robotics

MDPI

Article

Explanations from a Robotic Partner Build Trust on the Robot's Decisions for Collaborative Human-Humanoid Interaction

Misbah Javaid [1,*] and Vladimir Estivill-Castro [2]

[1] School of Information and Communication Technology, Griffith University, Brisbane 4111, Australia
[2] Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, 08018 Barcelona, Spain; vladimir.estivill@upf.edu
* Correspondence: misbah.javaid@griffithuni.edu.au

Abstract: Typically, humans interact with a humanoid robot with apprehension. This lack of trust can seriously affect the effectiveness of a team of robots and humans. We can create effective interactions that generate trust by augmenting robots with an explanation capability. The explanations provide *justification* and *transparency* to the robot's decisions. To demonstrate such effective interaction, we tested this with an interactive, game-playing environment with partial information that requires team collaboration, using a game called *Spanish Domino*. We partner a robot with a human to form a pair, and this team opposes a team of two humans. We performed a user study with sixty-three human participants in different settings, investigating the effect of the robot's explanations on the humans' trust and perception of the robot's behaviour. Our explanation-generation mechanism produces natural-language sentences that translate the decision taken by the robot into human-understandable terms. We video-recorded all interactions to analyse factors such as the participants' relational behaviours with the robot, and we also used questionnaires to measure the participants' explicit trust in the robot. Overall, our main results demonstrate that explanations enhanced the participants' understandability of the robot's decisions, because we observed a significant increase in the participants' level of trust in their robotic partner. These results suggest that explanations, stating the reason(s) for a decision, combined with the transparency of the decision-making process, facilitate collaborative human–humanoid interactions.

Keywords: trust; decision explanations; human-humanoid interaction

Citation: Javaid, M.; Estivill-Castro, V. Explanations From a Robotic Partner Build Trust on the Robot's Decisions for Collaborative Human-Humanoid Interaction. *Robotics* **2021**, *10*, 51. https://doi.org/10.3390/robotics10010051

Received: 1 November 2020
Accepted: 16 March 2021
Published: 23 March 2021

1. Introduction

Social robots are now deployed in human environments, such as in hotels, shops, hospitals, and particularly in roles as co-workers. These robots complement humans' abilities with their own skills. Hence, robots are expected to cooperate and contribute productively with humans as teammates. In recent years, the technical abilities of robotic systems have immensely improved, which has led to an increase in the autonomy and functional abilities of existing robots [1]. As robots' abilities increase, their complexity also increases, but the increased ability of the robot often fails to improve the competency of a human–robot team [2]. Effective teamwork between humans and robots requires trust. In situations with incomplete information, where humans need to interact and work as teammates with a robot, humans' trust in their robot teammate is crucial. In such cases, autonomous decision-making by the robot creates unpredictable and inexplicable situations for human teammates. Consequently, humans' lack of insight into the robot's decision-making process leads to a loss of trust in their robot teammate. In critical situations, such as search-and-rescue or completion of a *time-sensitive task*, humans cannot afford to lose trust in robot teammates.

Robots shall be required to explain and justify their decisions to humans, and humans will be more likely to accept those decisions as they realise the reasoning behind them. We postulate that a robot's decisions (which generate the robot's actions) can be communicated

through explanations to humans. We hypothesise that the explanations that express *how* a decision is made and *why* the decision is made are the most promising to generate trust. We argue that these types of explanation will lead to a higher inclination in humans to accept the robot as a trustworthy teammate. Trust is an important aspect for the cooperation of humans and robots as a team [3]. Trust directly affects humans' willingness to receive and accept robot-produced information and suggestions [4,5]. The absence of trust in human–robot interaction leads to disuse of a robot [6]. Ensuring an appropriate level of trust is a challenge to the successful integration of robotic assets into collaborative teams because under-reliance or over-reliance on a robot can lead to misuse of the robot [2].

Trust among humans appears to require explanations [7]. In essence, trust-building encompasses a more or less detailed understanding of the motives of a person that we may or may not trust. We accept explanations, or we may cast a validity verdict on them. Artificial intelligence researchers, within the area of expert systems, have also provided sufficient motivation to consider the contribution of explanations [8] to building trust in humans [9,10] and leading to the acceptability of these systems [11]. Hand-crafted explanations have also shown to be promising in providing enough transparency to humans [12].

It seems timely and essential to understand how to promote effective interaction between robots and humans, especially when robots and humans share a common goal but not the same resources, and are required to work as a team. An interaction, by definition, requires communication between humans and robots [13]. We stipulate that explanations can be used as an effective communication modality for robots, earning humans' trust. Explanations shall enable humans to track the performance and abilities of the robots. We suggest that insight into the robots' decision-making process can also lead to a human desire for interaction and acceptability and will help to establish smooth and trustworthy human–robot interactions. This paper reports on a study which set out to examine the effect of a robot's explanations on humans' level of trust. The explanations use *English like sentences*, in understandable human terms. We expect that understandable explanations will induce humans to adjust their mental model of the robot's behaviour, becoming more predisposed to trust the robot's actions. Thus, humans will work together with robots as a team to achieve a common goal.

For human–robot interaction, there has been little empirical evaluation of the influence of explanations on the humans' level of trust. Sheh et.al. [14] supported the argument that explainability does help to establish trust, and that robots should be given the ability to explain their behaviour to human counterparts. Sheh et.al. [14]'s work focuses on generating explainable decision-tree models that can explain decision-making processes in the form of IF-THEN–ELSE statements. In this way, they developed an explainable artificial intelligent agent that can explain its decisions to humans [14]. However, the study did not discuss the effectiveness of explanations in improving system understandability since it did not include a user study (the presentation includes a hypothetical example of a dialog between a human and the explainable intelligent robot). Other research [15] presented a needs-based architecture to produce verbal "why explanations" for the transparency of the goal-directed behaviour of a social robot. This other study suggested that the generated explanations will give humans the opportunity to make informed behaviour assessments of the robot [15]. Using long-term feedback, humans will communicate individual preferences, observe the robot's adoption of them, and be offered the robot's explanations. The claim is that humans will trust more personalised robot behaviour. Although that research [15] evaluated the architecture by creating an imaginary scenario with a mobile social robot, no user study was performed to evaluate the influence of explanations on humans' trust. Wang et al. [9] used a different approach to increase transparency by using a simulated robot to provide explanations for its actions. Explanations did not improve the team's performance. For some specific high-reliability conditions, trust was identified as an influential factor. Moreover, Wang [9] used an online survey and the analysis of the survey's responses indicated improvements in humans' acceptance of the robot's suggestions. One of the disadvantages of conducting an online survey to evaluate humans' perception of the robot's

attributes is that the human participants act only as observers. Such a human perception is incomplete, since it is missing the robot's physical presence and interaction [2]. Thus, it is unclear what happens in settings where humans and a robot interact directly in the same environment.

We focus here on a physical setting where the robot has been endowed with the communication ability of explaining autonomous decisions via natural language utterances. Such explanations aim to justify decisions and provide transparency, reduce uncertainty, establish some understanding of the robot's behaviour, eventually increase the humans' trust, and shift the humans' perception of the robot from a tool to a trustworthy teammate.

Our contribution consists of a user study that takes a more socially relevant approach by focusing on the physical interaction between humans and an autonomous social robot. We chose *Domino*, a team-based partial-information game, to form the basis of the interaction between humans and the robot. Game-playing scenarios are useful and powerful environments to establish human–robot interaction [16] because games provide an external, quantifiable measure of the underlying psychological state of a human's trust [17]. In particular, multi-player game environments not only maintain social behaviour when played in teams, but also develop trust dynamics among teammates to achieve the common goal of winning the game.

We selected the *Domino* game as the basis for our experimental paradigm for the following reasons. The environment of the game *Domino* is partially observable. A game of *Domino* involves two teams, with two members on each team, where each participant has incomplete information (the hand of each player is not revealed to any other player), but cooperation is required by members of a team to achieve a win. Because each player has different tiles, each player has different resources, and in *Domino*, the resources of teammates and opponents are unknown when making a decision. We configured mixed teams: a human and robot, facing an adversary team of only human participants. In this setting, the robot has two roles: first, as a team partner with a human (the role of teammate), and second, as a member of a human–robot team that competes against a team of two humans (the role of adversary). Figure 1a illustrates the heterogeneous team versus homogeneous team setting.

We want to examine the effect of explanations on the humans' level of trust, in an environment where a robot makes decisions, and those decisions influence the outcome. The primary motivation behind this study is the interaction between humans and robots, which is distant from a *master–slave* relationship and is close to a *peer-to-peer* relationship. We note that, as is common at present [18,19], the *human-in-the-loop* concept presents humans as supervising (intelligent, smart, AI-incorporating) machines. Our setting presents the heterogeneous team of robots and humans with equivalent intellectual roles. The humans make decisions, and so do the robots (not about physical tasks). In our case study, rather than the human controlling, supervising or monitoring machine behaviour [20], the robot's decisions alternate with human's decisions. We augment the robot with the ability to provide different types of explanation, which will influence the human. Explanations should disregard the complex behaviour of the robot, but since the explanations are understandable and intuitive, the human will build a trustworthy mental model of the robot's decision-making mechanism, even if such a mechanism is beyond simple and intuitive explanations. We shall show that these explanations will lead to the development of human trust in the robot.

(**a**) Human participants are playing the game (in teams) with the robot.

