“An error occurred!” - Trust repair with virtual robot using levels of mistake explanation

Kasper Hald, Katharina Weitz, Elisabeth André, Matthias Rehm

Angaben zur Veröffentlichung / Publication details:

Hald, Kasper, Katharina Weitz, Elisabeth André, and Matthias Rehm. 2021. “‘An error occurred!’ - Trust repair with virtual robot using levels of mistake explanation.” In *HAI’21: Proceedings of the 9th International Conference on Human-Agent Interaction, Virtual Event, Japan, November 9 - 11, 2021*, edited by Kohei Ogawa, Tomoko Yonezawa, Gale M. Lucas, Hirotaka Osawa, Wafa Johal, and Masahiro Shiomi, 218–26. New York, NY: ACM. https://doi.org/10.1145/3472307.3484170.

"An Error Occurred!" - Trust Repair With Virtual Robot Using Levels of Mistake Explanation

Kasper Hald*
Dept. of Architecture, Design & Media Technology, Aalborg University
Aalborg, Denmark
kh@create.aau.dk

Katharina Weitz*
Elisabeth André
Lab for Human-Centered Artificial Intelligence, University of Augsburg
Augsburg, Germany
katharina.weitz@uni-a.de
elisabeth.andre@uni-a.de

Matthias Rehm
Dept. of Architecture, Design & Media Technology, Aalborg University
Aalborg, Denmark
matthias@create.aau.dk

ABSTRACT

Human-robot collaboration in industrial settings is an expanding research field in robotics. When working together, robot mistakes are an important factor to decrease trust and therefore interferes with cooperation. It is unclear whether explanations help to restore human-robot trust after a mistake. In our study, we investigate whether system explanations as a trust-repairing action after a robot makes a mistake in a collaborative task is helpful. Our pilot study revealed that users are more interested in solutions to errors than they are in just why the error happened. Therefore, in our main study, we evaluated three levels of mistake explanations (no explanation, explanation, and explanation with solution) after a robot in VR made a mistake in executing a shared objective. After testing with 30 participants we found that the robot making a mistake significantly affects trust toward the robot, compared to it completing the task successfully. While participants found the explanations helpful to trust or distrust the robot, the levels of the explanation did not lead to an increase in trust towards the robot after a mistake. In addition, we found no significant impact of explanations on self-efficacy and the emotional state of the participants. Our results show that explanations alone are not sufficient to increase human-computer trust after robot mistakes.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; *Empirical studies in interaction design*; *Virtual reality*.

KEYWORDS

human-robot collaboration, human-robot trust, XAI, virtual reality, robot mistakes, proximity

ACM Reference Format:
Kasper Hald, Katharina Weitz, Elisabeth André, and Matthias Rehm. 2021. "An Error Occurred!" - Trust Repair With Virtual Robot Using Levels of Mistake Explanation. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21), November 9–11, 2021, Virtual Event, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3472307.3484170

1 INTRODUCTION

The collaboration between humans and machines in the industrial setting is becoming more and more realised due to the enormous progress in the fields of robotics and machine learning. Introducing collaborative robots in the manual production can potentially help relieve the workers of strenuous and repetitive tasks. Through the use of machine learning methods like Deep Reinforcement Learning, it is possible for robots to interact autonomously in industry tasks and to adapt dynamically to the demands placed on them. In addition, an intuitive usage and interaction by humans become more and more common. However, the more natural the handling of robots in industry becomes, the more demands humans place on them. If these demands are not met, human-robot collaboration (HRC) can be disrupted. In addition to reduced trust and frustration, this can have serious consequences such as accidents and production losses [7, 12]. To enable successful HRC it is important that we maintain human-robot trust (HRT), especially when working with a robot at close proximity. To this end, we investigate the capabilities of system communication with the human collaborator to perform trust-repair through explanation in cases where the robot makes a mistake during the execution of a shared objective. We base the research on the context of a shared task where the human and robot have to move a collection of objects within a shared tabletop work space. To integrate system communication with non-obstructive output modalities we base the design of the communication system on projection-based augmented reality (AR), so that messages can be displayed directly on the work surface. To sum up, we investigate how we can use mistake explanations after a robot mistake as a trust-repairing action in order to maintain trust during close-proximity collaboration. Rather than implementing the robots' communication system using real hardware, we test our prototype iterations using computer-generated demonstrations and virtual reality (VR) testing environments[1]. With our work, we make the following contributions:

- With our pilot study we give novel insights about requirements and expectations of end-users towards robot explanations after mistakes.

