.1 we can see a graphic representation of the mean score

of Performance Trust between the three conditions.

Table 4.11: Kruskal-Wallis test results of the Performance Trust scale, Reliable and Competent subscales [2]

Scale	Mean LP	SE LP	Mean HP	SE HP	Mean HT	SE HT	H	df	p
Performance Scale	4.12	1.37	4.89	1.41	4.51	1.14	7.47	2	.024
Reliable Subscale	4.28	1.34	4.88	1.45	4.53	1.18	5.35	2	.069
Competent Subscale	3.98	1.48	4.91	1.44	4.49	1.22	9.26	2	.010

Table 4.12: Pairwise Comparison results of the Performance and Competent Trust Scales [6]

Scale / Pairs	p LP-HT	p LP-HP	p HT-HP
Performance Scale	.255	.006	.119
Competent Subscale	.144	.002	.125

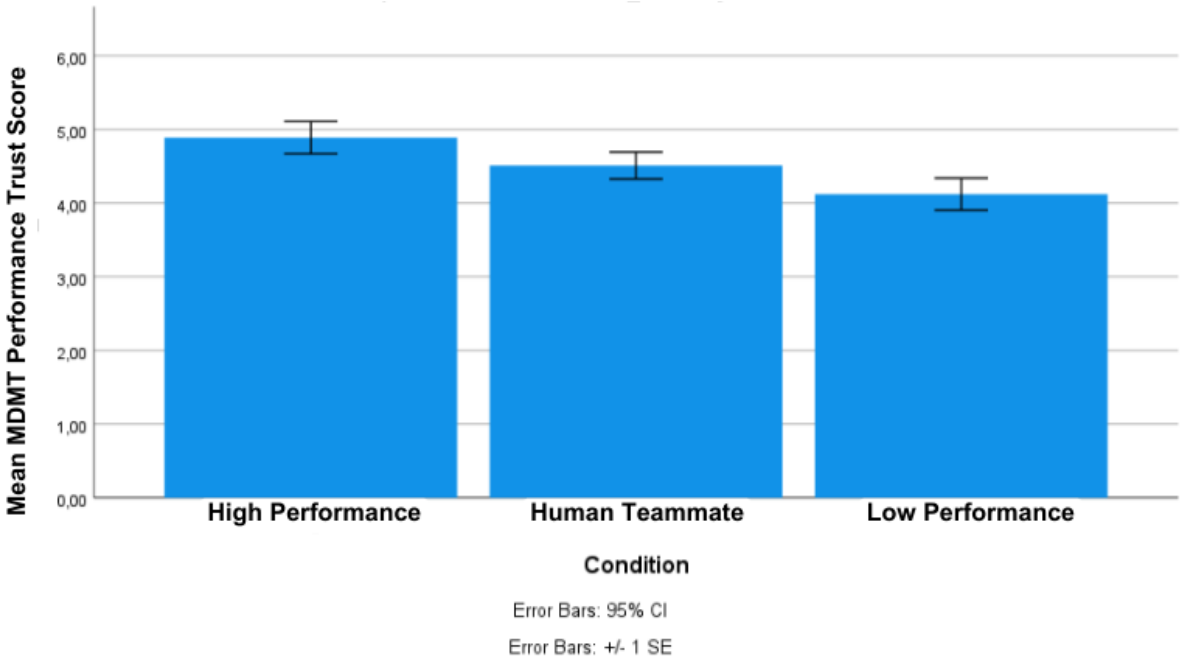


Figure 4.1: Graphic representation of the mean score of Performance Trust between the three conditions

4.2.3 Timesteps and Game Transitions analysis

For both levels of the game, we compared the number of timesteps that participants held the ball of toxic waste, the total number of timesteps they took to finish the game, and the number of transitions between different environment positions where they held the ball.

4.2.3.A Level 1 Results

The results indicated that there was no significant difference between the timesteps that the participants held the toxic waste of the three groups at level 1 of the game. The same conclusion was obtained regarding the total timesteps participants took to finish the game and regarding the number of transitions between different environment positions where participants held the ball.

In Table 4.13 we show the details of the results of the analysis performed, for each condition. The variable Hold represents the number of timesteps that the participants held the ball, the variable Total represents the total timesteps participants took to finish the game, and the variable Transitions represents the total number of transitions performed between different environment positions by the participants while they were holding the ball. For instance, the LP group had a mean number of 206.90 for the variable Hold with a standard error (SE) of 103.75, while the HP group had a mean score of 198.02 for the variable Hold with a standard error of 114.83, and the HT group had a mean number of 213.41 with a standard error of 91.05. In the variable Hold, the Kruskal-Wallis H value is 1.83, the degrees of freedom (df) are 2, and the corresponding p-value is .400. In addition, we show in Figures 4.2 to 4.4 a graphic representation of the mean values of each variable measured between the three conditions.

Table 4.13: Kruskal-Wallis test results of the timesteps and transition analysis of level 1

Variable	**Mean LP**	SE **LP**	**Mean HP**	SE **HP**	**Mean HT**	SE **HT**	H	df	p
Hold	206.90	103.75	198.02	114.83	213.41	91.05	1.83	2	.400
Total	989.85	678.02	852.80	423.51	1065.28	587.55	4.29	2	.117
Transitions	9.20	4.40	9.37	4.72	9.25	6.34	.800	2	.670

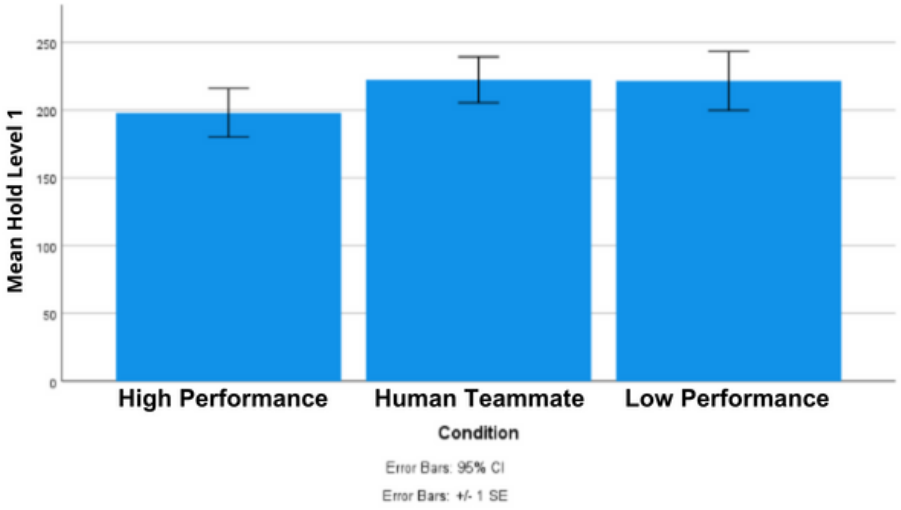


Figure 4.2: Graphic representation of the Hold Variable Mean by Condition (Level 1)