(**b**) The robot explains how to play the game (*Condition-2*)

(**c**) Human participants are attentive to the robot's explanations

Figure 1. Two types of team during the activities. Heterogeneous teams are composed of a robot and a human. Homogeneous teams have two humans.

Section 2 further surveys the literature on trust and explanations in the context of human–robot interaction. Although changing the perception of humans towards the robot is an important factor in trust, and we have evaluated such changes in perception [21], this paper focuses on the evaluation of trust alone. Section 3 presents our human–robot interaction scenario, followed by the design description of our robot as a team player. Section 4 discusses the *user study* in detail, as well as the experimental design and the measurement of dependent variables. Section 5 presents the results in detail, taking the proposed hypotheses into account. Section 6 presents a discussion and, finally, Section 7 considers the implications of this work for future work from the human–robot interaction community.

2. Related Work

For decades, trust has been studied in a variety of ways (i.e., interpersonal trust and trust in automation). However, in human–robot interaction, there is much space to study the trust that humans attribute to robots. There have been a growing number of investigations into and empirical explorations of the different factors that affect humans' trust in robots [22,23]. Hancock et. al. [4] reported on 29 empirical studies, and developed a triadic model of trust as a foundation providing a greater understanding of the different factors that facilitate the development of humans' trust in robots. The model's three groupings of factors are first, robot-related factors (anthropomorphism, performance and behaviour), second, environmental-related factors (task- and team-related factors) [4] and third, human-related factors (i.e., demographic attributes of humans) [1]. Robot-related factors [4], especially robot-performance-based factors, influence humans' trust most dramatically. These robot-performance-based factors include a robot's functional ability [24], and robot etiquette (i.e., remaining attentive of errors) [25,26], especially how

the robot casts blame [2], and its reliability and safety [5]. How a robot address the significance of errors, and what feedback humans recieve from error-prone robots radically influences humans' trust [5]. However, in situations where the robot's low reliability was clearly evident, even from the early stages of interaction, human participants continued to follow the robot's instructions [5].

Most of the previous investigations regarding the influence of the explanations provide by an artificial system on humans' trust have been conducted in rule-based systems [10], intelligent tutoring systems [27], intelligent systems (i.e., neural networks, case-based reasoning systems, heuristic expert systems) [8] and knowledge-based systems [28]. Intelligent tutoring systems try to convey knowledge of a particular subject to a learning person. Nevertheless, intelligent tutoring systems cannot clarify their behaviour and remain restricted to particular tasks [29]. Expert systems [30] are systems that recommend answers to problems (i.e., financial decisions, industrial procedure investigations). The corresponding problems usually require a skilled human to solve them [7]. The rule-based expert system *Mycin* [31] was the first expert system to provide explanations for its reasoning in response to *Why*, *Why-Not* and *How-To* queries, but the comparative benefits of these explanations were limited [8,32]. Since *Mycin* was incapable of justifying its advice, it was observed that physicians were reluctant to use it in practice [33]. Earlier work [34] confirmed that different types of explanations not only improved the effectiveness of context-aware intelligent systems but also contributed to stronger feelings of human trust. Although the main focus was on the influence of the *How-to*, *What-if*, *Why* and *Why-not* explanations, the results showed that *Why* and *Why-not* explanations were an excellent type of explanation, It was shown [34] that these two types of explanation effectively helped to improve the overall understandability of the system. There has been little empirical evaluation of the impact of explanations on human–machine trust [11]. Dzindolet et. al. [12] explored manually crafted (and pre-recorded) explanations. Hand-crafted explanations have been shown to be effective in providing transparency and improved trust. However, since hand-crafted explanations are static and created manually, they fail to transfer the complexity of the decision-making process to humans. Nothdurft et. al. [35,36] focused on the transparency of and the justification for decisions made in human–computer interaction. Glass et. al. [37] studied trust issues in technical systems, analysing the features that may change the level of human trust in adaptive agents. They claim that designers should "supply the user of a system with access to information about the internal workings of the system", but the evidence to substantiate such a claim is limited.

The systems, as mentioned earlier, deliberately focused on the use of explanations to convey conceptual knowledge and increase the acceptability of these systems, such as the reliability and accuracy of the system's performance. However, the problem of the non-cooperative behaviour and trust of humans regarding robots remains largely unexplored. To the best of our knowledge, there is still a gap in the current human–robot interaction literature, and there is very little experimental verification that could show that explanations promote or affect humans' trust in robots. The systems mentioned earlier deliberately focused on reliability and accuracy, followed by explanations that convey their conceptual knowledge and increase acceptability.

In addition to the physical appearance of a robot, human perception of the robot's attributes can also affect trust [2]. For example, prior to interacting with a robot, humans develop a mental model of the expected functional and behavioural abilities of the robot. Nonetheless, the human's mental model evolves after interaction with the robot. A mismatch between the human's initial mental model and the later mental model can have a detrimental effect on the human's trust [38]. A human's mental model of a robot's functional and behavioural abilities also defines the human's intentions for future use of robot [8]. Therefore, explanations are valuable because explanations can shape the humans' mental model. Finally, we suggest that our approach (enabling a robot to provide explanations for *transparency* and *justification* of its decisions) is to be considered as the robot's functional ability, which should be categorised as a robot-related factor.

3. Human–Robot Interactive Scenario

Our human-robot interactive scenario is focused on a block-type game known as Spanish *Domino*. The match takes place between two teams, with two players in each team, and it consists of several *hands*; in each *hand*, each of the four player receives seven random domino tiles. Game players take their turn, moving clockwise, and aim for their pair to contain the first player to release all their *tiles*. The tiles in a *hand* are confidential to all but their owner. Thus, the decisions a player makes are made with partial information. During their turn, a game player can perform only two actions:

1. To release a tile (by putting down a tile with an endpoint matching one of the open ends of the current board);
2. To *pass* (because releasing a tile is impossible).

The hand ends when no player can play a domino tile or when a player releases all the domino tiles in their hand. *Domino* is a non-deterministic game, because of the random shuffling and dealing of tiles to the four players at the beginning of every hand. This initial shuffling is an element of non-determinism, but after each player receives their *hand*, all actions are deterministic and successful.

Figure 2 shows the complete set of domino tiles, ranging from (0,0) to (6,6), as used in the study. Because each tile is different, all players have different resources, and team members must cooperate without full knowledge of their partners' resources or the opposing teams' resources. During the match, the robot's behaviour is completely autonomous.

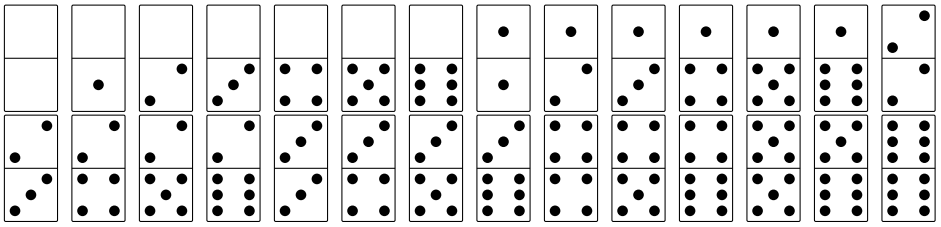

Figure 2. Representation of the 28 tiles of *Domino*.

Figure 3 shows the global architecture and the modules involved in our software for human–robot interaction [39]. Naoqi is the middleware provided by the robot's manufacturer. Our first module is the *Knowledge Base*, which records the rules of the game as well as applying Bayesian Inference to update the belief regarding which tile may be held by which player.

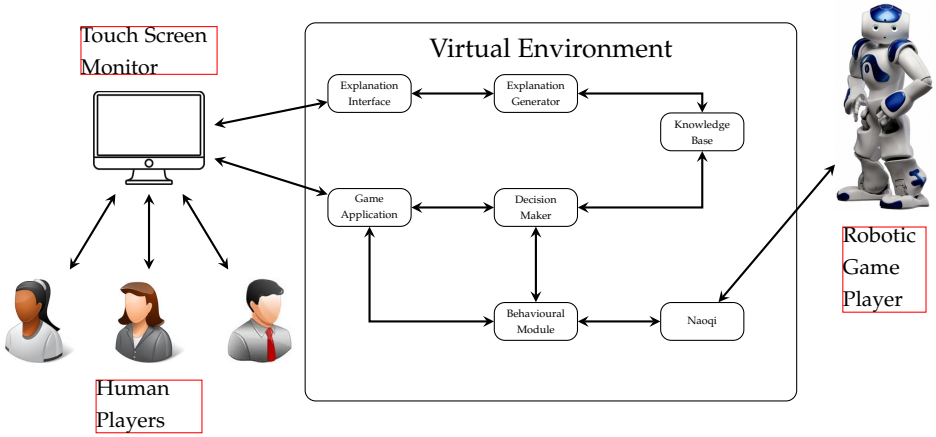

Figure 3. Complete architectural overview of our human–robot interaction scenario using components.