*Both authors contributed equally to this research.

[1] The presented studies in this paper as well as the gathered data have been collected with respect of the GDPR regulations

- We present a VR setup to research robot-mistakes in close-proximity collaboration-tasks.
- Our results about the impact of different levels of explanations after robot-mistakes on trust, explanation satisfaction, self-efficacy, and emotional state of end-users gain new insights regarding explainable AI (XAI) in HRC.
- We discuss the challenges using explanations in HRC and how our findings are useful for researchers to design XAI in HRC.

2 STATE OF THE ART

2.1 Human-Robot Trust & Robot Mistakes

De Visser et al. [7] use a definition of trust in the context of HRC as the human's willingness to engage in a situation characterized by vulnerability with another party based on their expectation toward that party. In this context the other party is the robot. In a meta-analysis Hancock et al. [12] categorized the constructs that affect the operator's perception of the robot into human-related, robot-related, and environmental factors. The robot-related factors are further split into performance-based and attribute-based factors, covering how the robot performs or behaves and how the robot looks or where it is, respectively. Looking at performance-based robot-related factors, reliability, dependability, and predictability have significant effect HRT. They also outline the importance of appropriate trust levels toward the robot in HRC, as too much trust may lead to dangerous situation as a result of misuse, whereas too little trust may lead to the robot not being utilized optimally. Schaefer [33] developed two HRT scales based on the operator's perception of the robot's characteristics, performance, predictability, and more. The long scale has 40 questions while the shorter version has 14 questions. Kessler et al. [21] compared these scales to a standardised scale of trust in automation with conflicting results, suggesting that the two scales evaluated different factors, making them not interchangeable.

In testing robot dependability and its effects on trust, Salem et al. [32] found that a home companion robot would be perceived as less trustworthy after making a mistake, even though the mistake did not significantly affect participants' willingness to follow the robot's instruction. In addition to the factors outlined by Hancock et al. [12], HRT has also been shown to be affected by the general transparency of the system controlling the robot. Boyce et al. [5] compared three transparency conditions in a simulation. Higher levels of transparency yielded higher trust measured using a modified automation trust scale. Due to the scale used one has to consider whether the trust pertains to the simulated robot or the communication system. Comparing decision explanations for a robot in a simulated reconnaissance mission, Wang et al. [38] found that low-ability robots gained more trust from explanation, as opposed to no explanation, whereas high-ability robots did not gain trust from them. When testing a robot that would assign blame after a mistake occurred, Kaniarasu and Steinfeld [20] found that people would be annoyed when the robot blamed them, but they trusted the robot less if it kept blaming itself.

On the importance of the presence of the robot, as we are testing using VR simulations, both Wainer et al. [36] and Bainbridge et al. [2] compared a co-located robot with a remote robot presented on a screen, and both found that the co-located robot was significantly favoured. However, Duguleana et al. [9] found, when comparing HRI with a real robot and with one presented in immersive VR, participants reported high engagement toward the virtual robot and rated it at 7.8 out of 10 in realism, relative to the real robot.

2.2 Explanations in Human-Robot Interactions

Evidence suggests that a lack of transparency, with respect to the decisions of an autonomous agent, might have a negative impact on the trustworthiness of a system, which in return hurts the overall user experience [16, 35].

The reemerging research field of explainable artificial intelligence (XAI) [11] investigates approaches to address this problem. Current research on XAI is mainly dealing with methods to explain the decisions of deep neural networks (e.g., [15, 19, 39]). Various promising approaches have meanwhile been developed for these use-cases (the interested reader is referred here to works of e.g., [1, 30]). In the field of human-robot interaction, different XAI approaches are discussed to gain insights in behaviour and goals of robots (e.g., the work of [34]).