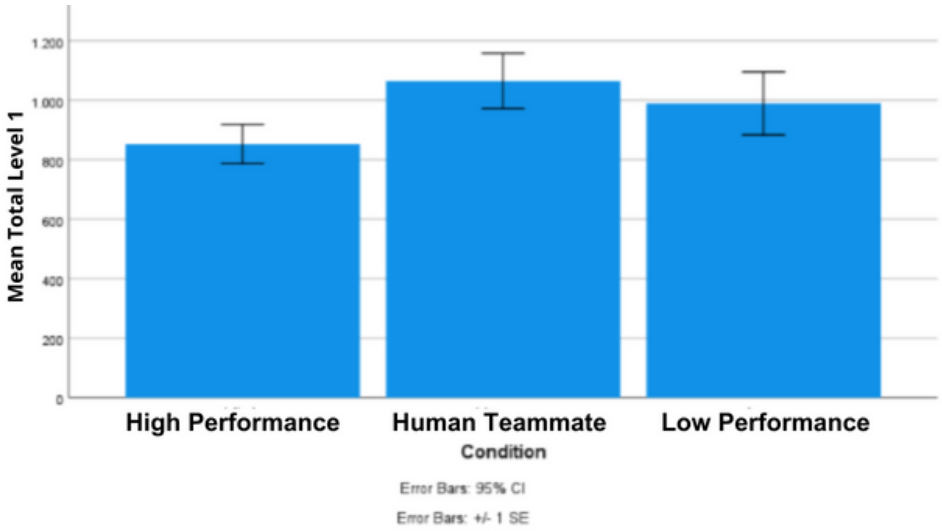


Figure 4.3: Graphic representation of the Total Variable Mean by Condition (Level 1)

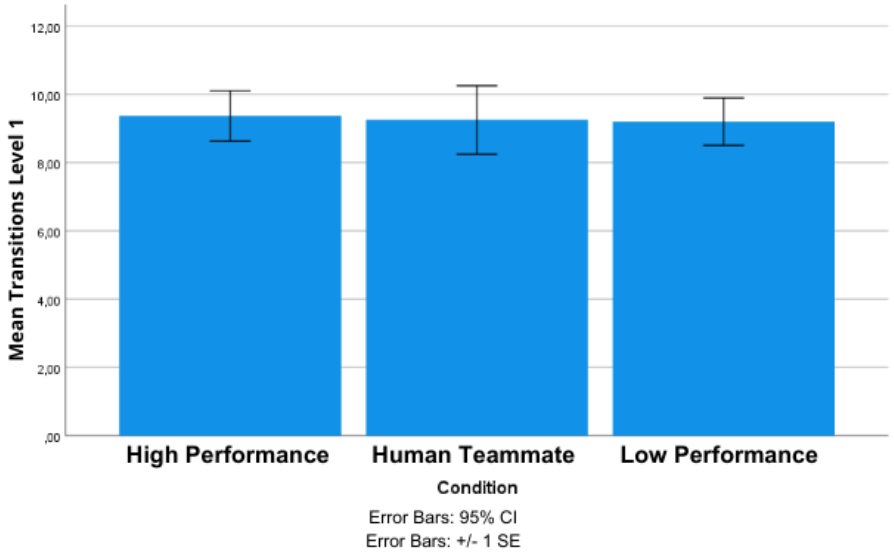


Figure 4.4: Graphic representation of the Transitions Variable Mean by Condition (Level 1)

4.2.3.B Level 2 Results

The results indicated that there was no significant difference between the timesteps that the participants held the toxic waste of the three groups at level 2 of the game. The same conclusion was obtained regarding the total timesteps participants took to finish the game and regarding the number of transitions between different environment positions where participants held the ball.

In Table 4.14 we show the details of the results of the analysis performed, for each condition. The variable Hold represents the number of timesteps that the participants held the ball, the variable Total

represents the total timesteps participants took to finish the game, and the variable Transitions represents the total number of transitions performed between different environment positions by the participants while they were holding the ball. For instance, the LP group had a mean number of 299.75 for the variable Hold with a standard error (SE) of 232.10, while the HP group had a mean score of 246.63 for the variable Hold with a standard error of 134.95, and the HT group had a mean number of 246.60 with a standard error of 125.06. In the variable Hold, the Kruskal-Wallis H value is 2.16, the degrees of freedom (df) are 2, and the corresponding p-value is .339. In addition, we show in Figures 4.5 to 4.7 a graphic representation of the mean values of each variable measured between the three conditions.

Table 4.14: Kruskal-Wallis test results of the timesteps and transition analysis of Level 2

Variable	**Mean LP**	SE **LP**	**Mean HP**	SE **HP**	**Mean HT**	SE **HT**	H	df	p
Hold	299.75	232.10	246.63	134.95	246.60	125.06	2.16	2	.339
Total	1591.28	1174.62	1428.38	695.90	1616.33	923.41	.945	2	.623
Transitions	12.06	6.42	11.43	7.03	9.73	5.06	5.43	2	.076

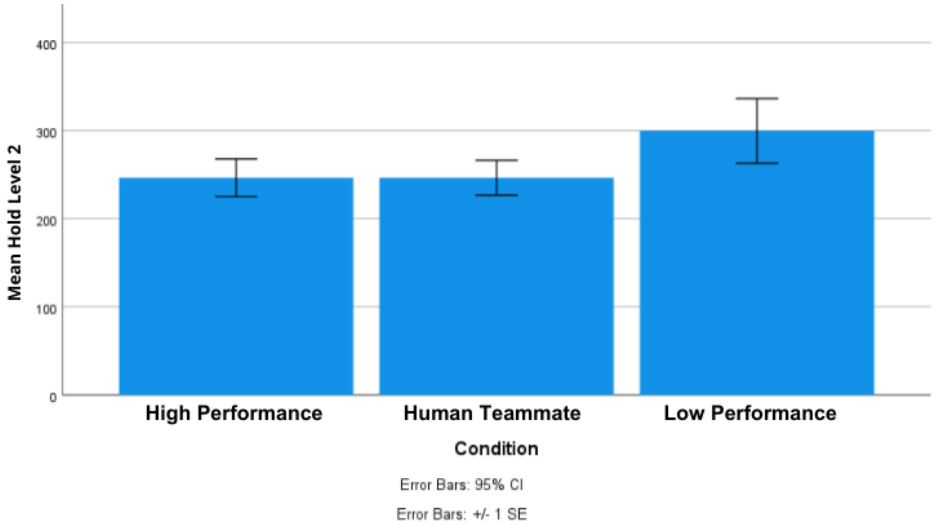


Figure 4.5: Graphic representation of the Hold Variable Mean by Condition (Level 2)