If we number the players as P_i (for $i \in \{0, 1, 2, 3\}$), then P_i is in the same team as P_j, if and only if, $i \cong j \bmod 2$. Each player holds $H_{P_i(u,v)}$, which is the probability that Player P_i was supplied (u, v) initially. We can assume that, initially, $H_{P_i(u,v)} = 1/4$, and as a player

P_i reveals its own hand, for instance, a tile (v_0, u_0), the value $H_{P_i(u_0,v_0)} = 1$ (I hold the tile) and $H_{P_j(u_0,v_0)} = 0$ (the others do not hold the tile) for $j \neq i$. Bayesian inference updates $H_{P_j(u,v)} = 1/3$ for any other tile that P_i does not hold (and $j \neq i$). Thus, Bayesian inference updates this array $H_{P_i(u,v)}$, which is individual and different for each robot; it is their partial information, and we could play games where all participants are robotic. Another part of the knowledge base is common knowledge and consists of the history of moves with the train of tiles (the commonly known board) that defines the current state of the game. After a player's turn, the new knowledge is based on observation. *Bayesian inference* is an effective way to deal with such a revision of belief and update the probabilistic belief $H_{P_i(u,v)}$ of the hand of the partner and their opponents. The *Decision Maker* is a module that approximates perfect play. It rounds the probabilities in $H_{P_i(u,v)}$ to zero or one (it simulates that hands of all players became known) and, in this approximation game with full information, uses the Min–Max algorithm with Alpha-beta pruning to decide on a move. The *Behaviour Module* controls the robot actions and behaviour according to game events, and communicates the speech and moves as well as capturing the moves of other players. The *Explanation Generator* uses a record of the reasoning traces behind a goal tree, to answer questions about the decisions and generate dynamic explanations. Usually, the explanation contrasts the available choices at the time the decision was made. The *Game Application* displays the game on a digital table (a horizontal touch monitor). Information becomes available to all players each time a player completes their turn, either by releasing a tile, or *passing*.

We developed the explanation-generation mechanism on top of the game-playing mechanism. We enable the robot to generate multiple *Static* and *Dynamic* explanations. Static explanations are pre-defined, but the agent still has to choose the relevance of the explanation to a situation; alternatively, dynamic explanations are generated [40] (see Table 1 for examples of these types of explanations). The *Static* explanations are based on (1) *history and facts* about the game, (2) *rules of the game* and (3) *game-play tips*. *Dynamic* explanations are "reactive" explanations that provide insight into the robot's previous decisions (actions), and are incorporated into the *justification* and *transparency* goals of an explanation. The *justification* explanations ("why"-explanations) are more about logical argument, and are the most natural and straightforward way of telling the human why a decision is correct, but these *justification* explanations do not necessarily aim to explain the actual decision-making process. Hence, another type of explanation is required to offer transparency regarding how did the decision was made; these are "how"-explanations. In this sense, the goal of explanations shifts from *justification* to *transparency*. This reflects that these two goals work together, and both serve the purpose of making a decision clear and understandable. Hence, these *Dynamic* explanations are suitable for answering *how-type*, and *why-type* questions. Finally, the robot can elicit the action outcome of its decision process by using a straight-forward "what"-explanation.

Dynamic explanations provide team members with the *transparency* regarding the different factors involved in the decision-making process of the robot. We focused on the intentional robot behaviour [41], towards which humans tend to acquire an intentional perspective, and the construction of explanations in terms of (human-like) reasons, i.e., intentions. Further, our approach to explanations for *Dynamic* explanations is more like a *contrastive approach*, which highlighted the intended *reason* of an event (decision-made) relative to some other event that did not occur. Providing human participants with the main reasoning behind the robot's decisions was the easiest way to inform them about the reliability of the robot's decisions. The robot's decisions were its actions, and those actions were based on the strategies that it applied to the information it possessed, in order to make a decision. Hence, the strategies included a causal structure composed of two steps: "*I [played double] because I [intended to block] my opponent who was [lack of tiles] of Suit 3*."

Table 1. Examples of *Static* Explanations in contrast to *dynamic* explanations. *dynamic* explanations require a reference to a recently played game. *Static* Explanations Refer to the Rules of the Game.

Examples	
***Static* Explanations (Rules of the Game)**	***Dynamic* Explanations (Decision Making)**
A player can make a pass only if the player cannot release a tile on the board.	I played (6,6) because it is a basic rule that, during the first game, the player holding the tile with the highest pips (dots on the tile) goes first.
The legal move is to play a piece in such a way that one of its numbers matches a number on one of the *ends* of the current board position.	I played double because doubles have the same suit value on either end, which provides fewer opportunities to set them down on the board. Therefore, it is a good idea to play them as soon as possible. It is too easy to get stuck with doubles.
The game will end with one of two possibilities. If a player plays his last tile on the board and is left with none, he is deemed the winner. Otherwi the game is blocked because neither player can play.	I tried to play strategically and compete as much as I could. To play the tile (1,4) was the best move based on the analysis, although it contained fewer points. However, it was not enough at the end to win the game.
	I played the tile (1,3) because I had only one choice to play.
	My strategy was to reduce the number of points the opposing team can win with. Therefore, getting rid of the highest-scoring tile was my strategic move.

4. User Study

Using the human–robot interaction scenario described in Section 3, we conducted a *user study*. Our primary objective was to establish an environment in which a social robot interaction with explanations to human players would shed some light on the hypothesis detailed next.

4.1. Hypotheses

To investigate the effect of the explanations given by a social robot on a human's trust, and how much the explanation influences the human's perception of the robot attributes, during an interactive task, we proposed three hypotheses for the experiment, as given below. We hypothesise that, in a team-based and incomplete-information environment:

Hypothesis H1. *A human participant would appreciate understanding about the (transparency) of the robot's decisions, and receiving informed justifications of the robot's choices communicated in terms of explanations from the robot. Such human inclination will be reflected by the human's attribution of more trust in the robot (Condition-2 and Condition-3) compared to no explanation at all (Condition-1).*

Hypothesis H2. *The tteam performance of human participants under Condition-2 and Condition-3 would be better compared to the performance of human participants under Condition-1.*

4.2. Design and Procedure of the Experiment

We adopted the approach of combining survey(s) with experiments to evaluate the effect of the explanations on the human participants' trust in the robot. We used between-subject and within-subject designs for our user study, in which the human participants play the *Domino* game with a stationery humanoid robot NAO. We conducted the experiment with activities organised in different stages, depending on the conditions we describe in Section 4.2.2. Several other attributes of participants are also captured in the questionnaires, to contrast their influence on trust with respect to the explanations.

4.2.1. Data Collection Approaches

Our experiment creates conditions out of the various activities. Several of the activities collect self-reported measures. However, to obtain behavioural data from our the human participants, we video-recorded the players while they played *Domino* matches. The initial video inspection of the behavioural data followed the techniques commonly used for this type of case study [42,43]. We used The Language Archive' ELAN platform [44] (from the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands) for behavioural coding (verbal and non-verbal actions of the human participants) to examine video material and perform a reliable organised analysis of the videos (no human subjectivity in extracting data from videos).

We also maintained a history at the backend of the system to record the moves played by the human participants in a *csv* file. We also kept the history of the human participants' examination of the robot's explanations.

We used annotations, which were executed to obtain evidence to challenge or support the self-reported data on which our investigations were established. To better recognise qualitative accounts, this method is widely used in ethnography [45]. Indeed, by using this method, we found patterns and relationships, based on which we inspected our observations and illustrated, complemented and presented support for the self-reported data.

4.2.2. Conditions and Activities

Activity 1 Pre-interaction set of questionnaires: This is a data-collection activity, consisting of self-reported and not video-recorded measures. Before starting the experiment, we collected human participants' demographics (i.e., previous experience with robots, initial impression of the robots, pet-ownership), initial perception, and trust towards the robot. In particular, this activity included the following sub-activity.

- Previous experience of human participants: Prior relationship with non-human agents such as pets [46] influences human interaction with a robot. Thus, to examine other factors such as prior experience with robots, we evaluated human participants' demographical information with the following questions.
  - Do you have any prior physical experience with a robot?
  - Have you ever watched a television show or a movie that involves robots?
  - Do you have any prior relationships with non-human agents such as pets [46]?

  Trust between humans and animals may be a suitable analogy for trust between humans and robots [2]. Examining the nature of a human–animal relationship can help to increase understanding of how a human interacts with, and trusts, a robot [24].

Trust is a dynamic attitude that changes over time [1,3]. To elucidate the changes in trust and perception of the robot's attributes, the human participants filled out questionnaires on their trust in, as well as their perception and impression of, the robots. In particular, the instruments used during this activity are listed here as informative of a corresponding dependent variable.

1. Dependent variable **Trust:** By using a 14-item subscale of Human–Robot Trust (HRT) questionnaire [1] we measure human teammates' trust, although this is not directly observable [47]. This questionnaire is also completed at the end of all interactions; see Activity 8.

Activity 2 Robot provides verbal static explanations (see Figure 1b): The robot provided verbal *Static* explanations of how to play the game to the human participants.

> ***"We will play the block-type game of Domino with double-six set of domino tiles. There are 28 tiles in the set ranging from (0,0) to (6, 6). There are four***

players in the game and each player will initially receive a set of seven random tiles. . .!"

Activity 3 A homogeneous team plays match facing a heterogeneous team: We say a team is homogeneous when both players in the team are humans. A team is heterogeneous when one participant is the robot and the other is a human. The match is played to 50 points, which usually means five hands (although there could be less or more).

Activity 4 Pre-explanation session questionnaire: Participants respond to questionnaire before receiving dynamic explanations from the robot. We aim to collect data to scrutinise the human participants' mental model of the robot's behaviour (the understandability and credibility of the robot's decisions); thus, we asked the following two questions

1. I was nervous with the robot's decisions during the game (by answering yes or no);
2. The robot's use of strategy (decisions) was correct, and helped me during the game (by answering yes or no or not sure).