Alongside the question of how explanations can be generated, the research field of XAI is also concerned with the question of how explanations can be communicated to users. In particular, communicating explanations to end-users is a challenge here, as they need to interact with the system (e.g., a robot) but have no knowledge how the system works. The work of Wang et al. [37] shows that explanations to end-users about a well working robot increases transparency, trust, and performance in human-robot interactions. But robots also make mistakes and are not free of errors. When an error occurs, without an explanation end-users are often unable to understand how the error arose, how to fix it, and how to avoid it in the future. This leads to performance losses as well as distrust [18]. But even with explanations, less accurate autonomous systems lead to a decrease of trust in robots abilities, and success of the task [37]. Therefore it is critical to investigate, whether it is possible to repair trust in the system and if so, which aspects of an explanation are relevant to increase trust.

3 PILOT STUDY

The scope of our work is to investigate HRT in an interaction scenario in which the robot makes a mistake. In the pilot study we conducted, we first wanted to investigate whether different *explanation modalities* (i.e., *textual* or *auditory*) are preferred by participants. We presented the participants with videos of a virtual robot performing a task of sorting bottles of different shapes at either side of a table. The setup is illustrated in Figure 1. In addition, we varied the *type of error*:

- *Colour vision error*: To illustrate the colour vision error, the robot shown is moving a bottle of incorrect shape. The explanation given was:"A computer vision error occurred. The system did not successfully distinguish the shapes in the current lighting conditions."
- *Calibration error*: Here the robot knocked over one of the bottle while moving them. The explanation given was: "A calibration error occurred. The motion planner did not properly compensate for the robot's momentum."

The pilot study was conducted to guarantee, that the different explanation modalities and types of error did not significantly differ in their impact on trust. Furthermore, we wanted to gain insights whether the given explanations were sufficient enough and whether/which additional information participant find helpful. In more detail, we formulated the following hypotheses:

- **H1: Robot Performance & Likeability**: The rating of robot performance and likeability will differ between the no-error and the two error conditions, where the ratings for the no-error robot will be higher.
- **H2: Explanation Quality**: After being presented with a robot mistake in videos of a virtual robot and a given modality of explaining the mistake, the user can describe the mistake accurately.
- **H3: Modality of Explanation**: There will be no difference between the modality of explanation (i.e., textual and auditory) regarding likeability, performance, trustworthiness, and understanding of the robot.
- **H4: Type of Robot Error**: There will be no difference between the types of mistake (i.e., calibration error and colour vision error) regarding likeability, performance, trustworthiness, and understanding of the robot.

To answer these hypotheses, we used a between-subjects design for the modality of explanation (i.e., textual or auditory), meaning that every participant saw one of the explanation modalities. For the two different robot mistakes (i.e., colour vision error and calibration error), we used a within-subjects design. Here, every participant saw both types of errors during the study[2].

3.1 Procedure

The pilot study took place as an online questionnaire. Within this questionnaire, the participants were shown a series of videos of a virtual robot modeled after the Rethink Robotics Sawyer[3] model. This robot had the task of sorting bottles at either end of a table based on their shape.

- **First video:** The first video showed the robot successfully completing the sorting task, switching the positions of two bottles, so that two round-based bottles are on the left side of table and two square-based bottles are on the right. Then, the participants rated the performance of the robot and their impression of the robot. They were then asked to briefly describe the robot, its behaviour, and the task it was performing.
- **Second video:** The second video showed the robot performing the same task again, but making a mistake (i.e., computer vision or calibration error). The participants then answered the same questions about the robot's performance and their impression. After that, they were asked to briefly describe what the difference was from the previous video.
- **First Explanation:** Subsequently, they were shown an explanation of the previously seen mistake (i.e., textual or auditory explanation). The textual explanation modality being shown in Figure 1. Next, the participants had to answer several questions about the explanation shown.
- **Third video:** After answering these questions, they were shown a third video, of the robot making the other type of mistake.
- **Second explanation:** Here, again, an explanation was shown to them afterwards and the participants had to evaluate it.

