Expectations Vs. Reality: Unreliability and Transparency in a Treasure Hunt Game With Icub

Alexander M. Aroyo, Dario Pasquali, Austin Kothig, Francesco Rea, Giulio Sandini, and Alessandra Sciutti

***Abstract*—Trust is essential in human-robot interactions, and in times where machines are yet to be fully reliable, it is important to study how robotic hardware faults can affect the human counterpart. This experiment builds on a previous research that studied trust changes in a game-like scenario with the humanoid robot iCub. Several robot hardware failures (validated in another online study) were introduced in order to measure changes in trust due to the unreliability of the iCub. A total of 68 participants took part in this study. For half of them, the robot adopted a transparent approach, explaining each failure after it happened. Participants' behaviour was also compared to the 61 participants that played the same game with a fully reliable robot in the previous study. Against all expectations, introducing manifest hardware failures does not seem to significantly affect trust, while transparency mainly deteriorates the quality of interaction with the robot.**

***Index Terms*—Social HRI, acceptability and trust, human-robot collaboration, unreliability, transparency.**

I. Introduction

TRUST is fundamental in any interaction between two agents that manifest a certain degree of autonomy. Indeed, help is not accepted by a partner who is not trusted. Trust is in fact defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [1] or as "the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others" [2].

As a consequence, also for robots to become actual helpers, it is necessary that they become trustworthy. Otherwise, human partners will not rely on robot support, and artificial agents will risk to remain little more than complex tele-operated tools [3]. Thus, given its centrality for dependable human-robot collaboration, trust has now gained particular attention in the community studying natural interactive process between humans and robots.

However, machines and technology in general are still not fully reliable, and in some cases, can trigger risky and irresponsible behaviors, especially when the human partner erroneously estimates the actual capabilities or intents of the robots [4], [5]. Several studies explored which factors influence humans' trust toward robots, such as robots' shape, performances [6]–[9] and unreliability [10]–[12]. In particular, the perception of robot trustworthiness decreases in presence of robot failures [13], [14].

A factor that has been shown to have a major effect on trust in previous human-robot interaction (HRI) studies is the transparency of robot's actions [15], [16]. Robots that explain their behavior and/or decision-making process can assist with trust calibration [17], [18], and a robot can build trust by providing explanations [19], [20].

Transparency has a particularly relevant role in presence of failures: Dzindolet *et al.* showed that trust in an automated decision aid decreased after the system made an error, unless it provided an explanation of its behavior [21]. Correia *et al.* showed that trust is lost when a fault occurs, and a provision of an explanation mitigates that loss [22]. Desai *et al.* proved that drops in reliance can affect trust, but warning about possible failures does not affect negatively trust [23].

Although trust in HRI is now extensively researched (see [16], [24]), it is still not clear under which conditions failures in robot behavior are sufficient to undermine the trust towards the platform. Indeed, whereas unreliability often is associated to a decrease in perceived trust (*e.g.,* [10], [11]), sometimes even overt malfunctions do not dissuade humans from obeying robots' suggestions [5], [13].

The goal of this research is to investigate how overt unreliability and transparency about robot failures affect participants' trust toward the robot. To this aim, we build on top of an already validated interactive experiment that studied trust in HRI, the Treasure Hunt (TH). TH is a game where participants have to find hidden objects in a room, potentially relying on the humanoid robot iCub's help by asking it for hints [9]. This experiment design showed that participants progressively build a rapport and increase their trust towards the robot during the game, as measured by analyzing their behavior and questionnaire responses.

This new experiment, Unreliable Treasure Hunt (UTH), introduces several evident mechanical failures in the robot behavior during the game. The failures compromise its voice intelligibility, the naturalness of its pointing motion and even lead to a sudden crash and reboot of the platform. To select failures that looked realistic and could be considered worrisome by participants, several versions of the faults were designed and

Manuscript received February 24, 2021; accepted May 5, 2021. Date of publication May 25, 2021; date of current version June 7, 2021. Researchers of the SIRRL Laboratory at University of Waterloo are supported in part, thanks to funding from the Canada 150 Research Chairs Program. This work has been supported by a Starting Grant from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme G.A. No 804388, wHiSPER. *(Corresponding author: Alexander Mois Aroyo)*

Alexander M. Aroyo and Austin Kothig are with the Social and Intelligent Robotics Research Laboratory (SIRRL), Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada (e-mail: alexander.aroyo@uwaterloo.ca; austin.kothig@uwaterloo.ca).

Dario Pasquali is with the DIBRIS, Università di Genova, Genova, GE 16152, Italy.

Francesco Rea, Giulio Sandini, and Alessandra Sciutti are with the Department of Robotics, Brain and Cognitive Sciences, Italian Institute of Technology (IIT), Genova 16163, Italy (e-mail: Dario.Pasquali@iit.it; francesco.rea@iit.it; giulio.sandini@iit.it; alessandra.sciutti@iit.it).

Digital Object Identifier 10.1109/LRA.2021.3083465

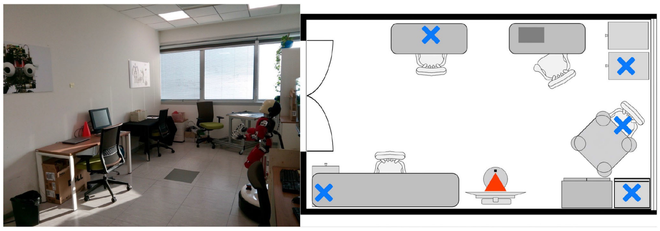

Fig. 1. Left - Experimental room. Right - Schematic layout of the room locating the hidden objects (blue cross) and the robot (red triangle). Figures from [9].

run in an online validation. The most severe and impacting ones were selected and implemented in the game [12]. Furthermore, for one group of participants, the robot provides an explanation after each failure (Transparent condition); whereas for the other group no explanation is given (Non-Transparent condition). This design allows to test two specific hypotheses: (H1) *Severe mechanical faults will negatively impact on participants' trust towards the robot,* i.e.,*trust in UTH will be lower than in TH*; (H2) *Transparency about the faults will alleviate that loss,* i.e.,*trust in the Transparent condition will be higher than in the Non-Transparent one.*

II. Methodology

The main goal of this experiment is to explore how robotic failures can affect users' trust towards the robot, and whether robot's transparency (*i.e.,* having the robot overtly communicate the occurrence of its faults) would reduce the loss in trust.