For *Condition-2* and *Condition-3* participants, we evaluated this twice: during this activity and after the *explanation-session* in Activity 5. However, for participants under *Condition-1*, since they did not receive explanations, we applied these questions after the first and third matches.

Activity 5 Post-explanation session questionnaire: Participants respond to questions after receiving dynamic explanations from the robot (see Activity 4).

Activity 6 Robot gives static explanations about the robot's decision: The robot makes observations about the recently played hands.

Activity 7 Robot gives dynamic explanations about the robot's decision (see Figure 1c): The robot debates the recently played games, providing information about the reasoning behind its decisions.

Activity 8 Post-interaction session questionnaire: Participants respond to questions that review the entire experience. The HRT questionnaire and the Godspeed questionnaire [48,49], which were applied before the experience, are applied here (see Activity 1). Similarly, the rating of images of robots is repeated here.

However, to evaluate the effect of explanations (only to those participants that received them, thus participants under *Condition-2* and *Condition-3* see below), we added five questions in which human participants were required to rate their *satisfaction-level* and the *understandability* of the robot's decisions, using explanations on a five-point Likert scale.

1. **Satisfaction:**
   - Explanations were clear and understandable (aimed at measuring the clarity of the robot's explanations);
   - Explanations helped me during the game (aimed at the usability of the robot's explanations).
2. **Understandability:**
   - Explanations helped me to understand the decisions of the robot;
   - The robot adequately justified its behaviour (decisions) through explanations;
   - I can now understand the robot's behaviour better.

Activity 9 These activity does not involve the participants. It consists of recording and analysing the player's decisions in all the matches (and under all conditions). The goal is to determine whether the robot's explanations improved the human participants' performance and decision-making in the game. We compared the task performance of human participants under three conditions (see below) based on the wins and losses of two teams (i.e., human–human and human–robot teams). In particular,

we examined the decision-making of human participants and their use of strategy in the games. A difference between the performances of the teams between the three conditions will suggest to us that the explanations have an impact on human participants' learning and performance improvement during the game.

There were different groups of human participants who experienced different variants of the same experiment that were controlled by the independent variable "explanations". In particular, the study had a total of three conditions:

- **Condition-1**: is our *Control Condition*. This consisted of the following activities, in this exact order: Activity 1, three iterations of Activity 3, and Activity 8. There were no explanations given by the robot, but more matches (Activity 3), with the expectation that we could compensate for learning with practice playing the game. That is, we expect that more hands, and more practice with the task, would not improve humans' performance as much as the advice received through the robot's explanations;
- **Condition-2**: Both types of explanation are provided. The activities and their order are Activity 1, Activity 2, Activity 3, Activity 7, Activity 2, Activity 3, and Activity 8;
- **Condition-3**: We remove static explanations and use only dynamic explanations. Thus, the activities and their order are Activity 1, Activity 3, Activity 4, Activity 7, Activity 5, Activity 3, and Activity 8.

For participants under *Condition-2* and *Condition-3* (those receiving explanations), the second game-playing session aims to observe the impact of the *decision-transparency*, and *justification* of the robot's decisions in improving the task performance of a team.

4.3. Nature of the Participants

We conducted the study in Griffith University Australia, and there were a total of 63 human participants, with ages ranging from 19 to 42 years old ($M = 30.60 \pm 7.08$). Out of the 63 participants, 42 played the game with a human partner and 21 played with the robot. These distributions aimed to collect a valid number of answers from the robot's partners. Additionally, out of the 63 subjects that revealed their gender, 25 were females and 38 were males. We recruited human participants through general advertising, using posters on a university notice board, and communicating directly with students. As many as 41.82% of human participants were students and involved professionally in a science, technology, engineering and mathematics subjects, and the rest were evenly distributed between employed and unemployed. Each human participant received an invitation letter with the main objective of conducting the experiment. Along with the invitation letter, we also attached a brochure with a brief description of the *Domino* (briefly describing the rules and mechanics of the game). Before taking part in the experiment, all human participants provided their consent. We offered a 10 AUD gift card as a *token of appreciation* to every human participant. There were five groups in *Condition-1*, 11 groups in *Condition-2*, and five groups in *Condition-3*. Each group in *Condition-1* played three matches with the robot; however, in *Condition-2*, and *Condition-3* each group played two matches, where a single match consisted of five hands in total, or until a pre-defined score was reached. Each human participant selected his/her team member in a draw. In addition, in each condition, before starting the formal game-play, the robot greeted and introduced itself to the human participants, saying, "Hello human! I am LAVA, today we are going to play *Domino* in teams."

Since the majority of human participants classified themselves new to the *Domino* game, a sequenced presentation reviewed the brochure presenting the case study. We included a description of the game on the user interface (i.e., on the touch screen), until everyone felt confident with the game's rules and the operation of the interface. Once participants confirmed that they felt competent with the rules and the interface, we gave the human participants a questionnaire based on the rules of the game to assess their initial level of understanding. After the assessment, the human participants started the first game-play session. We expected that all human participants would start with the same common-sense model of the task (i.e., *Domino* game), which also helped us to estimate

what knowledge the human participants possessed about the task. At the end of each session, the robot told the human participants that the interaction had ended, and thanked the human participants, saying, "Thank you for your participation. Hope you enjoyed playing with me."

5. Experiment Results

We investigated (1) the effect of the robot's explanations on the humans' level of trust, and (2) how much effective explanations were at changing humans' perception of the robot attributes based on self-reported, quantitative questionnaire data and objectively measured behavioural data. We obtained a series of results regarding the second point, the human's perception of the robot's attributes. As discussed in Section 2, the perception of robot's attributes is a factor in trust, and our results show that explanations impact attributes such as Animacy, Likeability, Perceived Safety, Perceived Intelligence, and Anthropomorphism. However, in this paper, we detail only the results on trust.

5.1. Self Reported Data Results

Prior to conducting any analysis, we performed a reliability analysis (Cronbach's α) to assess the internal reliability (an $\alpha > 0.7$ or higher is considered acceptable, indicating the reliability of the measuring scale) of the questionnaires. Cronbach's α for the HRT questionnaire [1] was $\alpha = 0.809$, and for the Godspeed Questionnaire [48,49], it was $\alpha = 0.881$.

5.2. Effect of Condition on the Dependent Variable Trust

We performed a normality analysis using the *Shapiro–Wilk Test*, to examine whether the dependent variable "Trust" followed a normal distribution under *Condition-1*, *Condition-2*, and *Condition-3*. The test reported a non-normal distribution for all three conditions. Hence, we performed a within-subject non-parametric *Wilcoxon Signed-Rank Test* to analyse the overall effect of the condition on human participants' trust in the robot and a between-subject non-parametric *Kruskal–Wallis H Test* to analyse the overall effect of the condition on the human participants' trust in the robot, controlling for the trust level reported before interaction with the robot.

Table 2 shows the results comparing the trust of human participants before and after interacting with the robot for *Condition-1* and *Condition-2*.

Table 2. Participants self-assessed dependable variable **Trust** before and after the interaction per condition (using within-subject *Wilcoxon Signed-Rank Test*).

Condition	Value before Interaction	Value after Interaction	Is the Difference Statistically Significant?
Condition-1	$M = 50.26 \pm 4.97$	$M = 69.93 \pm 7.95$	YES ($Z = -3.411$, $p = 0.001$)
Condition-2	$M = 52.24 \pm 9.44$	$M = 89.27 \pm 6.44$	YES ($Z = -5.014$, $p < 0.001$)

For *Condition-3*, we conducted a simple impact analysis by looking at the influence of the robot's explanations' of its *decision-transparency* and *justification* on the trust of the human participants. We used the non-parametric *Friedman Test* with repeated measures because of the violation of *Mauchly's Test of Sphericity*, and incorporated a Bonferroni correction to the significant findings.

Table 3 shows the results comparing the trust of human participants before interaction, after first interaction, and after the *Explanation-session* for *Condition-3*.

Table 3. Participants self assessed dependable variable Trust Before and after the Interaction for *Condition 3* (Using Within-Subject *Friedman Test* with Repeated Measures).

Condition	Value before Interaction	Value after First Interaction (By First Interaction, We Mean after First Match.)	Value after Explanation-Session	Is the Difference Statistically Significant?
Condition-3	$M = 54.03 \pm 3.57$	$M = 64.73 \pm 1.32$	$M = 77.06 \pm 2.87$	YES (*Chi-Square* = 30, $p = 0.01$)

5.2.1. Are the Changes in Trust, from Before to after the Experience of Interacting with Robot, Different Across Conditions?

The between-subject non-parametric *Kruskal–Wallis H Test* showed a highly significant result (Chi-Square = 37.54, $p < 0.01$) across the three conditions of human participants. We also ran a series of non-parametric, between-subject *Mann–Whitney U Test* for paired samples to compare the trust levels of the group of human participants under three conditions, and the results are summarised in Table 4, which shows the pair-wise results. Hence, our results suggested that there were statistically significant differences in the trust levels of all three groups after interacting with the robot in different settings.

Table 4. Results of non-parametric between-subject *Mann–Whitney U Test* across pair-wise comparison between *Condition-1 (Control condition)*, *Condition 2* and *Condition 3*.

