Enhancing Humans Trust and Perception of Robots Through Explanations

Misbah Javaid
School of ICT
Griffith University
Nathan, Queensland, Australia 4111
Email: misbah.javaid@griffithuni.edu.au

Vladimir Estivill-Castro
School of ICT
Griffith University
Nathan, Queensland, Australia 4111
Email: v.estivill@griffith.edu.au

Rene Hexel
School of ICT
Griffith University
Nathan, Queensland, Australia 4111
Email: r.hexel@griffith.edu.au

***Abstract*—To integrate robots into humans' environment, robots need to make their decision-making process transparent to increase humans' trust in robots. Explanations from a robot are a promising way to express "how" a decision is made and "why" the decision made is the best. We performed a user study investigating the effect of the explanations from a robot on humans' trust. Our setting consists of an interactive game-playing environment (the partial information game *Domino*), in which the robot partners with a human to form a team. Since in the game there are two adversarial teams, the robot plays two roles: the already mentioned partner with a human in a team, but also as an adversary facing the second team of two humans. The robot's explanations are provided in *human-understandable terms*. Explanations from the robot not only provide insight into the robot's decision-making process, but also help in improving humans' learning of the task. We evaluated the human participants' implicit trust in the robot by performing multi-modal scrutiny i.e., recording observations of facial expressions and affective states during the game-play sessions. We also used questionnaires to measure participants' explicit trust and perception of the robot attributes. Our results show that the human participants considered the robot with explanations' ability as a trustworthy team-mate. We conclude explanations can be used as an effective communication modality for robots to earn humans' trust in social environments.**

Keywords–Implicit Trust; Explicit Trust; Explanations; Human-Robot Physical Interaction.

I. INTRODUCTION

Social robots have moved from manufacturing environments and are now deployed into human environments, such as in hotels, shops, hospitals and as office 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 capabilities of robotic systems have immensely improved, which has led to an increase in the autonomy and functional capabilities of existing robots [1]. As robots' abilities increase, their complexity also increases, but increased capability in robots 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 humans' loss of trust in their robot teammate. In critical situations, such as *search-and-rescue* or to complete a time-sensitive task, humans cannot afford to lose trust in robot teammates.

We hypothesise that the explanations from a robot are a promising way to express *how* a decision is made and *why* the decision-made is the best. Robots shall be required to explain and justify their decisions to humans, and humans will tend 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. These explanations will also make it possible for humans to perceive and accept the robot as a trustworthy teammate.

Trust is an important aspect for humans and robots to perform cooperatively 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].

Humans are desirous of trusting other humans, particularly if explanations are provided. Trust appears to require explanations [7]. In essence, trust-building encompasses a more or less detailed understanding of the motives of a person we may or may not trust. We accept explanations, or we may cast a validity verdict upon them. Logically, trust and explanations seem to be mutual companions in everyday life.

Artificial intelligence researchers, within the area of expert systems, have also provided sufficient motivation to consider the contribution of explanations [8] to building humans trust [9] [10] and to the acceptability of these systems [11]. Hand-craft explanations have also shown to be promising in providing enough transparency to humans [12]. Robots have become increasingly important in human society, and it seems timely and essential to understanding how to promote their interactions with humans. An interaction, by definition, requires communication between humans and robots [13]. Hence, explanations can be used as an effective communication modality for robots, earning humans' trust in a social environment. By explanations, humans will also be able to track the performance and capabilities of the robots. Hence, a clear understanding of the robots' decision-making process can also lead to humans' desire for interaction and acceptability and will also help in establishing smooth and trustworthy human-robot interactions.

 ISBN: 978-1-61208-761-0

This study sets out to examine the effect of a robot's explanations on humans' level of trust. In addition, we refer to the explanations' approach as *English like sentences*, because in this way, humans can trace the performance of the robot. We expect that, when humans understand the behaviour of the robot, they will tend to trust the robot's actions and will work together as a team, to achieve a common goal.

For human-robot interaction, there has been a little empirical evaluation of the influence of explanations on humans' level of trust. Wang [9] used a different approach to increase transparency by using a simulated robot to provide explanations of its actions. Explanations did not improve the team's performance, although trust was identified as an influential factor only in the high-reliability conditions. Moreover, Wang [9] used an online survey because human participants were not present with a physical robot. Wang's analysis of the survey's responses indicates improvements in humans acceptance of the robot's suggestions. One of the disadvantages of conducting an online survey for evaluating humans' perception of a robot's attributes is that the human participants act only as observers. Such 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 communicates via explanations. We investigate the change in humans' perception of the robot from a tool to a trustworthy teammate. By addressing this question, findings from our research can serve to guide future work in recognition of specific robots' design metrics.

We explore the influence of explanations on humans' trust. 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 immerse interaction between humans and the social robot. Game-playing scenarios are useful and powerful environments to establish human-robot interaction [14] because games provide an external, quantifiable measure of the underlying psychological state of a human's trust [15]. Especially, multiplayer game environments, not only maintain social behaviour when played in teams, but also develop trust dynamics among teammates to achieve the common goal of wining the game. Besides, we hope to enhance the intelligibility of the robot by augmenting it with the communication ability through explanations, to improve the clarity of its decision-making.

We selected *Domino* game as the basis for our experimental paradigm for the following reasons. A game of *Domino* involves two teams with two members of 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. We configured mix-teams of a human and a robot. The robot plays two roles: first, team partner with a human, and second, member of a human-robot team that competes with a team of two humans. Because each player has different tiles, each player has different resources. The environment in the game *Domino* is partially observable.