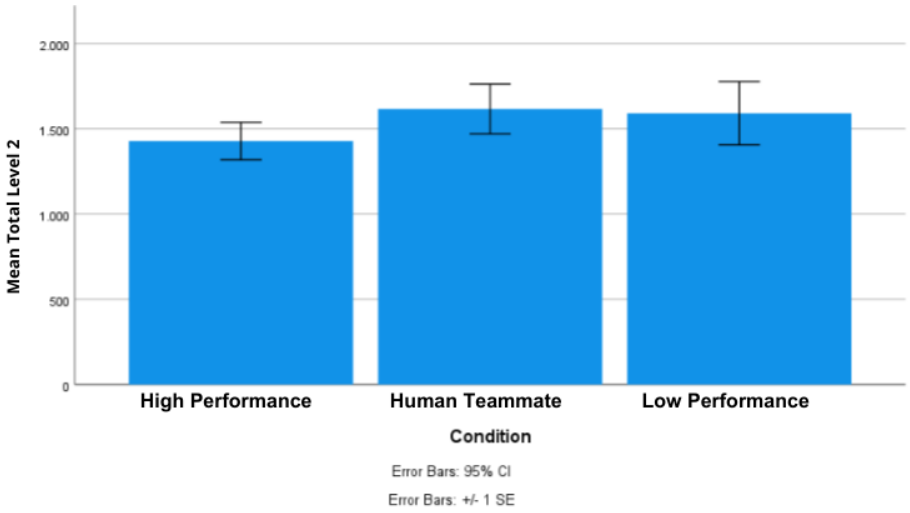


Figure 4.6: Graphic representation of the Total Variable Mean by Condition (Level 2)

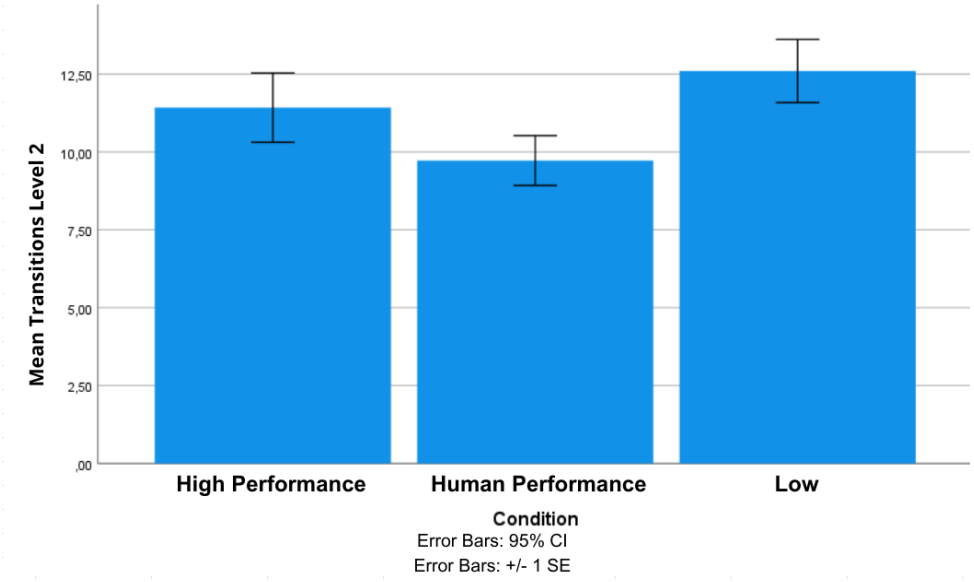


Figure 4.7: Graphic representation of the Transitions Variable Mean by Condition (Level 2)

4.2.4 P2T [3] questionnaire

We analyzed the propensity to trust of the participants of the three conditions, by comparing the average result of each item that composed the P2T scale [3]. The results indicated that there was no significant difference between the propensity to trust of the three groups.

In Table 4.15 we show the details of the results of the analysis performed, for each condition. Each sentence on the scale was scored on a scale from 1 to 5. The LP group had a mean P2T score of 3.07 with a standard error (SE) of .856, while the HP group had a mean P2T score of 3.28 with a standard error of .701, and the HT group had a mean P2T score of 2.92 with a standard error of .804. The one-

way ANOVA Z value is 2.083, the degrees of freedom (df) are (2, 118) and the corresponding p-value is .129. Additionally, we present in Figure 4.8 a graphic representation of the mean values of the propensity to trust measured in the three condition groups.

Table 4.15: One-way ANOVA test results of the P2T scale [3]

Variable	**Mean LP**	SE **LP**	**Mean HP**	SE **HP**	**Mean HT**	SE **HT**	Z	df	p
P2T	3.07	.856	3.28	.701	2.92	.804	2.08	2, 118	.129

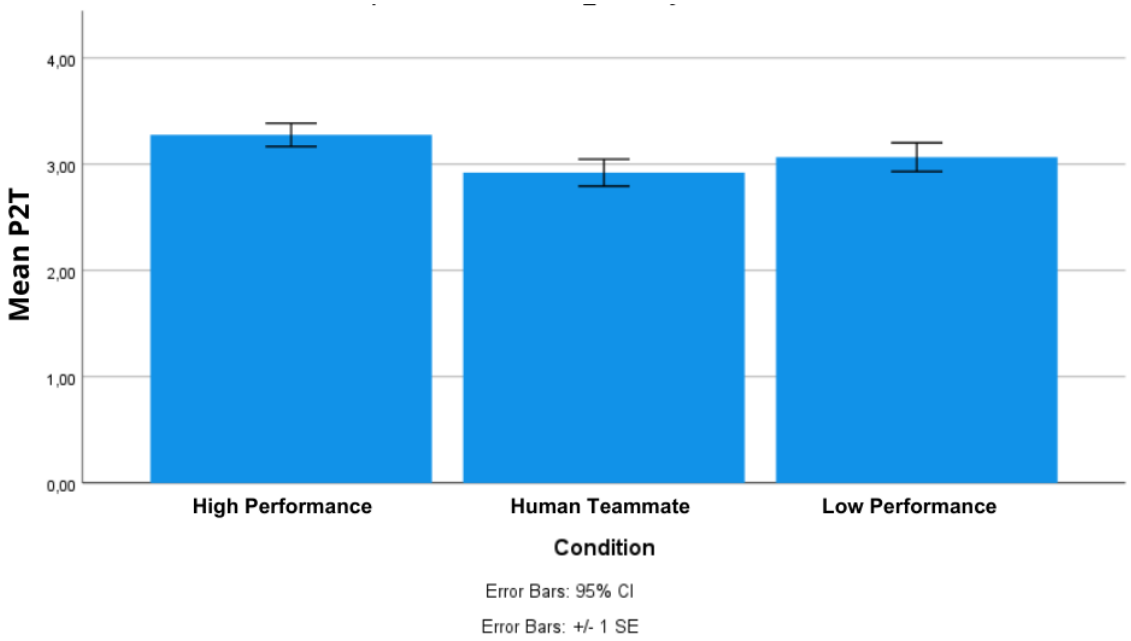


Figure 4.8: Graphic representation of the P2T Mean by Condition

Additionally, we thought of studying if a participant's propensity to trust influences their trust in the robot. To do that, we calculated the mean P2T of all the 121 participants, which was 3.09, and the respective standard error, which was 0.796. With these values, we decided that a participant had a high P2T if their P2T score was higher than the mean P2T score plus the standard error, i.e., 3.89. On the contrary, we considered a participant to have a low P2T if their P2T score was lower than the mean P2T score minus the standard error, i.e., 2.29.