Condition	Contrast with	Test Performed	*Is the Difference Statistically Significant?*
Condition-1 (Control condition)	*Condition-2*	non-parametric between-subject *Mann–Whitney U Test (This caused inflation inthe type 1 error rate, that we controlled using a Bonferroni adjustment.)*	YES ($U = 127$, $p < 0.01$)
Condition-1 (Control condition)	*Condition-3*	non-parametric between-subject *Mann–Whitney U Test*	YES ($U = 88$, $p < 0.01$)

5.2.2. Previous Experience of Human Participants on Trust

We divided the human participants into two sub-groups in each condition, based on (1) the participants that had previous experience with robots, and the ones without previous experience with robots, (2) the participants that had pet-ownership and the ones without pet-ownership. We examined the influence of previous experience with robots, and pet-ownership on the human participants' trust in the robot. We performed a between-subject *Mann–Whitney U Test* to examine the trust of human participants with and without previous experience with robots, and with and without pet-ownership. Table 5 reflects that the human participants' previous experience with robots and pet-ownership did not seem to influence their trust in the robot.

Table 5. Influence of participants' previous experience with robots, and pet-ownership on the Trust (using between-subject *Mann–Whitney U Test*).

Condition	Previous Experience with Robots		Pet-Ownership	
Condition-1	$U = 21$, $p > 0.05$	(non-significant)	$U = 19$, $p > 0.05$	(non-significant)
Condition-2	$U = 97$, $p > 0.05$	(non-significant)	$U = 85$, $p > 0.05$	(non-significant)
Condition-3	$U = 16$, $p > 0.05$	(non-significant)	$U = 13$, $p > 0.05$	(non-significant)

5.3. Are the Trust Levels Impacted by the Partner (Human or Robot)?

The constitution of a partnership (a team) also allows us to consider the two types of group. We divided the human participants into two types of group: those in a homogeneous team, that is, a human–human partnership, and those human participants that had a robotic partner (a heterogeneous team with a human–robot partnership). Table 6 shows the results when using the statistical test *Welch's T-Test* (unequal variances) with sub-group as a factor, and post-interaction trust as dependent variable. We found that the humans who partnered with a robot showed slightly lower levels of trust towards the robot than human adversaries in *Condition-1*, and *Condition-3*, but not in *Condition-2*. Overall, in *Condition-2* and *Condition-3*, we found significant differences in trust levels between different *Domino* players (i.e., team partner, adversary), indicating the crucial effect of the robot's explanations on trust.

Table 6. Summary of results contrasting the human versus robot partner.

Condition	Team Structure	# of Participants	Trust metric Mean	σ	*Welch's T-Test* F	p	
1	Human Partner	10	80.8	4.41	0.810	$p > 0.05$	not significant
	Robot Partner	5	71.6	4.56			
2	Human Partner	22	70.54	6.10	5.91	$p > 0.02$	significant
	Robot Partner	11	76.72	6.85			
3	Human Partner	10	70.59	1.10	21.22	$p > 0.02$	significant
	Robot Partner	5	69.48	3.921			

5.4. Human Participants' Assessment of the Robot's Behaviour

We did not inform the human participants about the mechanism guiding the robot's behaviour, to avoid bias in the explanations (*Condition-2* and *Condition-3*), and bias towards the perceived mental model of the robot's behaviour in different settings. Table 7 shows the total responses per condition to mental model questions.

In *Condition-1*, we marked the human participants' responses to the mental model questions concerning the robot's behaviour after the first match and the third match. We performed non-parametric, within-subject *Wilcoxon signed Ranks* test to investigate participants' mental model of the robot's behaviour as a consequence of "continuous interaction". The *Wilcoxon signed Ranks* test did not reveal a significant difference ($Z = -0.707$, $p > 0.05$) for the statement, *"I was nervous with the robot's decisions during the game"*, and ($Z = -0.365$, $p > 0.05$) or for the statement, *"The robot's use of strategy (decisions) was correct, and helped me during the game"*. Furthermore, we performed within-subject *Pearson's Chi-Square* test to investigate the relationship between the two questions (recorded after the first and the third match). The test revealed a significant association between the responses of the human participants in *Condition-1* for the mental model questions recorded after the first match ($X^2 = 5.455$, $p = 0.05$), and after the third match ($X^2 = 3.743$, $p = 0.05$).

In *Condition-2*, we recorded the human participants' responses to the mental model questions about the robot's behaviour before and after explanations. The non-parametric, within-subject *Wilcoxon signed Ranks* test revealed significant differences ($Z = -3.900$, $p < 0.01$) for the statement, *"I was nervous with the robot's decisions during the game"*. However, we found no significant difference ($Z = -1.33$, $p > 0.05$) for the statement, *"The robot's use of strategy (decisions) was correct, and helped me during the game"*. Furthermore, we performed within-subject *Pearson's Chi-Square* test to investigate association between the two questions (recorded before explanations and after explanations). The test revealed a non-significant association between the responses of the human participants in *Condition-2* for the pre-explanation mental model questions ($X^2 = 0.471$, $p > 0.05$), but we found a significant association for the post-explanation mental model questions ($X^2 = 5.238$, $p = 0.05$).

In *Condition-3*, we recorded the human participants' responses to mental model questions about the robot's behaviour before and after explanations. The non-parametric, within-subject *Wilcoxon signed Ranks* test revealed significant differences ($Z = -1.897$, $p = 0.05$) for the statement, *"I was nervous with the robot's decisions during the game"*, and ($Z = -1.61$, $p = 0.03$) for the statement,

"The robot's use of strategy (decisions) was correct, and helped me during the game". Furthermore, we performed a within-subject *Pearson's Chi-Square* test to investigate the association between the two questions (recorded before explanations and after explanations). The test revealed a non-significant association between the responses of the human participants in *Condition-3* for the pre-explanation mental model questions ($X^2 = 3.750$, $p > 0.05$), but we found a significant association for the post-explanation mental model questions ($X^2 = -0.784$, $p = 0.02$).

Table 7. Distribution of human participants' responses to mental model questions about the robot's behaviour (*Condition-1*: 15 participants, *Condition-2*: 33 participants, and *Condition-3*: 15 participants).

Condition		Total Number of Participants' Responses	Questions				
			I Was Nervous with the Robot's Decisions during the Task Possible Answers		The Robot's Use of Strategy Was Correct, and Helped Me during the Game Possible Answers		
			Yes	No	Yes	No	Not Sure
1	after firstmatch	15	11	4	3	5	7
	after thirdmatch		9	6	4	2	9
2	before explanations	33	22	11	21	7	5
	after explanations		5	28	27	3	3
3	before explanations	15	8	7	5	2	8
	after explanations		2	13	12	0	3

Furthermore, we performed a between-subject *Kruskal–Wallis H Test* (non-parametric ANOVA on ranks) to determine if there were statistically significant differences between the group of participants in different settings. We found significant differences between the three groups (Chi-Square = 19.03, $p < 0.01$) for the statement, *"I was nervous with the robot's decisions during the game"*, and (Chi-Square= 17.64, $p < 0.01$) for the statement, *"The robot's use of strategy (decisions) was correct, and helped me during the game"*.

5.5. Analysis of the Effect of Explanations Using Questionnaires

In *Condition-2* and *Condition-3*, we examined the effect of explanations on human participants' assessment of the robot's behaviour based on their responses to the questionnaires. We divided the explanations' questionnaires into two categories, i.e., category 1: level of satisfaction with the explanations (questions a and b), and category 2: behaviour understandability of the robot (questions c,d,e). The between-subject *Mann–Whitney Test* (non-parametric) was used to determine if there were statistically significant differences between the group of human participants in *Condition-2* and *Condition-3*, based on their *satisfaction-level* with the explanations, and the *understandability-level* of the explanations. The analysis captured the conditions on the grouping variable list, and the mean score of *satisfaction-level* and *understandability-level* questionnaires on the test variable list.

The non-parametric, between-subject *Mann–Whitney Test* reported statistically significant differences based on the explanation assessment questionnaires between the groups of human participants in two conditions ($U = 156.50$, $p = 0.02$) for *satisfaction-level*, and ($U = 20.50$, $p < 0.01$) for *understandability-level* questionnaires between the two conditions. Furthermore, we calculated the *Spearman's* correlations between the mean score of mental model questions (post explanations), and the mean score of explanations' questions. We found a strong positive correlation ($r_s = 0.245$*, $p = 0.05$) between the two dependent variables in *Condition-2*, and ($r_s = 0.599$*, $p = 0.01$) between the two dependent variables in *Condition-3*.

5.5.1. Do the Game Results Impact the Trust Levels across Conditions

We ran a Two-Way multivariate analysis of variance (MANOVA) on our data to investigate if the game's results impacted the trust levels. Since the trust metric and game result statistic violatied the assumption of multivariate normality, for the multivariate

test, we selected the *Pillai's Trace* and for the Post-Hoc tests, we selected the *Scheffe* test (equal variances not assumed). After applying a rank transformation to the data, we used the game result, and game teams (human–human, human–robot) as fixed factors, and post-interaction trust and conditions on the dependent variable list. The *Pillai's Trace* reported a significant value for the effect of game results on trust ($F = 9.519$, $p < 0.001$), and for the condition effect on trust ($F = 3.600$, $p = 0.009$). However, the interaction effect of both game result, and condition proved to be non-significant, with ($F = 1.468$, $p > 0.05$). Furthermore, the test of between-subject effect reported non-significant findings for game results and trust ($F = 1.423$, $p > 0.05$), and between condition and game teams ($F = 1.452$, $p > 0.05$). However, we found a significant impact between game results and game teams ($F = 1.953$, $p < 0.001$), and between condition and trust ($F = 7.119$, $p < 0.001$) (which has been shown before). Furthermore, we did not find a statistically significant result for the interaction effect of game results and condition on trust ($F = 0.996$, $p > 0.05$), and on game teams ($F = 1.447$, $p > 0.05$). Furthermore, the Post-Hoc tests *Scheffe* could not reject the null hypothesis with ($p > 0.05$), indicating a non-significant impact on trust. However, based on the *Pillai's Trace*, we found significant differences in the levels of trust based on different game results.