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 is changing from *master-slave* to *peer-to-peer*. Hence, to model effective human-robot interaction, the *human-in-the-loop* concept must be incorporated as frequently as possible. Hence, we adopted a *human-in-the-loop* approach by augmenting a robot with the capability of providing different types of explanations. Explanations shall make complex behaviour of the robot more understandable and intuitive for a human. We hope that explanations will lead to developing the humans' trust in the robot.

We divide this paper into different sections. Section II surveys the literature on trust and explanations in the context of human-robot interaction. Section III presents our human-robot interaction scenario followed by the design description of our robot as a team player. Section IV discusses the *User Study* in detail, as well as the experimental design and the measurement of dependent variables. Section V presents the results in detail, taking into account the proposed hypotheses. Section VI shows the correlation between the dependent variables. Section VII presents the discussion and finally, Section VIII considers the implications of this work on the human-robot interaction community.

II. 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 and empirical explorations on different factors that affect human-robot trust [16] [17]. Hancock [4] reported on 29 empirical studies and developed a triadic model of trust as a foundation to provide a greater understanding of 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. Robot performance-based factors comprised of a robot's functional capability [18], etiquette in a robot (i.e., remained attentive of errors) [19] [20], especially how the robot casts the blame of error [2], its reliability and safety [5]. Previous research [5] also provides additional support to precisely address the significance of errors and feedback from error-prone robots. In a situation where the robot's low reliability was clearly evident, even from early stages of interaction, human participants continued to follow the robot's instructions.

Most of the previous investigations regarding the influence of explanations on humans' trust have been conducted in rule-based systems [10], intelligent tutoring systems [21], intelligent systems (i.e., neural networks, case-based reasoning systems, heuristic expert systems) [8] and knowledge-based systems [22].

Intelligent tutoring systems try to convey knowledge on an exclusive subject to a learning person. Nevertheless, intelligent tutoring systems cannot clarify their behaviour and remain restricted to particular tasks [23]. Expert systems [24] 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].

 ISBN: 978-1-61208-761-0

The rule-based expert system *Mycin* [25] was the first expert system to provide trace explanations of its reasoning to respond to *Why*, *Why-Not* and *How-To* queries, but the comparative benefits of these explanations were limited [8] [26]. Since *Mycin* was incapable to justify its advice, it was observed that physicians were reluctant to use it in practice [27].

Earlier work [28] confirmed that different types of explanations not only improved the effectiveness of context-aware intelligent systems but also contributed to stronger feelings of humans' trust. Although the main focus was on the influence of the *How-to*, *What-if*, *Why* and *Why-not* explanations. However, the results showed that *Why* and *Why-not* explanations were an excellent type of explanation, which effectively helped to improve the overall understandability of the system.

For human-machine trust, there has been little empirical evaluation of the impact of explanations [11]. Dzindolet et. al. [12] explored manually crafted explanations. Hand-crafted explanations have been shown to be effective in providing transparency and improved trust. However, since hand-crafted explanations were static and created manually, they fail to transfer the complexity of the decision-making to the team members. Nothdurft et. al. [29] [30] focused on transparency and the justification of decisions in human-computer interaction. Glass et. al. [31] studied trust issues in technical systems, analysing the features that may change the level of humans 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 claim is limited.

The systems, as mentioned earlier, deliberately focused on the use of explanations to convey conceptual knowledge and acceptability of these systems, such as the reliability and accuracy of performance. However, the state-of-the-art may not resolve the problem of non-cooperative behaviour and trust of humans towards robots. To the best of our knowledge, there is still a gap in current human-robot interaction literature, and there is very little experimental verification that could show that explanations promote and certainly affect humans trust in and acceptance of robots.

Such systems, as mentioned earlier, deliberately focused on the reliability and accuracy followed by explanations to convey conceptual knowledge and their acceptability. However, the state-of-the-art leaves open the problem of non-cooperative behaviour and trust of humans towards robots. In particular, there is very little experimental verification that could show that explanations promote humans trust in robots.

It is important to realise that 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 capabilities of the robot. Nonetheless, the human's mental model evolves after interaction with the robot. The mismatch between the human's initial mental model and the later mental model creates a detrimental effect on the human's trust [32]. A human's mental model 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 that enables a robot to provide explanations for *transparency* and for *justification* of its reasoning is to be considered a robot's functional capability, which should be categorised as a robot-related factor.

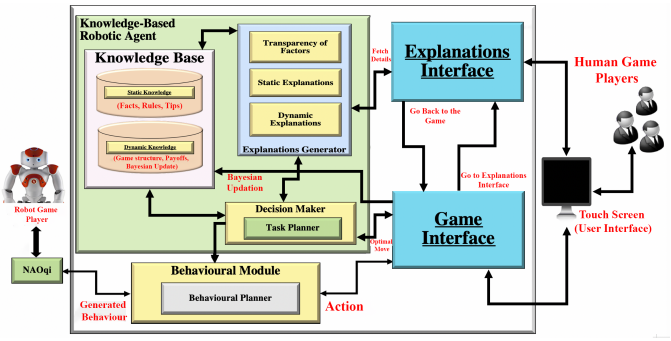

Figure 1. Complete architectural overview of our human-robot interaction scenario.

III. HUMAN-ROBOT INTERACTIVE SCENARIO

Our human-robot interactive scenario is around a block-type game known as Spanish *Domino*. A match is between two teams with two players in each team, and it consist of several *hands*; in each *hand*, each of the four player receives seven random domino tiles. Game players take their turn clockwise and aim for their couple to have the first player to release all its *hand*. The *hand* is confidential to its owner. Thus, the decisions a player makes are with partial information. At each turn, a game player can perform only two actions,

1) to release a tile (by putting a tile with an endpoint matching one of the open ends of the current board), or
2) to *pass* (because to release a tile is impossible).

The game ends when no player can play a domino tile or when a player runs out of the domino tiles.