For each of the three conditions (HP, LP, and HT groups), we compared separately the participants with High vs. Low propensity to trust, regarding their trust score in the robot. Beginning with the participants that belong to the HP group, 11 participants were classified as having a high propensity to trust (HighP2T group), while 3 had a low propensity to trust (LowP2T group). Despite being such a small sample, we tried to compare if there was any difference in terms of trust in the robot between these two groups. The results indicated group HighP2T had a significantly higher score on the three trust scales (Performance, Reliable, and Competent Scale) than group LowP2T.

In Table 4.16 we show the details of the results of the analysis performed for each trust scale, for both groups. We remind that each item was scored on a scale from 0 to 7, or with the option "Does not fit" whenever a participant believed that characteristic could not be applied to the agent. For example, for the Performance Scale, the LowP2T group had a mean score of 3.15 with a standard error (SE) of 1.61, while the HighP2T group had a mean score of 5.41 with a standard error of 1.06. The two-sample t-test value is 2.97, the degrees of freedom (df) are 12, and the corresponding p-value is .012.

Table 4.16: Two-sample t-test results of the Performance Trust Scale, the Reliable and Competent Trust subscales [2] (HP group)

Scale	**Mean LowP2T**	SE **LowP2T**	**Mean HighP2T**	SE **HighP2T**	t	df	p
Performance Scale	3.15	1.61	5.41	1.06	2.97	12	.012
Reliable Subscale	3.06	2.30	5.39	1.04	2.68	12	.020
Competent subscale	3.25	1.15	5.43	1.15	2.92	12	.013

Analyzing the participants that were part of the LP condition, 6 participants were classified as having a high propensity to trust, while 6 were considered to have a low propensity to trust. The results indicated that there was no significant difference between the trust score from all trust scales of the HighP2T group and the LowP2T group. Table 4.17 shows the details of the analysis performed. Each item composing the scales was scored on a scale from 0 to 7, or with the option "Does not fit" whenever a participant believed that characteristic could not be applied to the agent. For example, for the Performance Scale, the LowP2T group had a mean score of 3.78 with a standard error (SE) of 2.09, while the HighP2T group had a mean score of 4.08 with a standard error of 1.09. The two-sample t-test value is .324, the degrees of freedom (df) are 11, and the corresponding p-value is .752.

Table 4.17: Two-sample t-test results of the Performance Trust Scale, the Reliable and Competent Trust subscales [2] (LP group)

Scale	**Mean LowP2T**	SE **LowP2T**	**Mean HighP2T**	SE **HighP2T**	t	df	p
Performance Scale	3.78	2.09	4.08	1.09	.324	11	.752
Reliable Subscale	3.92	2.15	4.38	1.17	.465	11	.651
Competent subscale	3.68	2.15	3.79	1.10	.116	11	.910

Finally, among the participants of the HT group, 6 participants were classified as having a high propensity to trust, while 11 were considered to have a low propensity to trust. The results indicated that there was no significant difference between the trust score from all trust scales of the HighP2T group and the LowP2T group. Tables 4.18 and 4.19 show the details of analysis performed. Each item composing the scales was scored on a scale from 0 to 7, or with the option "Does not fit" whenever a participant believed that characteristic could not be applied to the agent. For example, when looking at Table 4.18 for the Performance Scale the LowP2T group had a mean score of 4.00 with a standard error (SE) of 1.18, while the HighP2T group had a mean score of 4.06 with a standard error of 1.68. The two-sample t-test value is .003, the degrees of freedom (df) are 14, and the corresponding p-value is

.998. In Table 4.19, U denotes the Mann-Whitney U value, and p is the corresponding p-value.

Table 4.18: Two-sample t-test results of the Performance Trust Scale, the Reliable and Competent Trust subscales [2] (HT group)

Scale	Mean LowP2T	SE LowP2T	Mean HighP2T	SE HighP2T	t	df	p
Performance Scale	4.00	1.18	4.06	1.68	.003	14	.998
Reliable Subscale	3.97	1.23	4.08	1.79	.056	14	.956

Table 4.19: Mann-Whitney U Test results of the Performance Trust Scale, the Reliable and Competent subscales [2] (HT group)

Scale / Item	Mean LowP2T	SE LowP2T	Mean HighP2T	SE HighP2T	U	p
Competent Subscale	4.03	1.25	4.08	1.71	23.00	.689

4.2.5 Main Study Discussion

The results showed that the participants found a high correlation between the RoSAS competence attributes [1], especially in the HP group, where this correlation was significantly higher than the other groups. An exception is found for the reliable item, where the LP group had a similar score to the HP group. This outcome is reflected in the types and levels of trust the participants felt about the robot.

Furthermore, the participants had relatively a high score of trust in the robot, but the HP group had a significantly higher performance and competent trust score than the LP group. The fact that the reliable trust was equal in all conditions highly correlates with the factors that affect trust, discussed in Section 2.2.2. We have discussed that automation faults cause a decline in trust [25] and that system predictability affects trust development [31]. Since the robot had areas where its behavior failed sometimes, and could not be predicted by users when it would happen, it is expected that the competence and performance trust are lower in the LP group. However, in the LP group, the highest trust score is regarding reliable trust since the robot is able to complete its task.

Regarding the timestep analysis, we could not obtain significant differences between the three groups. However, we made an effort to analyze the results obtained to understand if there was any tendency. By the graphics presented in Section 4.1.3, we could understand that the HP group, on average, held the ball less than the LP group, for both game levels. The same pattern was obtained for the total timesteps to finish the game. Regarding the number of transitions while holding the ball, we can not see any tendency in level 1, but interestingly, there is a difference in level two: the HT group transitioned less with the ball, and the LP group had a higher average of transitions. This evolution of difference in behavior shows the need to perform a longer experience, to allow people to form a game strategy.

When analyzing the figures presented in Appendix A, we concluded that for both levels, the participants showed some similar patterns of behavior, however the HP group seems to explore less the

environment, and has a lower frequency of timesteps (both in general and for holding the ball) than the other condition groups. In addition, the areas where the participants seem to hold more the ball are the radioactive zones, where the robot may suffer from interference, and, therefore, have behavior malfunctions. To deliver the toxic waste object to the robot, the participants had to be facing the robot, and with the malfunctions at the radioactive zones, that could be a difficult task to do.

Moreover, the propensity to trust was similar in the three conditions, which indicates that our priming of trust worked. However, we attempted to study if the participants' propensity to trust influences their trust in the robot. Due to the low number of participants in each condition, we can not conclude anything despite having differences in the results that suggest that people with a high propensity to trust trusted more the robot than people with a low propensity to trust. What we have noticed is that in the HP group, there were more cases of participants with a high propensity to trust than in the other conditions. On the contrary, there were more people with a low propensity to trust in the LP group. This suggests that even though the P2T questionnaire is independent of the scenario, the participants may have been influenced by their condition when answering this questionnaire.