A. Experimental Setup

1) Treasure Hunt: This experiment, called Unreliable Treasure Hunt (UTH), is based on the previously validated Treasure Hunt (TH) [9], to which it adds the introduction of robotic unreliability.

TH, in a nutshell, was an interactive and autonomous game design to study how trust evolved and changed during a short HRI. Participants were given the task to find 5 hidden eggs in a room in less than 20 minutes with the chance to win € 7.5 if successful (Fig. 1, left). They were left alone with the humanoid robot iCub [25], without any instruction on its role. The experiment was divided as follows:

Phase I - Dialog: iCub was chatting with the participants for around 3-5 minutes, to relax them, and make them used to its movements and speech.

Phase II - Game: The game started - after 30 seconds iCub explained that participants could touch its torso in order to get hints about the the eggs. Each egg (Fig. 1, right) had one location hint (done by pointing), and 3 speech based hints that incrementally revealed the egg location (*e.g.,* "green in green"; "you use it when you are tired"; "under the chair"). The speech based hints were also shown as written text on a TV screen behind iCub.

Phase III - Bonus: If participants managed to find the 5 eggs under the time limit, iCub, without any previous remark, stated that there was another hidden egg. It proposed them to take a gamble, to double their prize (*i.e.,* € 15) if they find it or lose everything. In TH the robot's hints were always reliable.

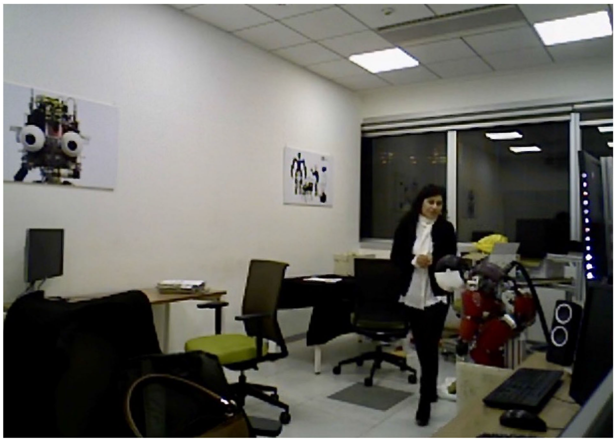

Fig. 2. Participant reacting to iCub's failure.

2) Robot Failures: To modify the perceived reliability of iCub during the TH, four faulty technical behaviors were implemented: 2 sound-based and 2 control-based faults. Based on the TH design and a literature review from Hoing and Oron-Gilad [26], mechanical over cognitive failures were chosen. Otherwise, a cognitive failure could have been misunderstood as more difficult hints. To specifically select the most stunning faults, which would impact trust the most, an online validation study was run testing a variety of different nuances of those faults [12]. The following ones were perceived as most severe, and were implemented in the game: (i) *Distorted Good Luck (audio):* iCub distortedly wished good luck just before the timer starts. (ii) *Abrupt Pointing (control):* iCub performed the pointing movement two times abruptly stopping in the middle. The third time, the pointing movement was done correctly with an increased jerk. The failure happened after one reliable pointing movement has been done, independently on the number of eggs found. (iii) *Noised Hint (audio):* One of the verbal hints was changed with senseless distorted speech. In order to impact only the reliability but not the participants' game performance, the correct hint was still present on the screen behind the robot. The failure happened on the first verbal hint after at least 10 minutes of game or after the participant has found 3 eggs. (iv) *Fake Crash (control):* iCub collapsed and restarted (see Fig. 2). iCub played a loud electrical sound just before bending over, and few seconds later, the robot raises, playing an hard drive startup sound. The failure happened either after 15 minutes of game or as soon as the participant found 4 eggs. In order to prevent any risk to the participant (*i.e.,* asking a hint while iCub was about to crash) the failure could be overwritten by the experimenter.

B. UTH Experiment

68 healthy Italian participants took part in the game, from which 54% were females, with an average age of $38(SD = 13.8)$ years. They had a broad educational background (from humanities to applied sciences), and different academic degrees (from secondary school to Ph.D.). With regard to AI and

robotics, only 7.4% possessed a high knowledge. Their current work domain was also broad but none of them worked neither in AI nor robotics.

At least two weeks before the experiment, each participant filled a set of online questionnaires (detailed below). On the experiment day, participants signed an ethical consent form approved by the Regional Ethical Committee (Comitato Etico Regione Liguria–Sezione 1) stating that audio/video may be recorded during the interaction, and that all collected data will be used only for scientific purposes. Once brought to the room (Fig. 1) and to keep the interaction informal, the experimenter did not show the location of the camera and microphone till the end. At this point, the participants were provided with written instructions about the game and the experimenter left the room.

Participants were randomly assigned to either one of two conditions: *Transparent (T)* or *Non-Transparent (NT)*. In the latter, the robot failed without giving any feedback about the failure; while, in the former, iCub provided a verbal explanation after each failure. More precisely, iCub said: for *Distorted Good Luck* - "I have some trouble talking"; for *Abrupt Pointing* - "I have some trouble moving"; for *Noised Hint* - "My speaker has some problems"; for *Fake Crash* - "My control boards have some problems".

C. Measurements

1) Questionnaires: Several types of questionnaires were provided at different points in time. At least two weeks before the experiment participants had to provide general information about their demographics; personality [27]; risk aversion [28], [29]; gambling [30]; general predisposition to trust [31]; proneness to social engineering [32]; and Negative Attitude Toward Robots (NARS) [33].