Domino is a non-deterministic game, because of the random shuffling and dealing of tiles to four players at the beginning of every game. This initial *hands'* aspect is an element of non-determinism, but after each player has received 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 1 shows the global architecture and the modules involved in our software for human-robot interaction [33]. Our *knowledge-based robotic agent* is capable of performing rapid updates of knowledge while playing the multi-player game of *Domino* with humans in the partial-information environment and in teams. Information becomes available to all players each time a player completes its turn; either by releasing a tile, or *passing*.

Bayesian inference is an effective way to deal with such partial observability. We incorporate *Bayesian inference* into our *knowledge-based robotic agent*. By using Bayesian inference, the *knowledge-based robotic agent* can update information about the environment (i.e., the current state of the game). The update is performed after an observation. An

 ISBN: 978-1-61208-761-0

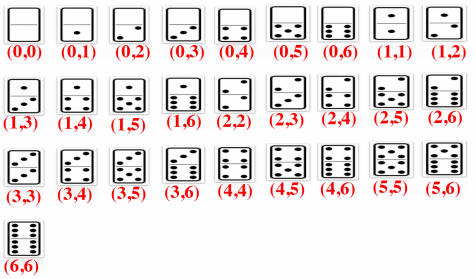

Figure 2. Complete Set of domino tiles used in the study ranging from (0,0) to (6,6).

observation provides new evidence, enabling the update of the belief representation. These observations are forwarded to the *Knowledge-Base* module.

The *knowledge-based robotic agent* controls the two roles of the robot: firstly, as an adversary with two humans and secondly, as a team partner with a human. Therefore, it displays cooperation with human team partners, but is goal oriented and competes with human opponents. We developed the explanation-generation mechanism on top of the game-playing mechanism.

We enable the robot to generates multiple *static* and *dynamic* explanations. The *static* explanations are based upon (1) *history and facts* about the game, (2) *rules of the game* and (3) *game-play tips*. While, *dynamic* explanations provide insight into the *knowledge-based robotic agent*'s decision-making process. These *dynamic* explanations would be suitable to answer *how-type* and *why-type* questions. Furthermore, *dynamic* explanations provide team members with the *transparency* for the different factors involved in the decision-making process of the *knowledge-based robotic agent*.

The mechanism for generating *dynamic* explanations is meaningful for the strategic aspect of the game.

IV. User Study

Using the human-robot interactive scenario discussed in Section III, we conducted a *User Study* to investigate the effect of a robot's explanations on the humans' level of trust and how much the explanations are effective in changing humans' perception of the robot attributes during an interactive task.

A. Hypotheses

Hypothesis 1 - Human participants would appreciate understanding about how the robot's decisions are made (*transparency*) and receiving informed *justifications* of the robot's choices in a partial information environment. Such human inclination will be reflected by an increase in humans' trust in the robot.

Hypothesis 2 - Explanations that provide *transparency* and supply *justification* for a robot's decisions in a collaborative (team-based) environment help in changing humans' perception of the robot attributes.

B. Variables

The independent variable is the explanations of the robot at the beginning of the first game and after the end of the match. For the quantitative assessment, both subjective and objective analysis of the interaction took place. The dependent variables fall into three categories to analyse the impact of explanations:

1) **Trust** - Human teammates' trust, which is not directly observable [34], by using a 14-items subscale of the Human-Robot Trust questionnaire [1] before and after interaction with the robot.
2) **Perception of Robot Attributes** - We used Godspeed questionnaire [35] [36] before and after interaction with the robot to evaluate human perception of the robot attributes related to trustworthiness [1]. We use the Godspeed questionnaire because it is a standardised measurement tool for interactive robots [35]. The Godspeed questionnaire uses a 5-point scale to measure five key concepts in human-robot interaction. (1) *Anthropomorphism* [37] is the characteristics of a human form. (2) *Animacy* [38] is the perception of a robot as a lifelike living entity. Perceiving things as living creatures allows humans to distinguish between humans and machines [38]. (3) *Likeability* [39] describes the first (positive) impression that humans make in their mind of others. Previous research investigated [40] that humans tend to consider robots as social agents; hence deal with them in a similar way. (4) *Perceived intelligence* [41] indicates how intelligent; the human participants judge the robot by its explanatory ability. (5) *Perceived safety* is the perception of danger attributable to the robot during the collaboration and the level of comfort the human participants' experience during the interaction [42].
3) **Previous experience of human participants** - Prior relationship with non-human agents such as pets [43] influence the interaction of a human 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 [43].

   We also showed two pictures, each with a social robot (i.e., *Nao* and *Pepper*) to human participants and assessed their initial impression of the robots. We also asked human participants to rate these images by classifying them as (1. human-like, 2. machine-like, 3. child-like, 4. toy-like, and 5. avatar). Trust between humans and animals may be a suitable analogy to trust between humans and robots [2]. To examine the nature of a human-animal relationship can help in increasing the understanding of how a human interacts with and trusts a robot [18].

C. Additional Measurement

Before starting the experiment, we instructed the human participants about the procedure of the experiment. The human participants were allowed to ask any relevant questions before starting the formal experiment. We asked the human participants to maintain a safe distance from the robot, so no human participant will push or damage the robot in any way.

 ISBN: 978-1-61208-761-0

In addition, no human participant can interrupt the robot and ask for explanations during a game.

The recognition of humans' affective states and emotions is one of the much-studied research questions at the moment [29] that can be recognised via vision-based, audio-based, and audio-visual recognition [44]. Therefore, we video-recorded the experiment to examine the affective states and behavioural responses of the human participants towards the robot. We also maintained a history at the backend of the system to record the moves played by the human participants. Also, we kept a history of the human participants examination and use of the robot's explanations. This record of explanation usage was used later to investigate, which type of explanations were accessed more i.e., *static* explanations or *dynamic* explanations.