Along these lines, we recall our hypotheses:

- **Hypothesis 1:** When people have a **low trust** level in their agent teammate, they will carry the toxic waste objects for longer periods of time, they will not wait for the agent and will go toward the agent's position while handling the toxic object.

- **Hypothesis 2:** When people have a **high trust** level in their agent teammate, they will wait for the agent to arrive closer to the objects and only then they will pick up the toxic object and drop in the agent's container.

With the results obtained, we can not confirm our hypotheses. Nevertheless, we could infer that the perceived competence attributes of a robot are highly correlated with the trust a person feels about it. Although not statistically significant, we understand that there were some behavior differences between the three groups and that some adjustments and new metrics must be performed in the future to obtain significant results.

We end this chapter by providing possible applications of this work. With these results, we can recognize that people are capable of feeling different levels of trust in a robot teammate who acts equally with all participants. With some improvements to the study, such as more interaction time and metrics that are not compromised by the environment characteristics, we will be able to collect different behavior clusters that represent different levels of trust. In Section 5.2 we will address some of the limitations that compromised this work and the modifications that are essential to achieve this objective. Finally, it will be possible to use these behavior patterns to infer the level of trust of the human teammate. As well, by following a similar concept as Chen et al. [18], we can adjust the behavior of the AHT agent according to

the level of trust of the human teammate. After such development, the following research question can be studied: if a trustworthy AHT algorithm increases or not human-robot team performance.

5

Conclusion

Contents

The increasing use of robots in society for crucial collaboration tasks, such as some healthcare services, demands a need for people to establish trusting relationships with their robot teammates.

In addition, some of these scenarios may require collaboration with unknown human teammates and their trust levels in a robot teammate may vary from person to person. These scenarios can be seen as an ad hoc teamwork problem, which consists of the problem of collaborating with other unknown agents without prior coordination methods [12]. In this work, we proposed a research method to identify how people behave with different levels of trust in a collaborative scenario, the Toxic Waste Game.

5.1 Conclusions

We concluded that people are able to feel different trust levels in a robot teammate. We also found that the trust level is correlated with the competence attributes: participants who had a higher trust in the robot, also perceived him as more capable, responsive, interactive, competent, and knowledgeable.

We could not obtain significant differences concerning the participant's behavior, but some patterns and tendencies in the results were found. Factors, such as the group with the highest trust in the robot held, on average, less toxic waste for both game levels or exploring less the game environment, indicate that some improvements to the study may be able to induce any difference in the conduct of people.

On top of that, participants' propensity to trust was not significantly different in all conditions, which means the priming done in our study was capable of influencing the participant's trust. Although we could not conclude if a participant's propensity to trust influenced their trust in the robot, our primary results point out that it is possible to have trust differences in the robot when dealing with participants with a different propensity to trust.

5.2 System Limitations and Future Work

Concerning the system limitations, we start by stating that the agent's policy for its trajectory was not smooth enough. The robot would not always follow the best trajectory due to the way the game was modeled. Although efforts were made to fix this, the time constraints did not allow the development of a good human policy that would permit the development of a new policy for the robot using different methods of training.

Furthermore, the experience was too short and did not allow a bigger evolution regarding the participant's behavior. We suggest, in the future, to increase the number of game levels and measure the trust of participants more often to evaluate how evolves through time.

As well, the game itself may have compromised the metrics used. The fact of having radioactive zones that cause robot behavior malfunctions makes the task of the participants to place the toxic object

inside the robot very difficult, which increases the number of transitions while holding the toxic waste.

Additionally, the P2T analysis suggested that the condition where the participants were placed seemed to influence how people answered the questionnaire. In the future, we believe it would be interesting to provide the questionnaire to the participants before and after the experiment in order to not only obtain a P2T questionnaire that was not influenced by the priming effect but also allow us to understand if there was any difference regarding the participants' propensity to trust and how did evolve.

For future work, the first step is to work on the previously mentioned issues to discover and form behavior clusters based on the participant's trust. This will allow us to answer our initial research question of how trust influences human behavior when cooperating with an unknown agent. Once this is achieved, we would like to propose the following research questions: How can the conclusions obtained from our initial research question be integrated into the decision-making process of an ad hoc teamwork algorithm? Does an ad hoc teamwork algorithm that is tailored for trustworthiness improve human-robot team performance? To study these questions, we recommend the development of a new AHT algorithm. This could be done by extending the work of Ribeiro et al. [14], where the agent must infer the level of trust at hand as the task identification step [13]. The states can be modeled with the agent's location, the location of the human teammate, and the position of the toxic waste objects. An additional variable to mark whether the human is holding an object can be added too. With this information, the robot can infer the trust level by identifying state patterns that represent the behavior clusters previously identified.

After the trust level is identified, the agent must act accordingly: have a more passive or active posture in the game, in order to improve task performance. If the trust level is low, the agent must let the human teammate complete the task without interviewing too much. On the other hand, if the trust level is high, the agent must take the initiative to go closer to the toxic waste objects. In Figure 5.1, we can see a simple diagram that explains the flow of the AHT agent behavior. As in BOPA [14], the AHT agent must infer both the task and the policy of the human teammate by observing how the state evolves during the interaction.

Finally, the last guideline we provide is concerning the evaluation methods necessary to perform a future study. To evaluate the trustworthy ad hoc teamwork algorithm, by effectiveness and efficiency. The algorithm is effective if it is capable of identifying correctly the level of trust of the human participant. Additionally, the algorithm is efficient if the AHT agent, together with the human, is capable of solving the task in near-optimal time. In addition, trust levels on the robot can be measured by the same questionnaires used in this project [1–3], before and after the interaction, to understand how trust evolved through time and also to check if the agent correctly identified the level of trust through time. For example, before the interaction, a video of the agent interacting in the environment should be shown to the participants in pursuance of establishing an initial level of trust. Lastly, to understand if accounting human trust in the decision-making process increases team performance, the trustworthy AHT agent