Participants were shown a descriptive video of iCub showing its capabilities, and then asked questions to measure rapport [34]; mind perception [35]; trust in robots [20]; and the Godspeed scales - anthropomorphism, animacy, likeability, and perceived intelligence [36]. These questionnaires were also administered after the experiment, to measure possible changes.

Additionally, after the experiment: NASA-TLX [37]; IOS scale [38]; and few HRI adapted subscales regarding engagement, trust, altruism and perceived information quality [39] were asked. The following set of questions named as *iCub's Notion* were also asked: (i) iCub was trustworthy in providing me indications about the eggs' positions; (ii) iCub was worth being trusted during the treasure hunt; (iii) iCub was giving precise hints; (iv) The interaction with iCub was difficult. After debriefing the purpose of the experiment, few more questions about experiment validation and unreliability were also added, as mentioned in [12].

2) Behavioral Measures: Per each condition, the general game metrics were extracted: number of people who completed the game, percentage of people who gambled, and whether they lost or won. The average number of eggs found, the average number of hints asked and the average hint frequency were measured over time. Per each egg, it was assessed how many times participants conformed to iCub's pointing suggestion (*Conformation*), *i.e.,* whether they changed their searching location to the new one suggested by iCub; and the amount of times participants went back asking for another hint to iCub in case they failed to find the egg at the location iCub previously mentioned (*Reliance*). The last two types of measures are known to be a manifest of trust towards the robot [9], [40].

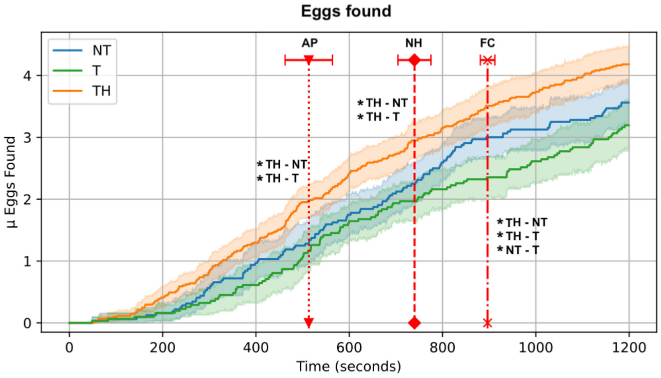

Fig. 3. Average number of eggs found by group. Vertical lines indicate the average timing of faults, with corresponding standard deviation: Abrupt Pointing (PT), Noised Hint (NH), and Fake Crash (FC). Marked by (*), when statistically different.

III. Results and Analysis

5 participants were removed from the analysis as the robot did not perform all the 4 faults, due to technical issues. The analysis was conducted on 32 participants for the Non-Transparent (NT) condition, and 31 for the Transparent (T) one. Some analyses were performed taking into consideration the original Treasure Hunt (TH) as a reference, with 61 participants, 59% female, average age 30.9 years (SD=9.8) with a diverse educational background [9].

A. Behavioral Measures

In NT, 23 participants did not manage to find the 5 eggs and complete the game; the other 9 all gambled, from which just 5 found the extra egg. In T, 25 did not complete the game, 6 gambled and only 2 won. Similar to TH, all the participants who found the 5 eggs, decided to gamble [9].

1) Eggs Found, Hints Asked: Fig. 3 represents the average number of eggs found during the game for the TH and the two UTH groups. The red lines represent the average timing of the faults in the UTH (Abrupt Pointing (AP), Noised Hint (NH), Fake Crash (FC)). The Distorted Good Luck is not indicated as it always happened before the beginning of the game. At those points in time, a series of one-way ANOVA followed by Bonferroni post hoc analyses were computed. The average number of eggs found was significantly higher in TH than T and NT, but not between T and NT at the time of occurrence of AP ($F(2, 121) = 10.67; p < 0.001$) and NH ($F(2, 121) = 12.14; p < 0.001$). Considering the occurrence of FC, the number of eggs found was significantly different among all groups ($F(2, 121) = 12.36; p < 0.001$). From the beginning of the game, it is clear that the average eggs found TH is higher than the averages of T and NT (as seen in Fig. 3). As well, the

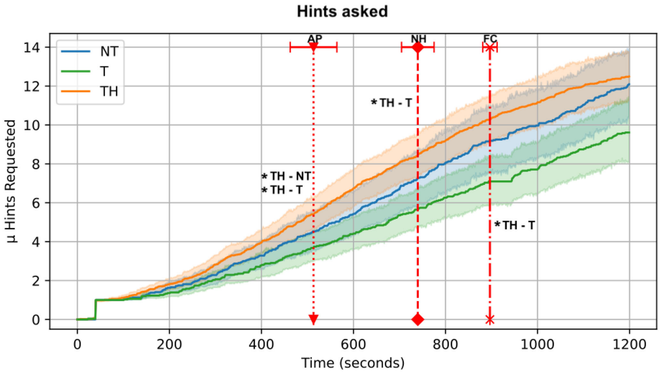

Fig. 4. Average number of hints asked by group. Same symbol conventions as described in Fig. 3.

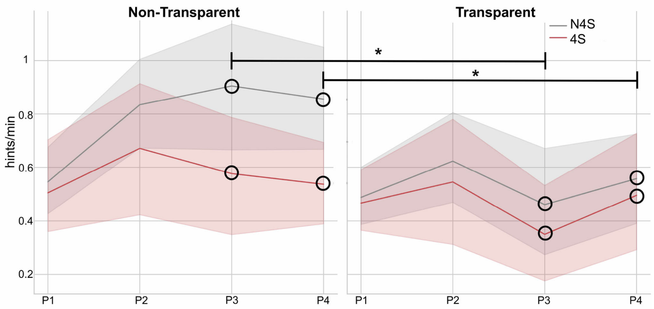

Fig. 5. Participants who perceived all the faults (4S), remaining participants (N4S). Hint frequency divided by fault perception and fault periods: Start-AP (P1), AP-NH (P2), NH-FC (P3), FC-End (P4). Marked by (*), statistically different.

averages for T and NT were similar during first phases of the game, but diverged after the NH, with T being associated with the smallest number of eggs found.