At the end of the online study, participants had to provide some personal information about themselves.

3.2 Evaluation Methods

To gain insights of the user's impressions regarding the robot errors and the explanation modalities, we used different scales.

Performance. To evaluate the perceived robot performance, we asked the participants after every video to rate the performance of the robot, using a 7-point Likert scale (1= not good, 7= very good).

Likeability. Similar to the measurement of the perceived robot performance, we asked the participants after the no-error video as well as after each explanation, how much they liked the robot and if they wanted to work with the robot[4] (7-point Likert scale; 1= not at all, 7= totally).

Explanation Quality. To measure the quality of the presented explanations, we used two items of the Explanation Satisfaction Scale (ESS)[17]. Here we asked the participants (1) whether the explanations helped to trust the robot and (2) whether they helped to understand how the robot worked (5-point Likert scale; 1= I disagree strongly, 5= I agree strongly). In addition, we asked two general yes / no questions regarding the explanations, i.e., "Have you learned anything because of the explanation?" and "Was the explanation easy to understand?". We also asked for free-form feedback. Here we wanted to know from the participants which parts of the explanation were easy/not easy to understand, whether they would have needed more/additional information and which one and why the explanation was not helpful (i.e., when participants answered the "Have you learned anything because of the explanation?"question with "no").

In addition, at the end of the pilot study we collected personal information (e.g., age, gender) from participants as well as their knowledge and attitudes toward AI and XAI.

3.3 Participants

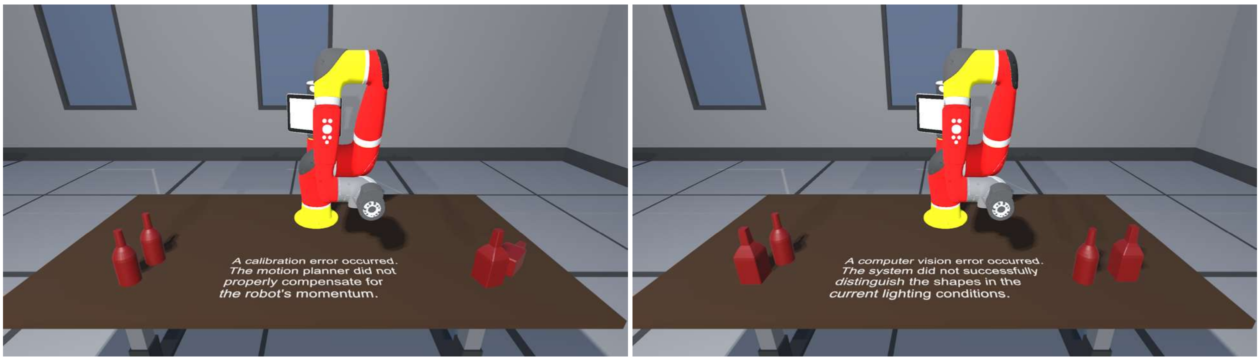
In our pilot study, 20 people between 21 and 54 years (M = 29.3, SD = 7.47) participated. 11 of them were male, 9 were female. All participants had heard about the term AI, but only 9 of them had heard about XAI.

3.4 Results

3.4.1 Rating of Robot Performance & Likeability.

To answer H1, we compared the variables likeability and performance between the no-error robot and the two error conditions. For this, we conducted paired t-tests. Here, the performance of the no-error robot was perceived significantly higher compared to the calibration error robot, $t(19) = 9.20$, $p = < .001$, $d = 2.06$ (large effect) as well as the colour vision error robot, $t(19) = 9.11$, $p = < .001$, $d = 2.04$ (large effect). Similar results were found for the likeability. The no-error

[2] We randomized the order of the presented errors to control for sequence effects
[3] https://www.rethinkrobotics.com/sawyer/
[4] We calculated a mean value from both ratings


A calibration error occurred.
The motion planner did not
properly compensate for
the robot's momentum.
A computer vision error occurred.
The system did not successfully
distinguish the shapes in the
current lighting conditions.