D. Procedure of the Experiment

We adopted the approach of combining survey(s) with an experiment to evaluate the humans' perception and trust towards the robot. We experimented in three-stages.

1) ***Stage-1 of the Experiment****:* During Stage 1, we evaluated human participants' demographics, initial perception and trust towards the robot.

2) ***Stage-2 of the Experiment****:* Before starting the formal game activity, the robot greeted the human participants and provided verbal static explanations of how to play the game.

> ***"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. . . !"***

Next, human participants played repeated hands with the robot in teams, until reaching a pre-defined score. After the match, the human participants examined the explanations. Following the explanations session, human participants played more hands with the robot, until reaching a pre-defined score. The second game-playing session aims at observing the effect of explanations, and whether explanations improve a team's performance.

3) ***Stage-3 of the Experiment****:* Trust is a dynamic attitude that changes over time [1] [3]. On the completion of Stage 2 of the experiment, to elucidate the changes in trust by human participants and their perception of the robot, human participants filled out another human-robot trust questionnaire and Godspeed questionnaire. Changes in the level of trust and perception of the robot attributes will elicit the influence of explanations from the robot. At the end of the experiment, the robot thanked all the human participants for their participation.

E. Recruitment and Participation

This study was conducted in Griffith University Australia, and there were a total of 33 human participants, (15 females and 18 males) with ages ranging from 19 to 35 years old (M = 28.33 $\pm$ 4.58). We recruited human participants through general advertising, using posters on university notice board, and communicating directly with students. Each human participant received an invitation letter for the main objective of conducting the experiment. Along with the invitation letter, we also attached a brochure with a brief description of the *Domino* game. We expected all human participants to start

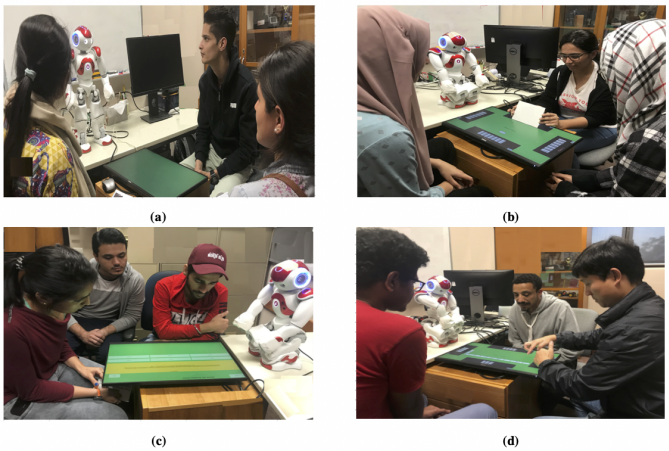

Figure 3. (a) Robot is explaining how to play the game (b) Player 2 is taking its turn (c) Human participants' are checking the robot's explanations (d) Player 3 is taking its turn after the explanations' session.

Cronbach's α for Scales used in the Experiment		
Sr. #	Scale	α
1	Shafear Human-Robot Trust Scale	0.729
2	Godspeed Questionnaire	0.893

Figure 4. Statistics for Cronbach's α for the customised scales used in the experiment.

with the same common-sense model of the task (i.e., the *Domino* game), which also helped us estimate what knowledge the human participants have already possessed about the task. Before taking part in the experiment, all human participants provided their consent.

We offered an *Aud* 10 gift card as a *token of appreciation* to every human participant. We configured human-robot matches, with four participants i.e., one robot and three humans in each team. There were 11 groups in total. Each group played two matches with the robot. A single match consists of a maximum of five hands in total, or until a pre-defined score is reached. Each group played two matches, the first match before the explanations' session and the second after the explanations' session.

V. Experiment Results

Prior to conducting any analysis, we performed a reliability analysis (Cronbach's α) to assess the internal reliability of the Human-Robot Trust Questionnaire [1] and Godspeed Questionnaire [35] [36]. An $\alpha > 0.7$ or higher is considered acceptable, which indicates the reliability of the measuring scale. Figure 4 shows Cronbach α for all the scales used in the experiment.

A. Effect of Robot Explanations on Humans' Trust

After performing the reliability analysis, we performed a normality analysis by applying the *Shapiro-Wilk test*. The *Shapiro-Wilk test* showed the dependent variable trust fit a normal distribution satisfactorily. Therefore, we performed a parametric paired sample *t-test* to analyse the effect of robot explanations. We compared the levels of trust that human participants had in their robot teammate after interaction,

 ISBN: 978-1-61208-761-0

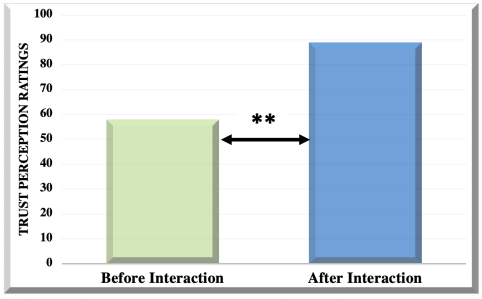

Figure 5. Difference in the level of trust of the human participants in the robot before and after interaction - (**$p < 0.01$).