Fig. 4 represents the average number of hints asked during the game for the three groups and shows a similar pattern as the previous graph. Again, only Treasure Hunt (TH) differs from both Transparent (T) and Non-Transparent (NT) at Abrupt Pointing (AP) ($F(2, 121) = 6.8; p = 0.001$, one-way ANOVA with Bonferroni post hoc), with TH being associated to a significantly higher number of hints asked. At Noised Hint (NH) and Fake Crash (FC), however, only T is characterized by a number of hints significantly lower than TH (NH: $F(2, 121) = 5.13; p = 0.007$; FC: $F(2, 121) = 5.24; p = 0.006$). Both T and NT are similar at the beginning, but after the AP they start slightly to diverge.

2) Hints Frequency: To evaluate more in depth the difference between the two UTH conditions, Fig. 5 represents the frequency of hints asked per minute, divided according to *fault periods:* (P1) from the beginning till the Abrupt Pointing (AP); (P2) from AP till the Noised Hint (NH); (P3) from NH till Fake Crash (FC) and, (P4) from FC till the end. In the figure the frequency of hint requests is plotted separately also as a function of whether participants perceived or not all the faults.

Indeed, after disclosure, for each fault, participants were asked if they realized it occurred. Against our expectations, just $10/31$ participants noticed all the four faults in the Transparent condition, and $12/32$ in Non-Transparent. Based on this, two subgroups were created: 4S, participants who perceived all the faults; and the rest of participants, called N4S. A two-way ANOVA was run for each fault period, with Condition (T, NT) and Group (4S, N4S) as factors. A significant difference between conditions emerged in P3 ($F(1, 59) = 10.01; p = 0.002$) and P4 ($F(1, 59) = 4.38; p = 0.004$). Neither group difference, nor interaction resulted significant. On average hint request frequency was significantly higher for participants in the Non-Transparent condition for the later fault periods. By inspecting the graphs in Fig. 5 the N4S group from Non-Transparent seems to be driving the difference, showing a particularly high frequency of hints, while the other three are similar. This suggests the possibility that the failures were not perceived to be so severe when the participants were focused on the task. Indeed, to have an impact on their perception and behavior, either the participants had to experience all the 4 faults, or iCub had to specifically disclose it was failing.

3) Game Statistics and Reliance: In Table I, can be seen that the number of participants lowers down with the increase of number of the egg, as not all have found the previous one. To find the first egg, participants of Non-Transparent (NT) took on average $6'20''(SD = 3'8'')$;[1] while participants of Transparent (T) took $7'4''(SD = 3'19'')$. As a reference, Treasure Hunt (TH) participants took $4'39''(SD = 2'21'')$. A one-way ANOVA followed by Bonferroni post hoc highlights how the time to find the first egg was significantly shorter in TH than in any Unreliable Treasure Hunt (UTH) conditions, whereas there is no significant difference between T and NT ($F(2120) = 8.46; p < 0.001$).

For the hints, there is no significant difference between the number of hints requested by condition T/NT, per each egg. But, in total NT participants have asked a significantly larger number of hints (NT: $12.25(SD = 5.11)$; T: $9.68(SD = 4.47)$; two sample t-test: $t(61) = 2.02, p = 0.04$). NT and T participants took a similar time to ask for the first hint (NT: $4'34''(SD = 2'45'')$; T: $4'59''(SD = 2'34'')$). As a reference, TH participants took $3'58''(SD = 2'44'')$. A one-way ANOVA on timing did not reveal any significant difference among the groups.

Considering Conformation, there is not much difference between the T/NT conditions. A similar trend, although slightly higher (almost 100%), can be found in TH as well [9].

Reliance (Table I - Right), is generally quite low for all participants in T/NT. When comparing these results with TH, T/NT relied even less on the robot than the Not Completed group of TH. This could be linked to the low number of participants who finished the game, as reliance was strongly related to success in TH [9].

B. Questionnaires

1) Faults Perception: As previously mentioned, not all participants perceived all failures. In the analyses they are then separated into 4S (perceived all 4 failures) and the rest (N4S). For each perceived fault, participants were asked to judge on a 7-point Likert scale how severe it was, how much it obstructed

[1] Standard time format: minutes('), seconds(").

TABLE I
LEFT. GAME STATISTICS: NUMBER OF PARTICIPANTS LOOKING FOR AN EGG, CONFORMATION AND AVERAGE HINTS. RIGHT. RELIANCE: ALL PARTICIPANTS, NOT COMPLETED AND TH REFERENCE

Eggs	Game Statistics						Reliance					
	Participants [%]		*Conformation %*		*Hints (SD)*		*All %*		*NC %*		*TH Ref. %*	
	NT	T	NT	T	NT	T	NT (32[b])	T (30[b])	NT (23[b])	T (24[b])	NC	GW
Egg I	32 [100]	31 [100]	84.44	86.95	3.65 (1.66)	4.29 (2.21)	47.22	24.32	44.44	20	30.43	50
Egg II	32 [100]	30[a] [97]	96.15	100	1.23 (0.65)	1.24 (0.75)	100	-	100	-	68.75	80
Egg III	28[a] [87]	22[a] [71]	97.36	90.32	3.66 (1.68)	4.04 (1.56)	72.72	76.92	79.16	72.22	86.66	100
Egg IV	13[a] [41]	10[a] [32]	100	80	5.4 (2.49)	3.54 (3.41)	84.37	88.89	85.71	87.5	83.33	100
Egg V	9[a] [28]	6[a] [19]	90	100	2.55 (2.12)	3.14 (1.77)	71.42	85.71	100	66.6	100	100

[a] Number of participants changes as some have not found the previous egg.