5.6. Effect of Explanations on Team Performance

We recorded the players during each match, which consisted of several games, and the game statistics are given in the Table 8 for *Condition 2* and *Condition 3*. We compared the performance of the teams in the three conditions, to see whether receiving an explanation or not receiving an explanation impacted the performance of the teams. We also examined the usability of these explanations for participants during the games, in terms of their use of strategies and the selection of their next move (*Condition-2* and *Condition-3*). Due to the introductory brochure, our expectation was that all human participants had a similar starting playing ability, and approached the matches with the same common-sense model. We also evaluated participants' basic understanding of the game based on the assessment questionnaire containing basic questions regarding the game; all the participants provided correct answers in the questionnaire.

Table 8. Game statistics: What type of team obtained what result (*Condition 2* and *Condition 3*).

Condition 2				**Condition 3**			
33 human participants (11 groups)				15 human participants (5 groups)			
Teams	Before explanation session	After explanation session	Total Number of Games	Teams	Before explanation session	After explanation session	Total Number of Games
Human-Human Team Wins	11	14	110	Human-Human Team Wins	5	7	50
Human-Robot Team Wins	29	34		Human-Robot Team Wins	14	14	
Game Stuck	15	7		Game Stuck	6	4	

Table 9 shows the performance of each team in three matches under *Condition-1*, where human participants were not provided with explanations, and they only knew the *rules* of the game. The same teammates and adversaries played repeated hands in each group, but we did not see any significant improvement in the three matches. Further, the decrease in the number of games won in the second match also shows that the human participants' udnerstanding of the game was not enhanced. Conversely, the performance of teams

under *Condition-2* and *Condition-3* was better, compared to *Condition-1*, which showed the impact of the explanations. The comparison between both teams' performances in the first match (before the *explanation-session*) with that of the second match (after the *explanation-session*) also demonstrated that the performance of teams in both conditions improved significantly, which was noticeable in the second match, played after the *explanation-session*. The difference between the teams in both conditions is quite straightforward in the sense that, under *Condition-3*, participants were only provided with *Dynamic* explanations, hence their improvement in the game is based on learning from the robot's decisions, communicated by way of explanation. Conversely, under *Condition-2*, human participants were not only provided with the *Dynamic* explanations, they were also exposed to the *Static* explanations augmented by the *rules of the game*, and *game-play tips*, as well as examining the robot's decisions. We noticed that many participants during the *explanation-session* revised the rules, and checked different tips. Hence, these aspects also played a role in their improved performance in the matches. Hence, the performance of teams was better in *Condition-2* and *Condition-3* compared to *Condition-1*.

Table 9. *Condition-1*: performance comparison of both teams in each match (human participants were not provided with explanations).

Match	Human–Human Team	Human–Robot Team	Game Stuck
First Match	8	13	4
Second Match	7	14	4
Third Match	8	16	1

By evaluating the moves of the players, stored in our records, we observed the implicit trust between human partners in a team. The records also showed moves where human participants exhibited cooperation, and sacrifices for their robotic partner. Furthermore, under *Condition-2* and *Condition-3*, we observed that human participants considered playing random tiles in the first match. After the *explanation session*, they used some of the strategies demonstrated, i.e., they preferred to play tiles with the highest points and place doubles on the board during the early stages of the hand.

We performed between-subject, one-way MANOVA to test if there were any differences in the group of human participants under *Condition-1*, *Condition-2*, and *Condition-3*, regarding the linear combination of the two dependent variables, i.e., game results and strategy used. We considered two strategies (1) playing tiles with doubles, and (2) playing tiles with the highest points. Before performing *One-Way MANOVA*, we performed certain assumption tests, which disclosed the violation of the assumption of multivariate normality. The assumption of multivariate normality was violated as game results had the value of the significance level of the *Shapiro–Wilk Test* $p < 0.05$, reflecting the non-normal distribution of the variable. Hence, for the multivariate test, we selected the *Pillai's Trace*, which reported that the three groups, under *Condition-1*, *Condition-2*, and *Condition-3*, were different across the levels of the independent variable. In a linear combination of the two dependent variables (game results and strategy used), the value obtained was ($F = 8.630$, $p < 0.001$). We also interpreted the *Levene's Test* of equality of variances , which reported non-significant results ($p > 0.05$). Hence, we performed the tests of between-subject effects that reported statistically significant findings ($F = 24.057$, $p < 0.01$) for the strategy used. However, we found non-significant findings for game results ($F = 0.455$, $p > 0.05$). The Multivariate test reported the same, statistically significant findings (*Partial Eta Squared* $= 0.223$, $F = 8.630$, $p < 0.01$). Furthermore, the Univariate test reported statistically significant findings for strategy used (*Partial Eta Squared* $= 0.445$, $F = 24.05$, $p < 0.01$), but not for game results (*Partial Eta Squared* $= 0.015$, $F = 0.455$, $p < 0.05$). This shows that we had statistically significant findings for the *Pillai's Trace*, so interpreting the post-hoc test also helped us to interpret the Univariate test results. The post-hoc *Scheffe Test* (equal variances not assumed) reported a statistically significant difference between the group of

human participants under *Condition-1* and *Condition-2* ($p < 0.01$), and under *Condition-1* and *Condition-3* ($p < 0.01$) on the strategy used. Furthermore, we found a non-statistically significant difference between the group of human participants under *Condition-2* and *Condition-3* ($p > 0.01$) on the strategy used. We also found a non-statistically significant difference between the group of human participants under *Condition-1* and *Condition-2*, and *Condition-1* and *Condition-3*, and *Condition-2* and *Condition-3* ($p > 0.05$) on the game results.

Hence, this analysis shows that, with the linear combination of dependent variables (game results and strategy used), all three groups are different. However, for the strategy using a dependent variable, we found differences between the group of human participants under *Condition-1* and *Condition-2*, and *Condition-1* and *Condition-3*, but we found no difference between the group of human participants under *Condition-2* and *Condition-3*. This findings provide us with evidence that, after the *explanation-session*, the use of strategy was improved in *Condition-2*, and *Condition-3*.

6. Discussion

Our study investigated how explanations from a social robot can influence how humans perceive a robotic partner in a task, and thus how they judge the robot's behaviour. We observed evidence that humans showed increased trust levels, and a change in the perception of the robot. With this experimental procedure, we distinguished the impact of interaction with the robot on human's trust and the perception of the robot, without explanations *Condition-1* and with explanations *Condition-2* (assuming game-play and the *explanation-session* as continuous interaction), and *Condition-3* (by distinguishing the separate influence of the explanations).

6.1. Effect of Results on Our Hypotheses

Overall, the difference in the trust levels in different settings suggested that human participants trusted the robot when playing *Domino*, and the trust levels were upgraded after playing the game. However, under *Condition-2* and *Condition-3*, when participants were provided with explanations, their level of trust increased significantly compared to *Condition-1*. Hence, results from our preliminary analysis strongly support our Hypothesis H1. We also observed that the trust levels were also impacted by the game result (Section 5.5.1), which indicated that, in a team-based environment, winning or losing a game destabilised the perception of trust in the robotic game player. This observation also reinforces the consequence of condition on this measure, based on the better team performance in *Condition-2*, and *Condition-3*. Furthermore, the element of partnership (human–human team, human–robot team) allowed us to calculate the effect of every defined measure on humans' trust. Hence, we investigated differences in trust levels between different *Domino* human players (i.e., team partner, adversary), and found significant differences in trust levels under *Condition-2* and *Condition-3*. This findings confirm the profound impact of the robot explanations on trust.

We found that, irrespective of the condition, the participants had a higher degree of trust in the robot after interacting with the robot. However, Figure 4 shows that items measuring the robot's explanatory capability, such as *Providing Feedback*, *Providing Appropriate Information*, and *Communicating with People*, are tightly connected with the outcome. Human participants under *Condition-2*, and under *Condition-3* rated these items very highly when compared to the participants' ratings in *Condition-1*. Furthermore, we did not find a significant difference for the items *Malfunctioning*, and *Have Errors* under *Condition-2* and *Condition-3*, but participants in *Condition-1* gave slightly higher ratings to the *Have Errors* item, reflecting that they were not confident with the robot's decisions. In addition, the human participants also considered the robot less *Reliable*, more *Unresponsive*, and less *Predictable* under *Condition-1*. For robots, the factors *Predictability* and *Dependability* can be taken as a basis of trust. Contrarily, human participants in *Condition-2* and *Condition-3*, trusted the robot, and thereby considered it more *Reliable*, less *Unresponsive*, and more *Predictable*. Furthermore, we found human participants gave high ratings to the item *Dependable* in all three conditions, indicating that the physical behaviour of the robot did not

meet their expectations. Hence, our findings suggest the human participants in *Condition-2* perceived the behaviour of the robot to be more intelligent and *trustworthy* compared to other conditions.