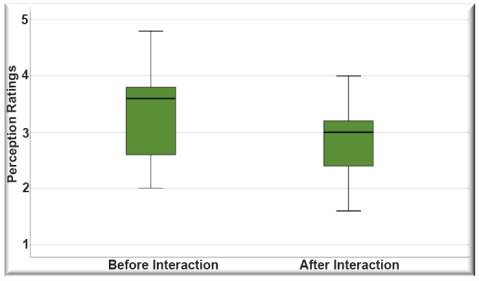

Figure 6. After interacting with the robot, the *anthropomorphism* ratings of the robot decreased.

controlling for the levels of trust reported before interaction. Results showed a significant difference ($t(32)$ = -7.729, $p<0.001$); Figure 5 displays much higher trust levels of human participants towards the robot after interaction (M = 89.27 $\pm$ 6.44), when compared with their respective trust levels before interaction (M = 58.24 $\pm$ 9.44). Overall, the analysis indicates that the robot is successful in earning the trust of the human participants' based on the notable distinction between the trust levels before interaction and after interaction.

B. Effect of Explanations in changing humans' perception of the robot.

We performed the *Shapiro-Wilk test*, which indicated that the Godspeed questionnaire follows a normal distribution. Following this, we performed paired sample *t-test* to scrutinize the effect of explanations from the robot in changing humans' perception of the robot.

1) Anthropomorphism: We analysed the decline in the degree of anthropomorphism after interacting with the robot: $t(32) = 4.389$, $p<0.001$ (refer to Figure 6). These values reflect that the humans' perception of anthropomorphism of the robot was reduced significantly after interaction (M = 2.9 $\pm$ 0.59) when compared with before interaction (M = 3.4 $\pm$ 0.79). The results indicate that the human participants considered the robot less human-like, less natural and less conscious.

2) Animacy: The robot's explanations created a positive effect on the perception of the robot's *animacy* : $t(32)$ = -4.884, $p<0.001$ (refer to Figure 7). We observed higher perception ratings of the robot's animacy after the interaction (M = 3.6 $\pm$ 0.71), when compared to before the interaction (M = 2.88 $\pm$ 0.50). The results show that the human participants appraise the robot as more interactive and responsive.

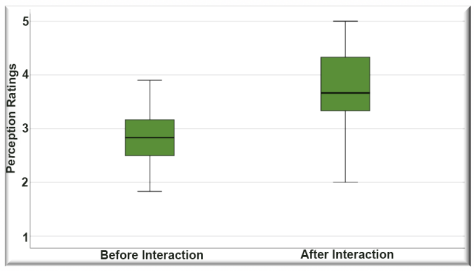

Figure 7. The *animacy* ratings of the robot significantly increased, after interacting with the robot.

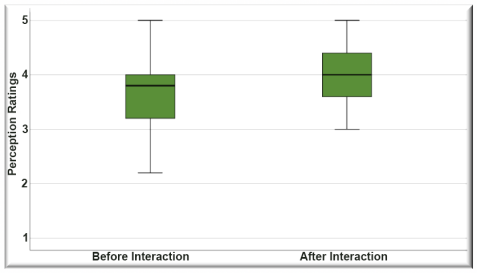

Figure 8. After interacting with the robot, the *Likeability* ratings of the robot greatly increased.

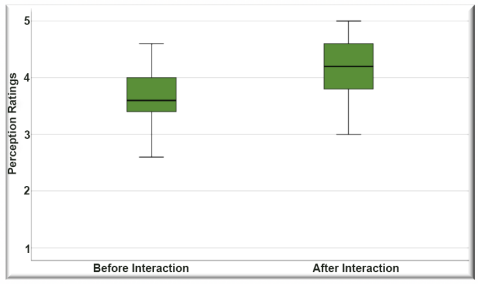

Figure 9. Difference in the perception ratings show the rise of the robot's *Perceived Intelligence*.

3) Likeability: Significant differences were found in the *likeability* of the robot : $t(32)$ = -3.522, p = 0.001. Figure 8 shows a significant difference in the perception ratings of the robot after interaction (M = 4.07 $\pm$ 0.55), when compared with the perception ratings before interaction (M = 3.60 $\pm$ 0.69). The results show that the human participants considered the robot pleasant and friendly.

4) Perceived Intelligence: Figure 9 shows the rise of the robot's perceived intelligence: $t(32)$ = -5.502, $p < 0.001$. We observed a significant difference between the pre-interaction ratings (M = 3.70 $\pm$ 0.41) and the post-interaction ratings (M = 4.23 $\pm$ 0.50). The results provide evidence that the human participants considered a robot with explanatory capability to be more intelligent, knowledgeable and competent.

5) Perceived Safety: Figure 10 shows that there is no significant differences between the perceived safety levels before and after interacting with the robot. Consequently, there were no significant changes in this aspect as a result of interaction with the robot.

 ISBN: 978-1-61208-761-0

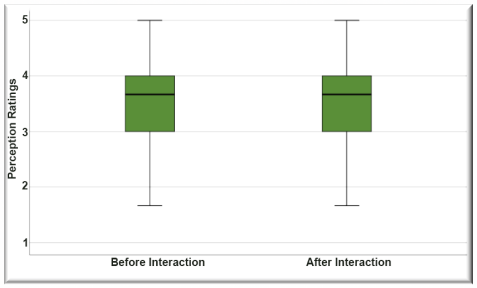

Figure 10. After interacting with the robot, there was no significant difference in the level of *Perceived Safety*.

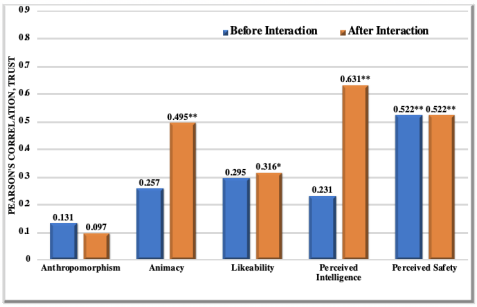

Figure 11. Correlation between trust and humans' perception of the robot, before interaction and after interaction - ($\star\star$Correlation is significant at $p < 0.01$, $\star$Correlation is significant at $p < 0.05$).