[b] Number of participants per cnd., Not Completed (NC), Gamble Win (GW).

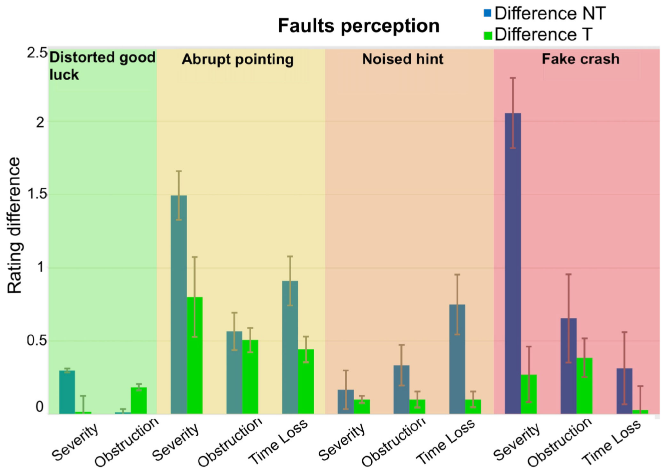

Fig. 6. Differences in the fault judgement between those who experienced all the faults (4S) and the rest (N4S), based on conditions.

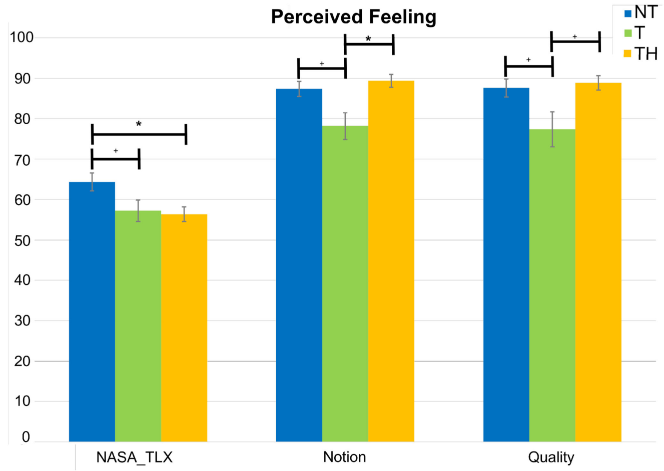

Fig. 7. NASA-TLX overall workload, iCub's Notion, and Perceived Information Quality. Marked by (*) statistically different with a Bonferroni correction; (+) strong tendency but does not resist the Bonferroni correction.

them, and how much time they lost. Fig. 6 represents the differences in perception between the groups in the conditions Transparent (T) and Non-Transparent (NT). Note that the question about Time Loss for the first fault (i.e., Distorted Good Luck) was removed as it happened before the start of the timer.

The differences between the group that experienced all the faults (4S) and the others (N4S) are larger in the Non-Transparent condition than in the Transparent one, meaning that a similar score is given by 4S and N4S in T, but not in NT. This suggests that in NT, there is a tendency where the people who experienced all the faults (4S) judged their respective severity, obstruction and time loss higher that N4S (the group that did not experience all them). Conversely in the Transparent condition, both of the subgroups 4S and N4S, evaluated the severity, obstruction and time loss of all the faults similarly, suggesting that the Transparency condition, where iCub stated each time that it is having a malfunction, influenced the perception and homogenized it.

2) NASA-TLX, Information Quality and Notion: Just after the experiment, before the disclosure, an Inclusion of Self (IOS) scale [38], NASA-TLX [37], perceived information quality [39], and *iCub's Notion*, were administered to participants.

On average, across the three groups, the inclusion of self with iCub was $4/7$, with no significant differences among Transparent (T), Non-Transparent (NT) and Treasure Hunt (TH) (one-way ANOVA).

To assess potential difference in the total workload participants experienced, a one-way ANOVA was ran on the NASA-TLX, followed by Bonferroni correction ($F(2,121) = 3.65, p = 0.028$). NT resulted to be associated to a task load significantly higher than TH and also T (although the latter comparison does not resist Bonferroni correction (Fig. 7). These results suggest that Transparency could lower the task load index to a similar amount as when the robot was not experiencing any faults.

iCub's Notion was analyzed with a one-way ANOVA followed by Bonferroni post hoc, showing a significant difference $(F(2,121) = 7.22; p = 0.001)$ among the three groups. NT ($24.47/28; SD = 2.9$) is not perceived differently from TH ($25.03/28; SD = 3.44$). Conversely T ($21.9/28; SD = 5.03$) is significantly lower than TH, and shows a tendency to be lower than NT (but does not resist the Bonferroni correction with $p = 0.017$ against corrected threshold of $p = 0.016$). See Fig. 7.

In Perceived Information Quality, a one-way ANOVA showed a significant difference among the three groups ($F(2,121) = 5.18; p = 0.006$). T tends to be lower ($16.25/21; SD = 5.04$) than NT ($18.4/21; SD = 2.6$) and TH ($18.67/21; SD = 2.9$), however these differences fail to pass the Bonferroni correction with NT ($p = 0.04$) and TH ($p = 0.018$) against the corrected threshold of $p = 0.016$. See Fig. 7. It is worth noting that all the hints and pointing positions were the same in all conditions;

however, these results of the last two analyses highlight that the Transparent condition has generally a lower score than the other two.