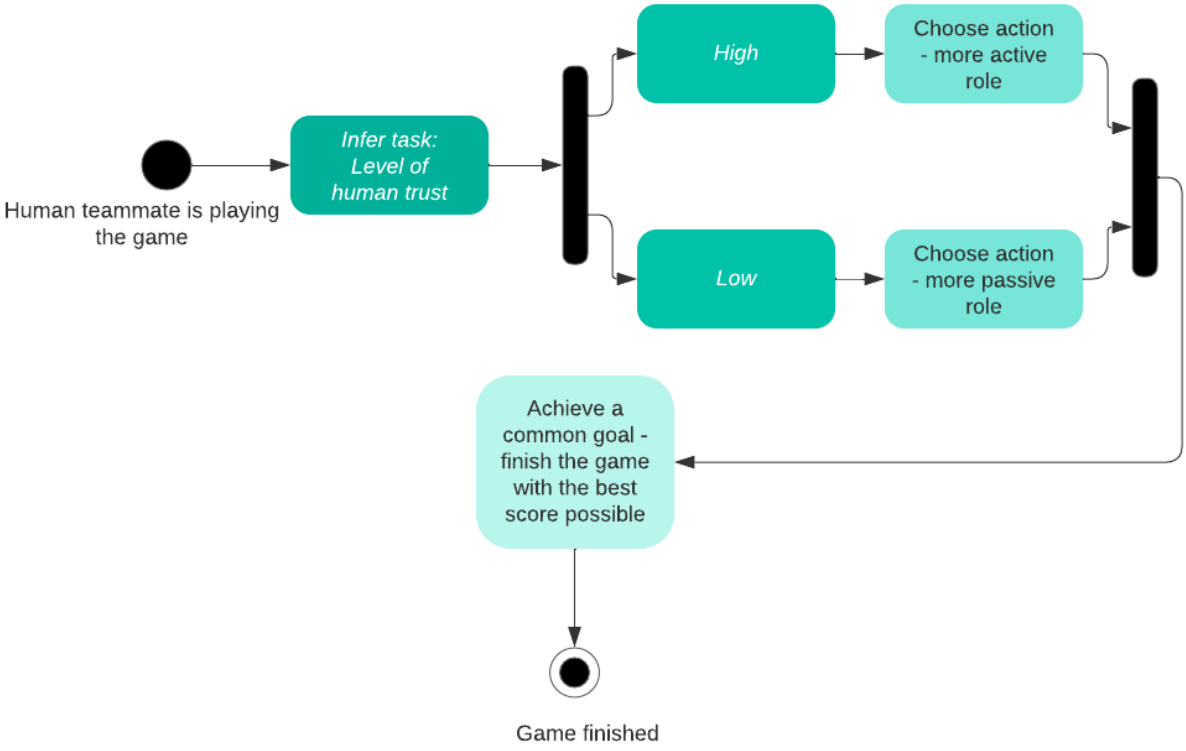


Figure 5.1: Simple diagram of the trustworthy AHT agent algorithm

should be compared to an AHT agent that does not take into account the trust level and will always act in an optimal way, regardless of how much trust the human participant has in it.

Bibliography

[1] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The robotic social attributes scale (rosas): Development and validation," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 254–262.

[2] D. Ullman and B. F. Malle, "What does it mean to trust a robot? steps toward a multidimensional measure of trust," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 263–264.

[3] M. L. Frazier, P. D. Johnson, and S. Fainshmidt, "Development and validation of a propensity to trust scale," *Journal of Trust Research*, vol. 3, no. 2, pp. 76–97, 2013.

[4] 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*, ser. HRI '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 92–100.

[5] A. Carrasco, "Adapting behaviour based on trust in human-agent ad hoc teamwork," 2022.

[6] D. Ullman and B. F. Malle, "Mdmt: multi-dimensional measure of trust," 2019.

[7] M. K. X. J. Pan, E. A. Croft, and G. Niemeyer, "Validation of the robot social attributes scale (rosas) for human-robot interaction through a human-to-robot handover use case," 2017.

[8] J. A. Pepito and R. Locsin, "Can nurses remain relevant in a technologically advanced future?" *International Journal of Nursing Sciences*, vol. 6, no. 1, pp. 106–110, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352013218301765

[9] R. Tasaki, M. Kitazaki, J. Miura, and K. Terashima, "Prototype design of medical round supporting robot "terapio"," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 829–834.

[10] N. Ruangpayoongsak, H. Roth, and J. Chudoba, "Mobile robots for search and rescue," in *IEEE International Safety, Security and Rescue Rototics, Workshop, 2005.*, 2005, pp. 212–217.

[11] A. Hong, O. Igharoro, Y. Liu, F. Niroui, G. Nejat, and B. Benhabib, "Investigating human-robot teams for learning-based semi-autonomous control in urban search and rescue environments," *Journal of Intelligent & Robotic Systems*, vol. 94, no. 3, pp. 669–686, 2019.

[12] P. Stone, G. Kaminka, S. Kraus, and J. Rosenschein, "Ad hoc autonomous agent teams: Collaboration without pre-coordination," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 24, no. 1, pp. 1504–1509, Jul. 2010. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/7529

[13] F. S. Melo and A. Sardinha, "Ad hoc teamwork by learning teammates' task," *Autonomous Agents and Multi-Agent Systems*, vol. 30, no. 2, pp. 175–219, 2016.

[14] J. G. Ribeiro, M. Faria, A. Sardinha, and F. S. Melo, "Helping people on the fly: Ad hoc teamwork for human-robot teams," in *Progress in Artificial Intelligence*, G. Marreiros, F. S. Melo, N. Lau, H. Lopes Cardoso, and L. P. Reis, Eds. Cham: Springer International Publishing, 2021, pp. 635–647.

[15] J. G. Ribeiro, L. M. Henriques, S. Colcher, J. C. Duarte, F. S. Melo, R. L. Milidiú, and A. Sardinha, "HOTSPOT: An Ad Hoc Teamwork Platform for Mixed Human-Robot Teams," 11 2021. [Online]. Available: https://www.techrxiv.org/articles/preprint/HOTSPOT_An_Ad_Hoc_Teamwork_Platform_for_Mixed_Human-Robot_Teams/17026013

[16] J. G. Ribeiro, C. Martinho, A. Sardinha, and F. S. Melo, "Assisting unknown teammates in unknown tasks: Ad hoc teamwork under partial observability," *CoRR*, vol. abs/2201.03538, 2022. [Online]. Available: https://arxiv.org/abs/2201.03538

[17] S. Herse, J. Vitale, B. Johnston, and M.-A. Williams, "Using trust to determine user decision making task outcome during a human-agent collaborative task," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 73–82.

[18] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with trust for human-robot collaboration," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 307–315.

[19] R. Mirsky, I. Carlucho, A. Rahman, E. Fosong, W. Macke, M. Sridharan, P. Stone, and S. V. Albrecht, "A survey of ad hoc teamwork research," 2022.

[20] S. Barrett, *Making friends on the fly: Advances in ad hoc teamwork*. Springer, 2015, vol. 603.

[21] S. Barrett and P. Stone, "Cooperating with unknown teammates in complex domains: A robot soccer case study of ad hoc teamwork," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[22] J. Suriadinata, W. Macke, R. Mirsky, and P. Stone, "Reasoning about human behavior in ad hoc teamwork," in *Adaptive and learning Agents Workshop at AAMAS 202*, 2021.

[23] Y. Hanina, R. Mirsky, W. Macke, and P. Stone, "Quantifying human rationality in ad-hoc teamwork."

[24] 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)*, 2016, pp. 101–108.

[25] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[26] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the "trust perception scale-hri"," in *Robust intelligence and trust in autonomous systems*. Springer, 2016, pp. 191–218.

[27] A. R. Wagner, P. Robinette, and A. Howard, "Modeling the human-robot trust phenomenon: A conceptual framework based on risk," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 4, nov 2018.

[28] T. Law and M. Scheutz, *Trust: Recent concepts and evaluations in human-robot interaction*, 01 2021, pp. 27–57.