VI. Correlation Between Dependent Variables

We also conducted *Pearson's* (parametric) correlation to analyse (1) how much humans' trust and perception of the robot are correlated with each other and (2) how much trust impacts in changing humans' perception of the robot.

We found no significant correlation between the dependent variable *trust* and the *anthropomorphism* attribute. This result applies to both cases, before and after interacting with the robot.

Before interacting with the robot, we did not find any correlation between *trust* and *animacy, likeability* and *perceived intelligence* attributes of the robot. However, after the interaction, as *trust* increased, we observed a significant positive correlation between *trust* and the robot's *animacy, likeability* and *perceived intelligence* attributes. We also observed a significant positive correlation between *trust* and the *perceived safety* attribute before interaction with the robot, and it did not change after interaction with the robot.

VII. Discussion

Results from our preliminary analysis strongly support our *Hypothesis 1* by indicating that explanations increased human participants' trust in the robot. Moreover, explanations also improved the humans' perception of the robot attributes associated with trust, which is our *Hypothesis 2*. However, after interacting with the robot, the perception ratings of the *anthropomorphism* attribute decreased, and in a sense, our results partially support *Hypothesis 2*. Furthermore, for the *perceived safety* attribute, we did not see any difference in the perception ratings, neither before interacting nor after interacting with the robot. We suggest that the human participants considered that

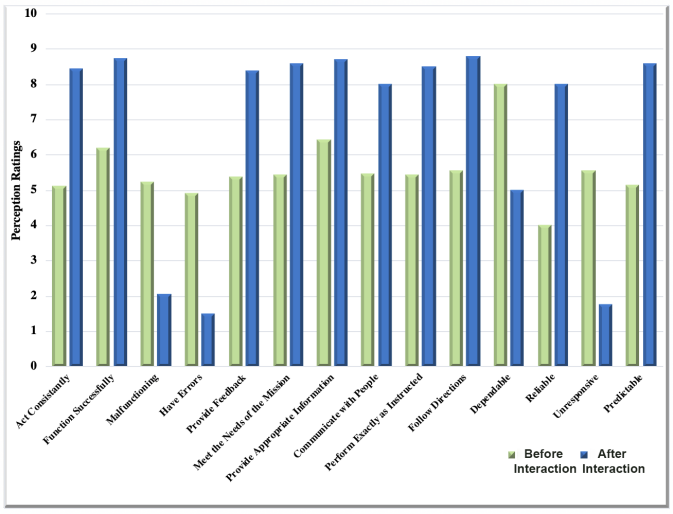

Figure 12. A summary of the quantitative data analysis results for trust.

the robot in the experiments conditions was not dangerous at all.

Previous studies have shown the role of transparency in building trust [12]. However, transparency alone may not be sufficient to establish trust. Hence, we designed our explanations to provide not only *transparency* about the mechanism for the robot's decisions, but also communicate *justification*s for the underlying sophisticated reasoning. We aimed at explaining the robot's motive for each of the decisions. Additionally, we believe explanations provide the human participants' with an insight into the concrete and individual factors involved in the decision-making process of the robot. Therefore, in the current study, explanations not only improved the trust of human participants' but also changed their overall impression of the robot. The results help us gain insight into how to design explanations to increase humans' trust.

Furthermore, Figure 12 shows that items measuring the robot's explanatory ability: *Provide Feedback*, *Provide Appropriate Information* and *Communicate with People* are tightly connected with the outcome.

Similarly, Figure 13 displays an increase in the ratings of the attributes *animacy*, *likeability* and *perceived intelligence*. This increase reflects that human participants' adopted a model of the robot that is more interactive, competent, knowledgeable and intelligent. In terms of *anthopomorphism*, human participants showed more concern by lowering the level of ratings associated with anthropomorphism. Even if a robot looks like a human, humans do not consider its capabilities to be human-like. This is an interesting result, because regardless of the less anthropomorphic perception, human participants still trusted the robot.

Most of the human participants had their previous interaction with robots through fictitious media or movies; thus, we believe that our results are not biased (or affected) by the human participants' previous experience with a physical present robot. Similarly, we did not find any partiality or differences in the results, for no-pet ownership with respect to pet ownership.

We also examined human participants' multi-modal

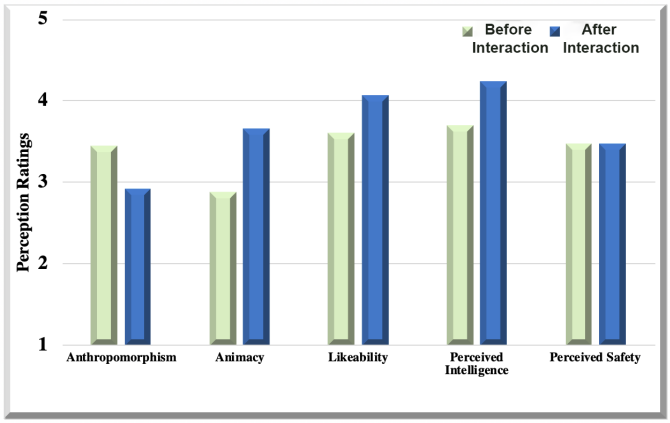

Figure 13. Summary of results for the quantitative data analysis of human participants' perception of the robot.

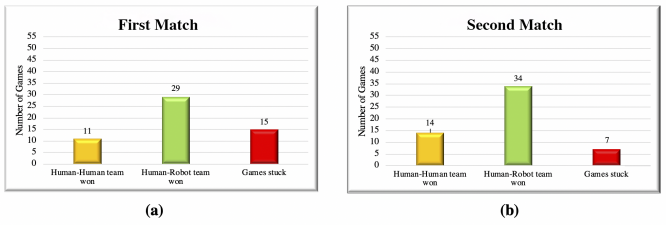

Figure 14. Change in the impression of the robot (a) before interaction (b) after interaction.