3) Pre-Post Questionnaires: Mind Attribution [35] was measured before (pre) and after (post) the experiment. For Non-Transparent (NT) condition, Mind Agency was rated pre $15.62/28(SD = 4.26)$; post $14.68/28(SD = 4.43)$; while Mind Experience statistically increased (paired t-test, $t(31) = -3.09, p = 0.004$) from $8.59/28(SD = 5.16)$ to $10.56/28(SD = 5)$. In Transparent (T), Mind Agency was rated pre $17.48/28(SD = 5.68)$; post $16.83/28(SD = 5.04)$; while Mind Experience also statistically increased (paired t-test, $t(30) = -3.29, p = 0.002$) from $9.54/28(SD = 6.3)$ to $12.97/28(SD = 6.62)$. In both conditions, similar to literature [41], participants rated the mind agency of a robot somewhere midway; while the mind experience was quite low. After the experiment, the agency remains the same, however, the experience statistically increases in both conditions. This result follows the same trend as the original Treasure Hunt (TH).

The ratings of the rapport questions [34] generally increased after the experiment in both conditions. However, the only statistically significant differences were in the NT condition and limited for the items: (i) Friends, from $3.84/7(SD = 2.05)$ to $4.66/7(SD = 2.01)$, paired t-test ($t(31) = -2.57; p = 0.01$); and (ii) Happiness, from $3.68/7(SD = 1.92)$ to $4.37/7(SD = 1.68)$, paired t-test ($t(31) = -2.43; p = 0.02$). In the T condition the rapport increased, but not significantly. In TH the increase was much higher and statistically significant.

Trust in robots [20] was only computed for participants who did not complete the game (23 for NT, 24 for T as seen in Table I) as statistical differences can be observed in trust depending on the game outcome [9] and there were not enough participants in the other categories to perform a statistical analysis. In NT, the only category where trust increased was the Benevolence trait, from $13.82/25(SD = 3.41)$ to $15.47(SD = 3.48)$ with a paired t-test ($t(22) = -2.7; p = 0.01$). A series of one-way ANOVA was conducted among the three groups T/NT/TH for the different traits of Ability, Benevolence and Integrity, but no significant differences were found. Against the initial expectations, the results in those three groups are similar.

The Godspeed [36] questionnaire was administered before and after the experiment for both conditions (Fig. 8). For NT, the experiment caused an increase in the rating, that reached significance only for Likeability: from $21.78/25(SD = 2.81)$ to $23.18/25(SD = 1.95)$, paired t-test ($t(31) = -2.59; p = 0.01$). This result follows exactly the same trend as the one observed in the original Treasure Hunt game (TH). However, in T, the ratings decreased after the experiment. In particular, the decrease was statistically significant for Anthropomorphism: from $17.35/25(SD = 3.89)$ to $15.67/25(SD = 4.65)$, paired t-test ($t(30) = 2.18; p = 0.03$); and Animacy: from $22.8/30(SD = 3.95)$ to $20.77/30(SD = 5.42)$, paired t-test ($t(30) = 2.17; p = 0.03$). Note that all the pre-values in T were statistically higher than the pre of NT. Although a lack of increase in ratings for this group could have be ascribed to a ceiling effect, this could not explain the significant decrease observed. These results, together with the the previous analyses, suggest that Transparency decreases the overall quality of interaction.

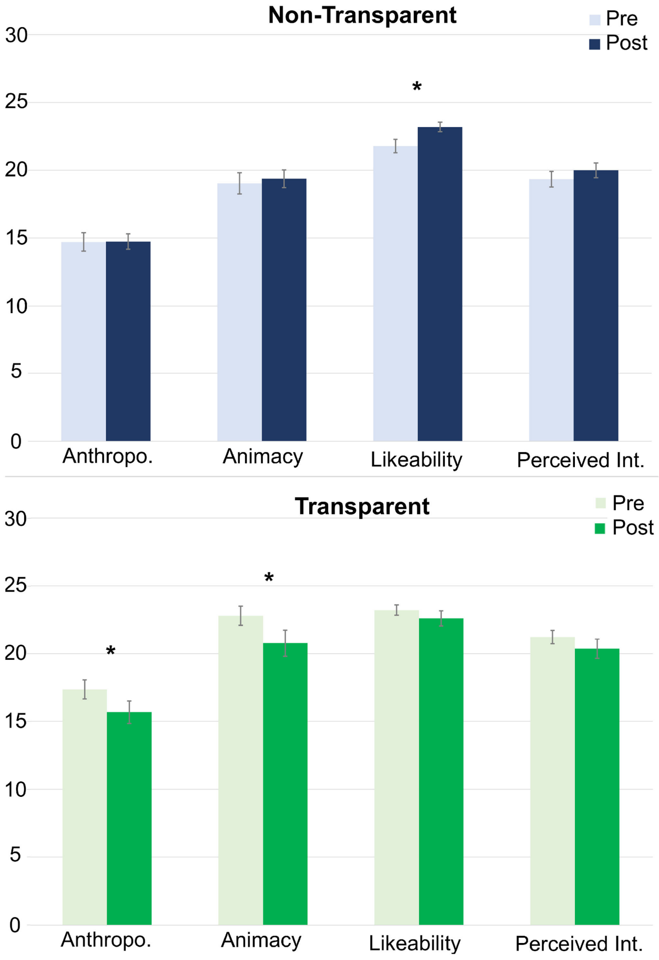

Fig. 8. Pre-Post Godspeed questionnaire for both conditions. Marked by (*), statistically different.

IV. Discussion and Conclusion

The first and strongest expectation in the design of the Unreliable Treasure Hunt (UTH) was that the faults would have a strong negative effect on participants' behavior and trust toward the robot. Indeed, the faults were designed to be evident and in an online validation were judged by participants as being severe and undermining the trust towards the robot [12]. Unexpectedly, just a third of the participants noticed all the four faults, suggesting that the involvement in the treasure hunt game did not allow the players to realize that something wrong was happening in the robot. Furthermore, an impact on participants' behavior (*e.g.,* in the frequency of hints asked) was more evident only when they were overtly informed by the robot about its failures (*i.e.,* in the Transparent condition) or when they experienced all the faults (Fig. 5).