[29] 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)*, 2015, pp. 1–8.

[30] V. Groom and C. Nass, "Can robots be teammates?: Benchmarks in human–robot teams," *Interaction studies*, vol. 8, no. 3, pp. 483–500, 2007.

[31] 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.

[32] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill, "Is it my looks? or something i said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams," in *Persuasive Technology*, J. Ham, E. Karapanos, P. P. Morita, and C. M. Burns, Eds. Cham: Springer International Publishing, 2018, pp. 56–69.

[33] T. Bridgwater, M. Giuliani, A. v. Maris, G. Baker, A. Winfield, and T. Pipe, "Examining profiles for robotic risk assessment: Does a robot's approach to risk affect user trust?" in *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 23–31.

[34] P. Shu, C. Min, I. Bodala, S. Nikolaidis, D. Hsu, and H. Soh, "Human trust in robot capabilities across tasks," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 241–242.

[35] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011, pMID: 22046724.

[36] D. V. Pynadath, N. Wang, and S. Kamireddy, "A markovian method for predicting trust behavior in human-agent interaction," in *Proceedings of the 7th International Conference on Human-Agent Interaction*, ser. HAI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 171–178.

[37] J. J. Lee, B. Knox, J. Baumann, C. Breazeal, and D. DeSteno, "Computationally modeling interpersonal trust," *Frontiers in psychology*, p. 893, 2013.

[38] C. Breazeal, C. Kidd, A. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 708–713.

[39] S. Habibian and D. P. Losey, "Encouraging human interaction with robot teams: Legible and fair subtask allocations," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6685–6692, 2022.

[40] "Prolific," https://www.prolific.co/.

[41] K. Vodrahalli, R. Daneshjou, T. Gerstenberg, and J. Zou, "Do humans trust advice more if it comes from ai? an analysis of human-ai interactions," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 763–777. [Online]. Available: https://doi.org/10.1145/3514094.3534150

A

Visual Representation of the Participant's Behavior

Below we show in Figures A.1 to A.4 a visual representation of the total timesteps participants spent either in general or holding the ball in each condition and each game level.

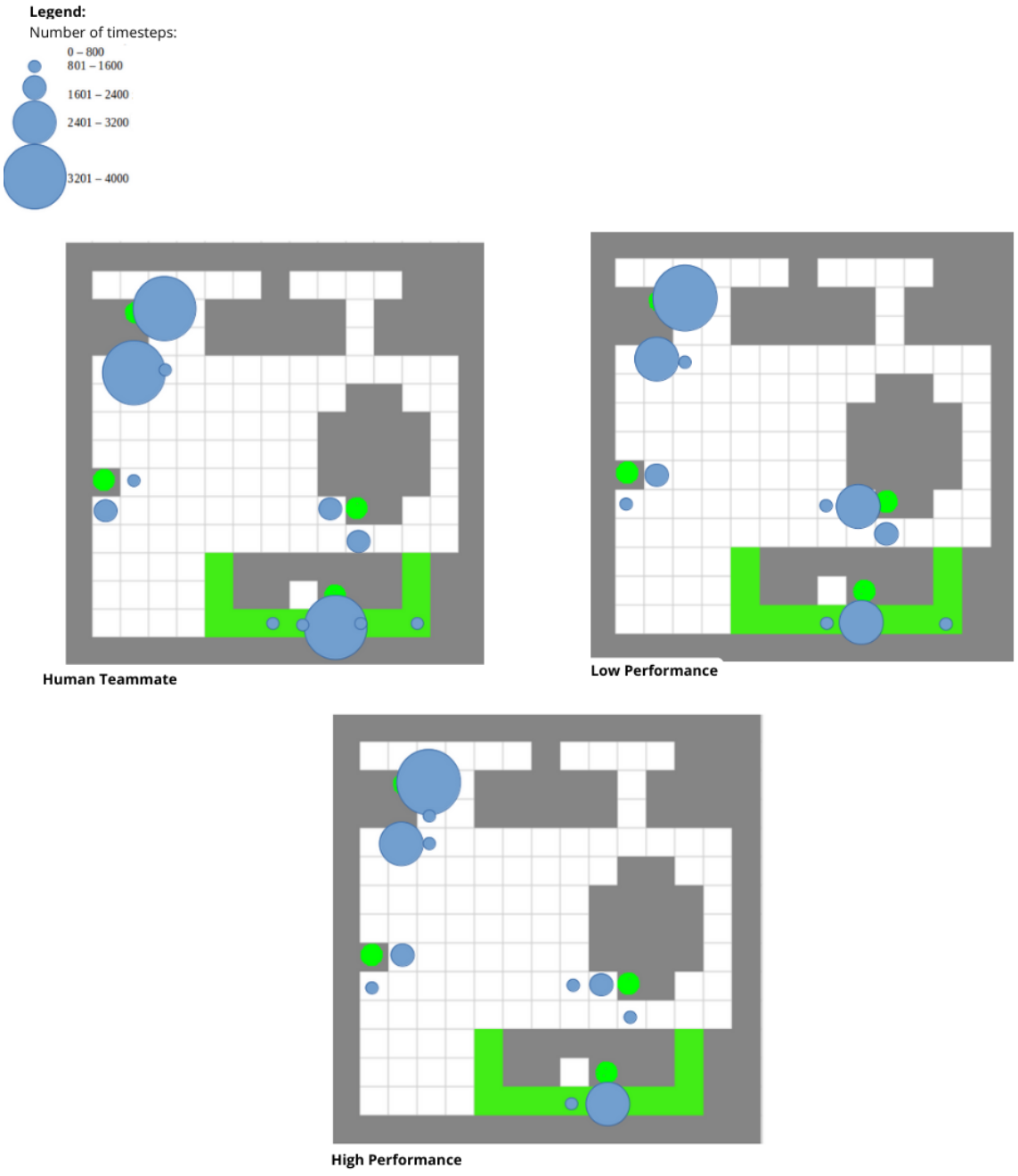


Figure A.1: Number of timesteps participants spent in each position by group (Level 1)