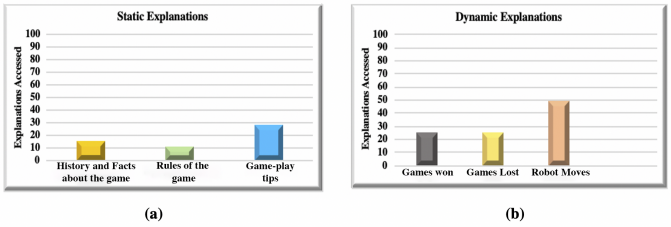

Figure 15. Change in the impression of the robot (a) before interaction (b) after interaction.

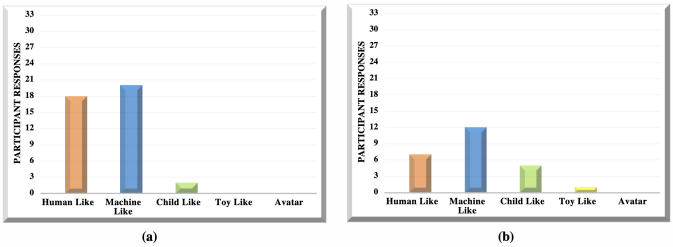

Figure 16. Change in the impression of the robot (a) before interaction (b) after interaction.

scrutiny i.e., facial expressions and affective states during the match and the explanations' sessions. We observed that human participants were unreserved, open to the robot, and even trying to cuddle it from a distance. When the robot was explaining the rules of the game, we noticed that human participants maintained eye contact with the robot, which is another signal of willingness to interact and affects trust. During the explanations' session, we observed human participants' facial expressions. These facial expressions show interest and engagement in explanations. We examined human participants' gestures and body movements while involved in a game. Players struggle to hide tiles, but reflected before making a move, were attentive towards the robot when the robot was speaking and describing its move and after the robot finished its turn, participants focused on the released tile, assimilating further the robot decision and play. As we mentioned, we provided the human participants with a brochure that briefly described the rules and mechanics of the *Domino* game. In addition, the robot also provided explanations for the mechanics of the game. Hence, our expectation is that all human participants starting playing ability is similar, and approached the matches with the same common-sense model. By evaluating the moves of the players stored in our records, we observed the implicit trust of human partners in a team. The records also show moves where humans exhibit cooperation and sacrifice also to their robotic partner.

Furthermore, the human participants' learning of the task domain enhanced, which is reflected by the increase in the number of games the human-human team won, after the explanations' session. We also investigated the human participants' use of strategies to select their moves, which was significantly improved and became visible in the second match. For example, the human participants considered playing random tiles in the first match. After the explanations' session, human participants' used some of the strategies i.e., preferred to play tiles with the highest points and put doubles on the board during the early stages of the hand.

In addition, we also kept a record of the number of times a human participant (partner/adversary) accessed explanations i.e., *static* or *dynamic*. We examined that the human participants (regardless of team partners or opponents) accessed the *dynamic* explanations more, to investigate the robot's decisions.

VIII. CONCLUSION

Overall, our results confirm that, in a team-based collaborative environment, the explanations that disseminate transparency and justification of a robot's decisions facilitate human-robot interaction.

Significant differences in the level of trust and perception of the robot, before and after the interaction, confirm that the robot has successfully earned the trust of the human participants through its explanations' ability. Besides, the strong correlation between trust and perception of the robot also suggests that the explanations helped change the overall impression of the robot.

To date, humans have rarely encountered physical robots in their lives, so their perception of robots may be affected by fictitious media. We expect that, as the opportunity for interaction with physically present robots increase, our study will be taken into account for future robot design metrics. Consequently, the findings of this study can be used to guide future work to determine specific robot design standards. So far, our work is the first to study the impact of explanations from a robot on humans' trust, by establishing peer-to-peer human-robot interaction. Overall, the results suggest that explanations

 ISBN: 978-1-61208-761-0

can potentially relieve the issue of misusing or under-utilizing a robot, which usually happens in the "absence" of trust.

REFERENCES

[1] K. E. Schaefer, "The perception and measurement of human-robot trust," Ph.D. dissertation, University of Central Florida Orlando, Florida, 2013.

[2] N. Wang, D. Pynadath, S. Hill, and A. P. Ground, "Building trust in a human-robot team with automatically generated explanations," in Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), December 2015, pp. 1–12, paper No. 15315.

[3] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 46, no. 1, 2004, pp. 50–80.

[4] Hancock et al., "A meta-analysis of factors affecting trust in human-robot interaction," Human Factors, vol. 53, no. 5, 2011, pp. 517–527.

[5] M. Salem and K. Dautenhahn, "Evaluating trust and safety in hri: Practical issues and ethical challenges," Emerging Policy and Ethics of Human-Robot Interaction, 2015.

[6] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," Human Factors, vol. 39, no. 2, 1997, pp. 230–253.

[7] W. Pieters, "Explanation and trust: what to tell the user in security and ai?" Ethics and information technology, vol. 13, no. 1, 2011, pp. 53–64.

[8] K. Darlington, "Aspects of intelligent systems explanation," Universal Journal of Control and Automation, vol. 1, no. 2, 2013, pp. 40–51.

[9] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in The Eleventh ACM/IEEE International Conference on Human Robot Interaction. IEEE Press, 2016, pp. 109–116.

[10] W. R. Swartout and J. D. Moore, "Explanation in second generation expert systems," in Second generation expert systems. Springer, 1993, pp. 543–585.