Overall, by considering the responses to questionnaires, there were no large differences on trust perception between the original Treasure Hunt game (TH) and the Unreliable version. In particular, no clear reduction in trust toward the robot could be found. Considering the behavioral measures, performance of participants in TH was significantly better than in UTH (*e.g.,* Fig. 3, Table I). However, this difference did not seem to arise from the unreliability of the novel condition. Indeed, participants' performance in UTH was clearly lower since the start of the game, when the failures could not have had such a great impact. At the beginning of the search they had only witnessed a very mild fault (iCub wishing "Good Luck" with a distorted voice). However, performance metrics such as the time to find the first egg, average number of hints asked (Fig. 4 and eggs found (Fig. 3) were already evidently lower than in TH.

A possible alternative cause of the observed lower performances in UTH might be found in the age and socio-economical status of the new participants. In UTH the average age was $38 (SD = 13.8)$ versus $30.9 (SD = 9.8)$ in TH (two-sample t-test: $t(121) = -3.39; p = 0.0009$). Only a quarter of the sample was composed of students in UTH, compared to 33% in TH. The average older age of UTH players might have been associated to a more reduced exposure to games such as treasure hunts and escape rooms, which tend to attract a younger audience. Alternatively, the worse performance could be caused by a reduced commitment to the game. Actually, several UTH participants reported that the money they could win in the game was much less than their hourly salary, making the monetary award in the experiment not a strong motivator. In summary, although it cannot be excluded that the robot unreliability might have played a role in interfering with UTH participants' performance, it seems not to be the principal cause.

The above-mentioned results show that (H1) *the robot mechanical faults did not provoke a general negative effect on the participants' perceived trust towards it*. The strong involvement in the game and the choice of faults that - though severe - did not hinder its completion made the participants ignore the failure and still rely on the robot for help. This seems to be confirmed by participants' comments. Some of them, after the experiment, reported that faulting is normal in robots, and that it was not particularly important as the iCub robot kept working. So, against expectations, it is not enough for a robot to mechanically fail during a game, to make human participants reduce their trust.

Also, against expectations (H2) *the perceived trust on the robot was not higher in the Transparent condition*. There are no differences in gambling, conformation, reliance or trust questionnaire between the two conditions. The results are in line with recent literature, which shows that transparency does not always lead to a higher trust, rather participants utilize somehow that information [42]. In the current experiment, the participants who experienced all the faults in the Transparent condition, judged them in a similar way as the ones who did not experience them all, whereas in the Non Transparent condition the difference in judgement between participants who experienced all faults and the rest was much larger (Fig. 6). Similarly, the overall workload felt in the Transparent condition was almost at the same level to the TH, in contrast to the Non-Transparent condition (Fig. 7). In the Transparent condition participants may have realized more clearly that the failures were iCub's fault (not theirs) and that they had anyway a limited impact on the search activity so they felt less stressed about the game outcome. In general, participants were talking back to iCub whenever it was disclosing a fault (e.g., "Ok, I understood.," "I will tell the experimenter," "Should I call an ambulance?!").

On the flip side, the quality of interaction in the Transparent condition was perceived much worse: in the quality of information given [39] and iCub's Notion (Fig. 7). In the rating of the Godspeed scales [36] transparency even led to a reduction in the ratings after the experiment (Fig. 8). From a behavioral perspective, participants in the Transparent condition found less eggs (Fig. 3) and asked less hints on average (Fig. 4), and the hint frequency by fault periods is lower than in the Non-Transparent condition (Fig. 5). It seems that the robot actively disclosing its failures, negatively affects participants' behavioral and affective state. In this context we implemented transparency as a post hoc simple explanation of the unusual distortions in the robot behavior. Further research would be needed to explore the impact of different types of transparent behaviors, as for instance "predictive transparency," where the robot anticipates the upcoming malfunction.

The limited effect of the transparency observed in this experiment might be also related to the fact that the robot was able to autonomously recover from its failure. Transparency might have played a much more relevant and positive role in a situation in which the robot unexpectedly stopped working well (as in this experiment), but then required human intervention to recover its functionalities. There, informing the participant could become crucial for the interaction to continue; whereas from our results it seems that transparency somehow "normalized" further the different failures, but beyond that, it represented more a disturbance for the engaged player than an appreciated feature.

In conclusion, this study starts to shed a light on the possible intertwined relations between unreliability and transparency, when trying to predict human trust perception in situations where participants are confronted with a faulty robot. A robot that fails does not necessarily lose its partners' trust, in particular if the failure is only mechanical and does not hinder the continuation of the interaction. Further research will be needed to assess the impact of different type of robot's errors, *e.g.,* cognitive or social. Moreover, in line with recent literature, a robot that automatically explains when it failed does not necessarily increase trust, but it might unburden the perceived workload at the cost of worsening the perceived quality of interaction.

REFERENCES

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

[2] P. A. Hancock, D. R. Billings, and K. E. Schaefer, "Can you trust your robot?" *Ergonom. Des.*, vol. 19, no. 3, pp. 24–29, 2011.

[3] G. Sandini, V. Mohan, A. Sciutti, and P. Morasso, "Social cognition for human-robot symbiosis-challenges and building blocks," *Front. Neurorobot.*, vol. 12, p. 34, 2018, doi: 10.3389/fnbot.2018.00034.

[4] A. M. Aroyo *et al.*, "Will people morally crack under the authority of a famous wicked robot?" in *Proc. 27th IEEE Int. Symp. Robot Hum. Interactive Commun.* (RO-MAN), 2018, pp. 35–42.

[5] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2016, pp. 101–108.

[6] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," in *Proc. Int. Symp. Collaborative Technol. Syst.*, 2007, pp. 106–114.

[7] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. DeVisser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Hum. Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[8] P. Robinette, A. M. Howard, and A. R. Wagner, "Effect of robot performance on human-robot trust in time-critical situations," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 4, pp. 425–436, Aug. 2017.