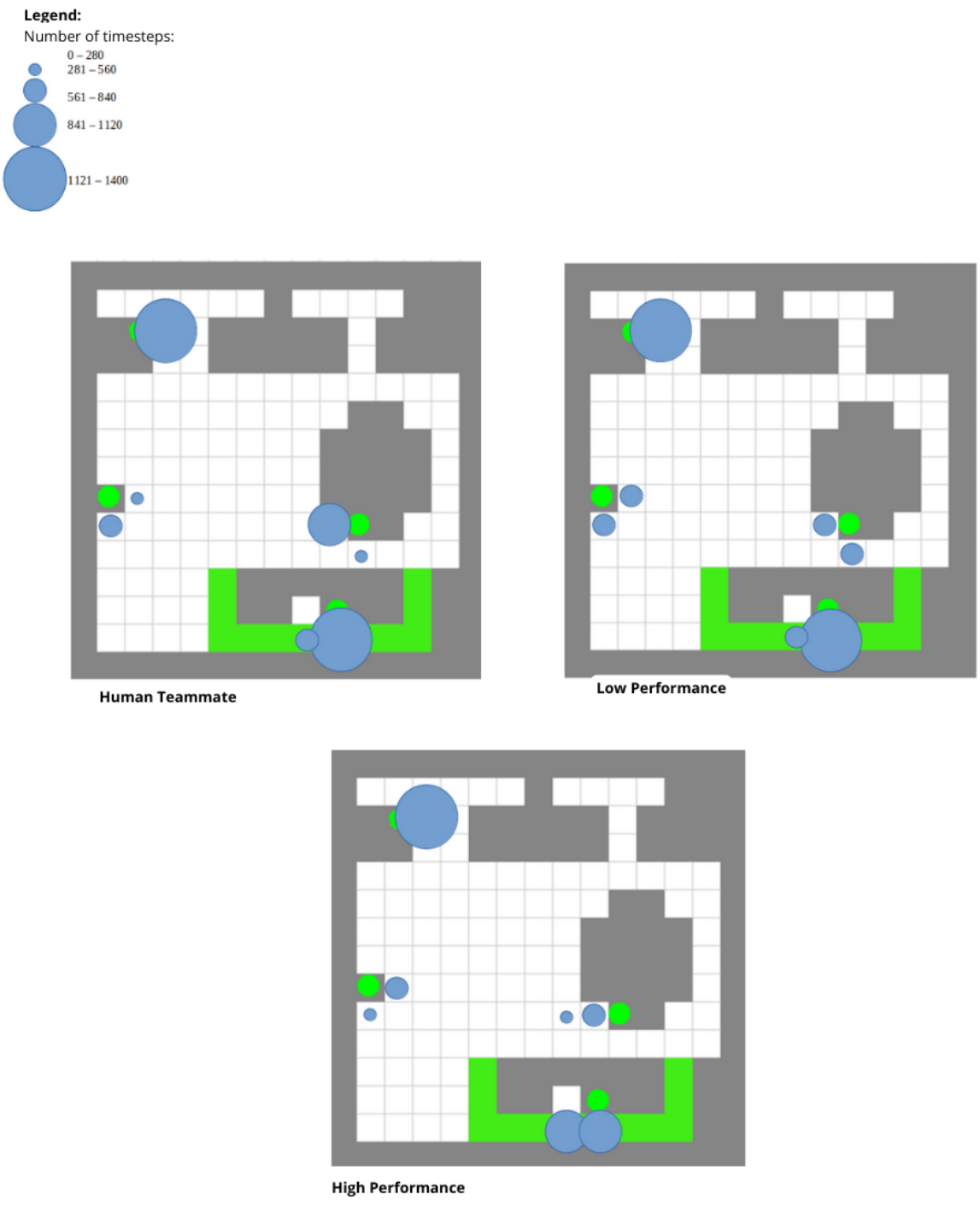


Figure A.2: Number of timesteps participants spent in each position while holding the toxic waste by group (Level 1)

Figure A.3: Number of timesteps participants spent in each position by group (Level 2)

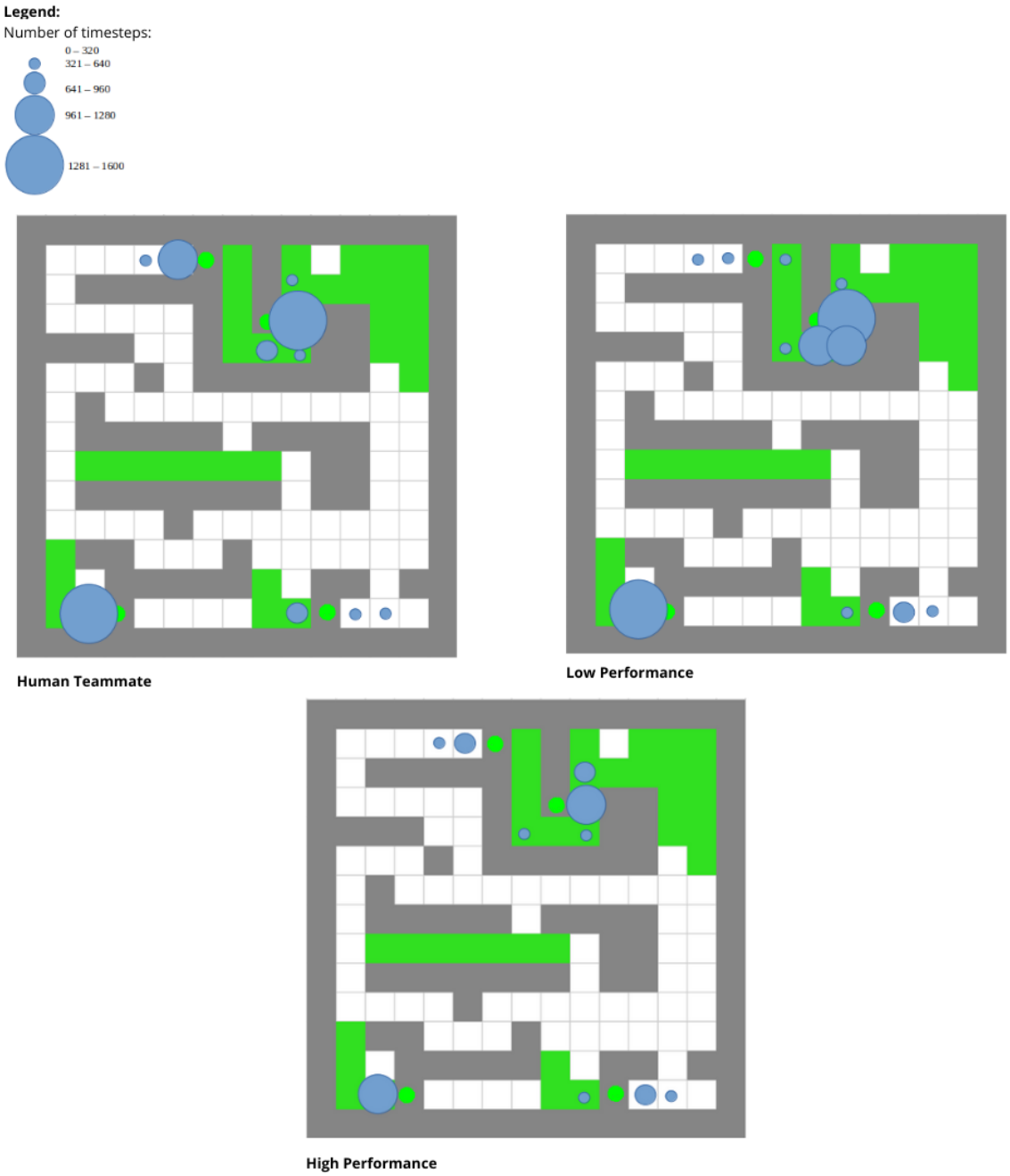


Figure A.4: Number of timesteps participants spent in each position while holding the toxic waste by group (Level 2)