[11] L. R. Ye and P. E. Johnson, "The impact of explanation facilities on user acceptance of expert systems advice," MIS Quarterly, vol. 19, no. 2, 1995, pp. 157–172.

[12] Dzindolet et al., "The role of trust in automation reliance," International Journal of Human-Computer Studies, vol. 58, no. 6, 2003, pp. 697–718.

[13] M. A. Goodrich, A. C. Schultz et al., "Human–robot interaction: a survey," Foundations and Trends® in Human–Computer Interaction, vol. 1, no. 3, 2008, pp. 203–275.

[14] F. Correia, P. Alves-Oliveira, N. Maia, T. Ribeiro, S. Petisca, F. S. Melo, and A. Paiva, "Just follow the suit! trust in human-robot interactions during card game playing," in Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on. IEEE, 2016, pp. 507–512.

[15] A. M. Evans and J. I. Krueger, "The psychology (and economics) of trust," Social and Personality Psychology Compass, vol. 3, no. 6, 2009, pp. 1003–1017.

[16] E. Paeng, J. Wu, and J. C. Boerkoel, "Human-robot trust and cooperation through a game theoretic framework," in AAAI, 2016, pp. 4246–4247.

[17] R. E. Yagoda and D. J. Gillan, "You want me to trust a robot? the development of a human–robot interaction trust scale," International Journal of Social Robotics, vol. 4, no. 3, 2012, pp. 235–248.

[18] Billings et al., "Human-animal trust as an analog for human-robot trust: A review of current evidence," University Of Central Florida Orlando, Tech. Rep., 2012.

[19] C. A. Miller, "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, 2005, pp. 22–27.

[20] 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 Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. ACM, 2015, pp. 141–148.

[21] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier, "Cognitive tutors: Lessons learned," The journal of the learning sciences, vol. 4, no. 2, 1995, pp. 167–207.

[22] S. Gregor and I. Benbasat, "Explanations from intelligent systems: Theoretical foundations and implications for practice," The Journal of MIS quarterly, 1999, pp. 497–530.

[23] J. R. Anderson, C. F. Boyle, and B. J. Reiser, "Intelligent tutoring systems," Science, vol. 228, no. 4698, 1985, pp. 456–462.

[24] P. Jackson, Introduction to expert systems. Addison-Wesley Longman Publishing Co., Inc., 1998, vol. 6.

[25] C. Lacave and F. J. Díez, "A review of explanation methods for bayesian networks," The Knowledge Engineering Review, vol. 17, no. 2, 2002, pp. 107–127.

[26] F. Sørmo and J. Cassens, "Explanation goals in case-based reasoning," in Proceedings of the ECCBR 2004 Workshops, 2004, pp. 165–174, 142-04.

[27] C. Yuan, H. Lim, and T.-C. Lu, "Most relevant explanation in bayesian networks," Journal of Artificial Intelligence Research, vol. 42, 2011, pp. 309–352.

[28] B. Y. Lim, A. K. Dey, and D. Avrahami, "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, 2009, pp. 2119–2128.

[29] F. Nothdurft, S. Ultes, and W. Minker, "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, vol. 110. Linköping University Electronic Press, August 6th-8th, 2014 2015, pp. 73–80.

[30] F. Nothdurft and W. Minker, "Justification and transparency explanations in dialogue systems to maintain human-computer trust," in Situated Dialog in Speech-Based Human-Computer Interaction. Springer, 2016, pp. 41–50.

[31] A. Glass, D. L. McGuinness, and M. Wolverton, "Toward establishing trust in adaptive agents," in Proceedings of the 13th international conference on Intelligent user interfaces. ACM, 2008, pp. 227–236.

[32] F. Nothdurft, F. Richter, and W. Minker, "Probabilistic human-computer trust handling." in SIGDIAL Conference, 2014, pp. 51–59.

[33] M. Javaid, V. Estivill-Castro, and R. Hexel, "Knowledge-based robotic agent as a game player," in Pacific Rim : Trends In Artificial Intelligence. Springer, 2019, pp. 322–336.

[34] Chen et al., "The role of trust in decision-making for human robot collaboration," in Workshop on Human-Centered Robotics, RSS, 2017.

[35] K. S. Haring, Y. Matsumoto, and K. Watanabe, "Perception and trust towards a lifelike android robot in japan," in Transactions on Engineering Technologies. Springer, 2014, pp. 485–497.

[36] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," International journal of social robotics, vol. 1, no. 1, 2009, pp. 71–81.

[37] A. Powers and S. Kiesler, "The advisor robot: tracing people's mental model from a robot's physical attributes," in Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction. ACM, 2006, pp. 218–225.

[38] K. M. Lee, N. Park, and H. Song, "Can a robot be perceived as a developing creature? effects of a robot's long-term cognitive developments on its social presence and people's social responses toward it," Human communication research, vol. 31, no. 4, 2005, pp. 538–563.

[39] J. L. Monahan, "I don't know it but I like you: The influence of nonconscious affect on person perception," Human Communication Research, vol. 24, no. 4, 1998, pp. 480–500.

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

[41] R. M. Warner and D. B. Sugarman, "Attributions of personality based on physical appearance, speech, and handwriting." Journal of Personality and Social Psychology, vol. 50, no. 4, 1986, pp. 792–799.

[42] D. Kulic and E. A. Croft, "Affective state estimation for human–robot interaction," IEEE Transactions on Robotics, vol. 23, no. 5, 2007, pp. 991–1000.

[43] M. L. Walter et al., "The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment," in

 ISBN: 978-1-61208-761-0

Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on. IEEE, 2005, pp. 347–352.

[44] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, 2009, pp. 39–58.

 ISBN: 978-1-61208-761-0