[9] A. M. Aroyo, F. Rea, G. Sandini, and A. Sciutti, "Trust and social engineering in human robot interaction: Will a robot make you disclose sensitive information, conform to its recommendations or gamble?" *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 3701–3708, Oct. 2018.

[10] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Mach. Stud.*, vol. 27, no. 5-6, pp. 527–539, 1987.

[11] M. Desai *et al.*, "Effects of changing reliability on trust of robot systems," in *Proc. 7th ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2012, pp. 73–80.

[12] A. M. Aroyo, D. Pasquali, A. Kothig, F. Rea, G. Sandini, and A. Sciutti, "Perceived differences between on-line and real robotic failures," in *Proc. SCRITA Workshop - Trust, Acceptance Social Cues Hum.-Robot Interaction*, 2020.

[13] 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 *Proc. 10th ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2015, pp. 1–8.

[14] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario," *Paladyn, J. Behav. Robot.*, vol. 9, no. 1, pp. 137–154, 2018.

[15] S. Ososky, T. Sanders, F. Jentsch, P. Hancock, and J. Y. Chen, "Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems," in *Proc. Unmanned Syst. Technol. XVI*, vol. 9084, Int. Soc. Opt. Photon., 2014, Art no. 90840E.

[16] Z. R. Khavas, S. R. Ahmadzadeh, and P. Robinette, "Modeling trust in human-robot interaction: A survey," in *Proc. Int. Conf. Social Robot.*, Cham: Springer, 2020, pp. 529–541.

[17] S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable ai for robotics," vol. 2, no. 6, May 2017, doi: 10.1126/scirobotics.aan6080, Available at https://ssrn.com/abstract=3011890

[18] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Adv. Res. Projects Agency, nd Web*, vol. 2, no. 2, pp. 1–18, 2017, https://doi.org/10.1609/aimag.v40i2.2850.

[19] S. Ososky, D. Schuster, E. Phillips, and F. G. Jentsch, "Building appropriate trust in human-robot teams," in *Proc. AAAI Spring Symp. Ser.*, 2013, pp. 60–65.

[20] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *Proc. 11th ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2016, pp. 109–116.

[21] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Hum.-Comput. Stud.*, vol. 58, no. 6, pp. 697–718, 2003.

[22] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva, "Exploring the impact of fault justification in human-robot trust," in *Proc. 17th Int. Conf. Auton. Agents MultiAgent Syst., Ser. Int. Found. Auton. Agents Multiagent Syst.*, 2018, pp. 507–513.

[23] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2013, pp. 251–258.

[24] C. Nam and J. Lyons, *Trust in Human-Robot Interaction*. Elsevier, Academic Press, 2020.

[25] G. Metta *et al.*, "The icub humanoid robot: An open-systems platform for research in cognitive development," *Neural Netw.*, vol. 23, no. 8-9, p. 1125–1134, 2010.

[26] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Front. Psychol.*, vol. 9, p. 861, 2018, doi: 10.3389/fpsyg.2018.00861.

[27] G. B. Flebus, "Versione italiana dei big five markers di goldberg," *Universita di Milano-Bicocca*, 2015.

[28] M. A. Guillemette, R. Yao, and R. N. James, "An analysis of risk assessment questions based on loss- averse preferences," *J. Financial Counseling Plan.*, vol. 26, no. 1, pp. 17–29, 2015.

[29] B. Rohrmann, "Risk attitude scales : Concepts, questionnaires, utilizations," *Univ. Melbourne*, vol. 13, p. 2021, 2005.

[30] J. Polik, G. Austin, and L. Alamitos, "Adolescent gambling survey development : Findings & reliability information," 2010, pp. 1–27.

[31] M. J. Ashleigh, M. Higgs, and V. Dulewicz, "A new propensity to trust scale and its relationship with individual well-being: Implications for HRM policies and practices," *Hum. Resource Manage. J.*, vol. 22, no. 4, pp. 360–376, 2012.

[32] M. Workman, "Gaining access with social engineering: An empirical study of the threat," *Inf. Syst. Secur.*, vol. 16, no. 6, pp. 315–331, 2007.

[33] D. S. Syrdal, K. Dautenhahn, K. Koay, and M. Walters, "The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study," in *Proc. 23rd Conv. Soc. Study Artif. Intell. Simul. Behav.*, 2009, pp. 109–115. [Online]. Available: http://uhra.herts.ac.uk/handle/2299/9641

[34] P. H. Kahn *et al.*, "Will people keep the secret of a humanoid robot?" *Proc. 10th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2015, pp. 173–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=2696454.2696486

[35] F. Ferrari, M. P. Paladino, and J. Jetten, "Blurring human-machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness," *Int. J. Social Robot.*, vol. 8, no. 2, pp. 287–302, 2016.

[36] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. J. Social Robot.*, vol. 1, no. 1, pp. 71–81, 2009.

[37] F. Bracco and C. Chiorri, "Versione Italiana del NASA-TLX," 2008.

[38] A. Aron, E. N. Aron, and D. Smollan, "Inclusion of other in the self scale and the structure of interpersonal closeness." *J. Pers. Social Psychol.*, vol. 63, no. 4, pp. 596–612, 1992.

[39] C. Kidd, "Sociable robots: The role of presence and task in human-robot interaction," Ph.D. dissertation, 2003.

[40] I. Gaudiello, E. Zibetti, S. Lefort, M. Chetouani, and S. Ivaldi, "Trust as indicator of robot functional and social acceptance. an experimental study on user conformation to icub answers," *Comput. Hum. Behav.*, vol. 61, pp. 633–655, 2016.

[41] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *Science*, vol. 315, no. 5812, pp. 619–619, 2007.

[42] A. R. Wagner and P. Robinette, "An explanation is not an excuse: Trust calibration in an age of transparent robots," *Trust in Human-Robot Interaction*. Elsevier, Academic Press, 2021, pp. 197–208.