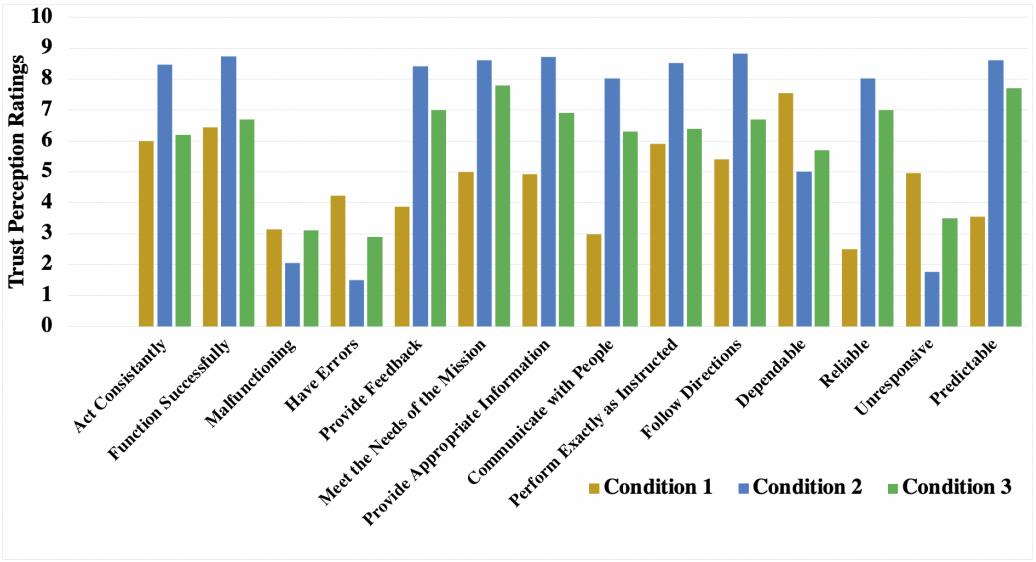

Figure 4. A summary of the quantitative data analysis results for trust.

Although detailed results are not presented here, our experiment enabled us to collect data and evaluate the participants' perception of the robot's attributes linked with trust: our data showed that human participants under the three conditions ranked attributes such as *Animacy*, *Likeability*, *Perceived Intelligence*, and the *Anthropomorphism* differently in ratings of the robot, but not for the *Perceived Safety* attributes of the robot. Thus, before interacting with it, the majority of participants did not know the robot, but, after interacting with it, they perceived it to be more intelligent and, at the same time, much safer.

Furthermore, the results in Section 5.6 show that the performance of teams was better under *Condition-2* and *Condition-3* compared to *Condition-1*, which is in accordance our Hypothesis H2.

Further, we also found no difference between the trust levels of human participants who had previous experience with robots, and rhose who had not interacted previously with robots. Additonally, participants who never had a pet and who had a pet equally perceived the robot as an intelligent and trustworthy partner, showing no difference in levels of trust in the robot.

We also compared human participants' mental models after direct experience with the robot under *Condition-1*, *Condition-2*, and *Condition-3*, based on their assessment of the robot's behaviour. In *Condition-1*, we found a slight change in participants' mental model of the robot's behaviour when examined after the third match, compared with the mental model examined after the first match. We investigated their confidence and satisfaction in the robot's decision making, which was slightly changed as a consequence of continuous interaction. Previous research [50] has also suggested that humans who interacted little with robots initially have a simple mental model. However, with experience, their mental model changes. In *Condition-2* and *Condition-3*, we measured the mental model of participants "after first match", and then examined the effect of explanations on changes in their mental model. We investigated how much a human's score varied via the mental model questionnaires completed after the *explanation-session*. Our results suggested that the impact of "explanations" was significant and valid, and changed participants' mental models when compared to mental models examined before the explanations. Additionally, the explanations also increased participants' trust and satisfaction in the robot's decisions. Although, under *Condition-1*, the mental model of participants also changed, thiis was a consequence of continuous interaction with the robot. Conversely, under *Condition-2* and

Condition-3, the change in the mental model of participants was purely the consequence of "explanations", as participants had interacted with the robot for some time, i.e., only during the first match. That is, participants under *Condition-2* and *Condition-3* had already incorporated the effect of playing three matches.

In general, the primary goal of explanations in human–robot interaction is to enable humans to appreciate and understand a robot's behaviour through an explanation that is given in "human-understandable terms". Previous research [51] has also suggested the use of mechanisms inspired by how humans explain behaviour to make an explanation useful to them, as humans assign similar levels of intentionality to robots as they attribute to other humans. Furthermore, in order to establish a social connection with a social robot, participants should perceive its behaviour as intentional and reasonable. Our results in Section 5.5 corroborated with the idea that participants under *Condition-2* and *Condition-3* confirmed that they not only understood the robot's behaviour through explanations, but also appreciated the content, quality and clarity of the explanations. Explanations changed their mental model and their perception of the robot's behaviour, and also upgraded their level of trust in the robot.

Under *Condition-2*, human participants were provided with *Static* and *Dynamic* explanations, while under *Condition-3*, participants were exposed to *Dynamic* explanations only. We kept a record of the number of times a human participant (partner/adversary) accessed explanations, i.e., *Static* or *Dynamic*, depending on the condition. We found that the human participants (regardless of whether they had a role as team partners or opponents) examined the robot's decisions for longer and in more detail in both of the more elaborate conditions that included explanations. Furthermore, we also observed that, after checking the *Robot Moves*, participants further examined the different *factors* involved in the decision-making process of the robot. The human participants showed less interest in how did the robot gathered the knowledge required to play the game in an incomplete-information environment, and investigated other factors, particularly players' *pass*, more than the robot.

7. Conclusions and Future Work

In this human-oriented world, while interactive robots are still in their emergent phase, initial misunderstandings, failures, and mistakes are likely to arise in human–robot collaborations. The extent to which robots can show human-like attributes, i.e., offer explanations to justify their behaviours, or offer apologies and corrective measures after faulty behaviours, can ameliorate dissatisfaction and increase humans' trust in them. This perspective is also vital to address the question how a robot rebuilds a human's trust after an error, and how successful and effective its attempt to rectify the erroneous situation, or offer an apology to mitigate the dissatisfaction resulting from its unpredictable behaviour, will be. We presented an experiment to further the understanding of how an explanation positively influenced participants' trust. The results suggest that augmenting robots with the ability to offer explanations can make robots more trustworthy.

We analysed individual demographics, i.e., previous experience with robots, and the impact of pet-ownership. For our current experiment, we used participants with a homogeneous profile, i.e., with little or no previous experience with robots, and no previous knowledge or experience with the *Domino* game. In future, we plan to carry out several interesting analyses, for example, the influence of what kind of explanation has a greater impact on the perception and trust that humans have in robots. Additionally, in our follow-up experiments, we aim to examine gender aspects, as well as participants' personality traits, technical background, and education. Furthermore, it would be interesting to investigate the possible influence of cultural background in the perception of and trust in robots. A investigation of how the personality traits of humans influence their perception of the explanations in a collaborative setting is also worthy of follow-up study. Additionally, our results indicated that, to enhance the social presence of a robot, its speech should be combined with additional measures, such as expressive facial features, and a rich set

of social behaviours. However, combining functional reliability with expressiveness is challenging; the design and implementation of a flawless system remains out of reach, as appropriate speech recognition systems are still emergent. Hence, we aim to conduct more research, focusing on specific design metrics of robots, in our future work.

Author Contributions: Conceptualization, V.E.-C.; methodology, M.J. and V.E.-C.; software, M.J. and V.E.-C.; validation, M.J. and V.E.-C.; formal analysis, M.J. and V.E.-C.; investigation, M.J. and V.E.-C.; resources, V.E.-C.; data curation, M.J.; writing—original draft preparation, M.J. and V.E.-C.; writing—review and editing, M.J. and V.E.-C.; supervision, V.E.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Griffith University Protocol Number: 2017/734, approval date: 05 October 2017, System ID: 12195.

Informed Consent Statement: All subjects gave their informed consent for inclusion before they participated in the user study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to Griffith University School of Information and Communication Technology's policy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schaefer, K.E. The Perception and Measurement of Human-Robot Trust. Ph.D. Thesis, University of Central Florida, Orlando, FL, USA, 2013.
2. Wang, N.; Pynadath, D.V.; Hill, S.G. Building Trust in a Human-Robot Team with Automatically Generated Explanations. In Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, FL, USA, 30 November–4 December 2015; pp. 1–12.
3. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef] [PubMed]
4. Hancock, P.A.; Billings, D.R.; Schaefer, K.E.; Chen, J.Y.; De Visser, E.J.; Parasuraman, R A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors* **2011**, *53*, 517–527. [CrossRef]
5. Salem, M.; Dautenhahn, K. Evaluating trust and safety in HRI: Practical issues and ethical challenges. In *Emerging Policy and Ethics of Human-Robot Interaction Workshop @HRI 2015*; 2nd ed.; ACM Press: Portland, OR, USA, 2015.
6. Parasuraman, R.; Riley, V. Humans and automation: Use, misuse, disuse, abuse. *Hum. Factors* **1997**, *39*, 230–253. [CrossRef]
7. Pieters, W. Explanation and trust: What to tell the user in security and AI? *Ethics Inform. Technol.* **2011**, *13*, 53–64. [CrossRef]
8. Darlington, K.W. Aspects of intelligent systems explanation. *Univers. J. Control Automat.* **2013**, *1*, 40–51.
9. Wang, N.; Pynadath, D.V.; Hill, S.G. Trust calibration within a human-robot team: Comparing automatically generated explanations. In Proceedings of the IEEE Eleventh ACM/IEEE International Conference on Human Robot Interaction, Christchurch, New Zealand, 7–10 March 2016; pp. 109–116.
10. Swartout, W.R.; Moore, J.D. Explanation in second generation expert systems. In *Second Generation Expert Systems*; Springer: Berlin/Heidelberg, Germany, 1993; pp. 543–585.
11. Ye, L.R.; Johnson, P.E. The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. *MIS Q.* **1995**, *19*, 157–172. [CrossRef]
12. Dzindolet, M.T.; Peterson, S.A.; Pomranky, R.A.; Pierce, L.G.; Beck, H.P. The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.* **2003**, *58*, 697–718. [CrossRef]
13. Goodrich, M.A.; Schultz, A.C. Human—robot interaction: A survey. In *Foundations and Trends® in Human–Computer Interaction*; Now Publishers Inc.: Delft, The Netherlands, 2008; Volume 1, pp. 203–275.
14. Sheh, R. "Why Did You Do That?" Explainable Intelligent Robots. In Proceedings of the AAMAS Workshops at the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–5 February 2017.
15. Stange, S.; Buschmeier, H.; Hassan, T.; Ritter, C.; Kopp, S. Towards self-explaining social robots. Verbal explanation strategies for a needs-based architecture. In Proceedings of the AAMAS 2019 Workshop on Cognitive Architectures for HRI: Embodied Models of Situated Natural Language Interactions, Montreal, QC, Canada, 13–17 May 2019.
16. Correia, F.; Alves-Oliveira, P.; Maia, N.; Ribeiro, T.; Petisca, S.; Melo, F.S.; Paiva, A. Just follow the suit! trust in human-robot interactions during card game playing. In Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, NY, USA, 26–31 August 2016; pp. 507–512.
17. Evans, A.M.; Krueger, J.I. The psychology (and economics) of trust. *Soc. Person. Psychol. Compass* **2009**, *3*, 1003–1017. [CrossRef]

18. Feng, L.; Wiltsche, C.; Humphrey, L.; Topcu, U. Synthesis of Human-in-the-Loop Control Protocols for Autonomous Systems. *IEEE Trans. Automat. Sci. Eng.* **2016**, *13*, 450–462. [CrossRef]
19. Orsag, M.; Haus, T.; Tolić, D.; Ivanovic, A.; Car, M.; Palunko, I.; Bogdan, S. Human-in-the-loop control of multi-agent aerial systems. In Proceedings of the 2016 European Control Conference (ECC), Aalborg, Denmark, 29 June–1 July 2016; pp. 2139–2145. [CrossRef]
20. Nunes, D.S.; Zhang, P.; Sá Silva, J. A Survey on Human-in-the-Loop Applications Towards an Internet of All. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 944–965. [CrossRef]
21. Javid, M.; Estivill-Castro, V.; Hexel, R. Enhancing Humans Trust and Perception of Robots Through Explanations. In Proceedings of the ACHI 2020: The Thirteenth International Conference on Advances in Computer-Human Interactions, Valencia, Spain, 21–25 November 2020; pp. 172–181.
22. Paeng, E.; Wu, J.; Boerkoel, J.C. Human-Robot Trust and Cooperation Through a Game Theoretic Framework. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence AAAI, Phoenix, AZ, USA, 12–17 February 2016; pp. 4246–4247.
23. Yagoda, R.E.; Gillan, D.J. You want me to trust a ROBOT? The development of a human–robot interaction trust scale. *Int. J. Soc. Robot.* **2012**, *4*, 235–248. [CrossRef]
24. Billings, D.R.; Schaefer, K.E.; Chen, J.Y.; Kocsis, V.; Barrera, M.L.; Cook, J.; Hancock, P.A. *Human-Animal Trust as an Analog for Human-Robot Trust: A Review of Current Evidence*; Technical Report; University of Central Florida: Orlando, FL, USA, 2012.
25. Miller, C.A. Trust in adaptive automation: The role of etiquette in tuning trust via analogic and affective methods. In Proceedings of the 1st International Conference on Augmented Cognition, Las Vegas, NV, USA, 22–27 July 2005; pp. 22–27.
26. Salem, M.; Lakatos, G.; Amirabdollahian, F.; Dautenhahn, K. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, ACM, Portland, OR, USA, 2–5 March 2015; pp. 141–148.
27. Anderson, J.R.; Corbett, A.T.; Koedinger, K.R.; Pelletier, R. Cognitive tutors: Lessons learned. *J. Learn. Sci.* **1995**, *4*, 167–207. [CrossRef]
28. Gregor, S.; Benbasat, I. Explanations from intelligent systems: Theoretical foundations and implications for practice. *J. MIS Q.* **1999**, *23*, 497–530. [CrossRef]
29. Anderson, J.R.; Boyle, C.F.; Reiser, B.J. Intelligent tutoring systems. *Science* **1985**, *228*, 456–462. [CrossRef] [PubMed]
30. Jackson, P. *Introduction to Expert Systems*; Addison-Wesley Longman Publishing Co. Inc.: Boston, MA, USA, 1998; Volume 6.
31. Lacave, C.; Díez, F.J. A review of explanation methods for Bayesian networks. *Know. Eng. Rev.* **2002**, *17*, 107–127. [CrossRef]
32. Sørmo, F.; Cassens, J. Explanation goals in case-based reasoning. In Proceedings of the ECCBR 2004 Workshops, Madrid, Spain, 30 August–2 September 2004; pp. 165–174.
33. Yuan, C.; Lim, H.; Lu, T.C. Most relevant explanation in Bayesian networks. *J. Artif. Intel. Res.* **2011**, *42*, 309–352.
34. Lim, B.Y.; Dey, A.K.; Avrahami, D. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Boston, MA, USA, 4–9 April 2009; pp. 2119–2128.
35. Nothdurft, F.; Ultes, S.; Minker, W. Finding appropriate interaction strategies for proactive dialogue systems—an open quest. In Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, Tartu, Estonia, 6–8 August 2014; Linköping University Electronic Press: Linköping, Sweden, 2015; Volume 110, pp. 73–80.
36. Nothdurft, F.; Minker, W. Justification and transparency explanations in dialogue systems to maintain human-computer trust. In *Situated Dialog in Speech-Based Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 41–50.
37. Glass, A.; McGuinness, D.L.; Wolverton, M. Toward establishing trust in adaptive agents. In Proceedings of the 13th International Conference on Intelligent user Interfaces, ACM, Gran Canaria, Spain, 13–16 January 2008; pp. 227–236.
38. Nothdurft, F.; Richter, F.; Minker, W. Probabilistic Human-Computer Trust Handling. In Proceedings of the SIGDIAL 2014 Conference, 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Philadelphia, PA, USA, 18–20 June 2014; pp. 51–59.
39. Javaid, M.; Estivill-Castro, V.; Hexel, R. Knowledge-Based Robotic Agent as a Game Player. In *Pacific Rim: Trends In Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 322–336.
40. Darlington, K.W. Designing for Explanation in Health Care Applications of Expert Systems. *SAGE Open* **2011**, *1*. [CrossRef] [CrossRef]
41. Malle, F. B. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*; Mit Press: London, UK, 2006.
42. Jordan, B.; Henderson, A. Interaction analysis: Foundations and practice. *J. Learn. Sci.* **1995**, *4*, 39–103. [CrossRef]
43. Heath, C.; Hindmarsh, J.; Luff, P. *Video in Qualitative Research*; Sage Publications: New York, NY, USA, 2010.
44. Lausberg, H.; Sloetjes, H. Coding gestural behavior with the NEUROGES-ELAN system. *Behav. Res. Methods* **2009**, *41*, 841–849. [CrossRef] [PubMed]
45. Hamacher, A.; Bianchi-Berthouze, N.; Pipe, A.G.; Eder, K. Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-Robot Interaction. In Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, NY, USA, 26–31 August 2016; pp. 493–500.
46. Walter, M.L.; Dautenhahn, K.; Te Boekhorst, R.; Koay, K.L.; Kaouri, C.; Woods, S.; Werry, I. The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. In Proceedings of the ROMAN 2005, IEEE International Workshop on Robot and Human Interactive Communication, Nashville, TN, USA, 13–15 August 2005; pp. 347–352.

47. Chen, M.; Nikolaidis, S.; Soh, H.; Hsu, D.; Srinivasa, S. The role of trust in decision-making for human robot collaboration. In Proceedings of the Human-Centered Robotics workshop of the 13th International Conference on Robotics: Science and System (RSS), Cambridge, MA, USA, 10–12 July 2017.
48. Haring, K.S.; Silvera-Tawil, D.; Matsumoto, Y.; Velonaki, M.; Watanabe, K. Perception of an android robot in Japan and Australia: A cross-cultural comparison. In Proceedings of the International Conference on Social Robotics, Sydney, NSW, Australia, 27–29 October 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 166–175.
49. Bartneck, C.; Kulić, D.; Croft, E.; Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **2009**, *1*, 71–81. [CrossRef]
50. Kiesler, S.; Goetz, J. Mental models of robotic assistants. Extended abstracts. In Proceedings of the CHI'02 Human Factors in Computing Systems, Minneapolis, MN, USA, 20–25 April 2002; pp. 576–577.
51. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2018**, *267*, 1–38. [CrossRef]