Calibration error explanation

Colour vision error explanation

Figure 1: Textual explanation modality. Two robot errors were explained during the pilot study: a *calibration error* (left) and a *computer vision error* (right).

robot was liked significantly more compared to the calibration error robot, $t(19) = 4.27$, $p < .001$, $d = 0.95$ (large effect) as well as the colour vision error robot, $t(19) = 6.06$, $p < .001$, $d = 1.35$ (large effect). These results are shown in Figure 2. Therefore, the results support our H1 that ratings for the the no-error robot were higher.

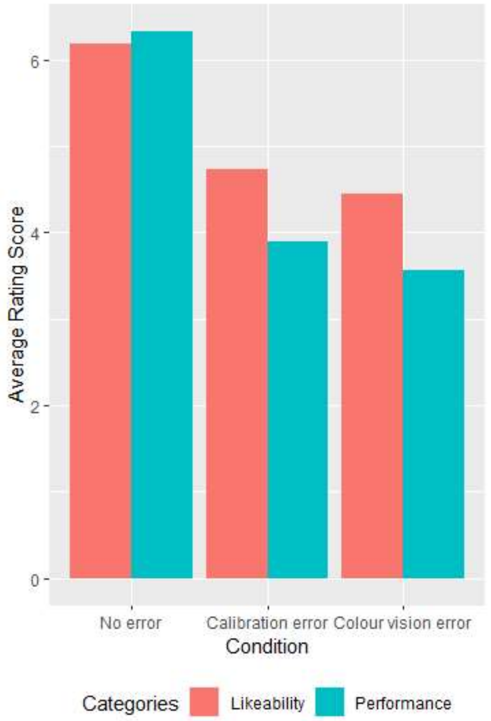


Figure 2: Rating of the robot in the no-error and the two error conditions. The ratings for the no-error condition were significantly higher than for the two error conditions.

3.4.2 Rating of Explanation Quality. To answer H2, we asked the participants whether they had learned something because of the explanation and whether the explanation was helpful or not to get a general impression of the explanation quality. Overall, we found evidence to support H2. 14 participants stated that they had learned something from the calibration error explanation. 17 participants stated they had learned something from the computer vision error explanation. Besides the quantitative feedback of the participants, we also analysed the qualitative free-form feedback. Here, participants mentioned for computer vision error, that it would be helpful to add information how to solve the error (e.g., information whether the error occurred because the environment was too dark or too bright). For the calibration error, participants mentioned that the explanation was too technical and they would have needed more information how to fix the error or how to calibrate the robot correctly to avoid similar errors in future.

To evaluate the explanation quality in more detail, we used two items ("help to trust or distrust the robot" and "help to understand how the robot works") proposed by Hoffman et al. [17]. Analyses on these scales will be reported in the next sections.

3.4.3 Comparison of Explanation Modalities. For answering H3, we used independent samples t-tests to analyse the impression of the two different explanation modalities (textual vs. auditory) regarding explanation quality, likeability, and performance of the participants towards the robot. We found no significant differences between the conditions (see Table 1), supporting our H3 that explanation modalities do not differ.

3.4.4 Comparison of Robot-Error Types. For H4, the conducted paired t-tests regarding explanation quality (i.e., trustworthiness & understandability), performance, and likeability. The analyses revealed (see Table 2) that the computer vision error explanation helped more to trust or distrust the robot compared to the calibration error explanation ($d = 0.62$ - medium effect). In addition, we found that participants rated the robots performance better in the calibration error condition ($d = 0.49$ - medium effect). This results does not support our H4 as we found differences between the error-types regarding performance and trustworthiness ratings.

Table 1: No significant differences in explanation quality (trustworthiness & understandability), performance and likeability between the two different explanation modalities (textual vs. auditory) for both types of robot error. Trustworthy refers to "help to trust or distrust the robot", Understandable refers to "helps to understand how the robot works".

Explanation Modalities	Measurement	*df*	*t*	*p*
calibration error	trustworthiness	18	-0.94	.36
text vs. audio	understandability	18	-1.40	.18
	performance	18	.30	.77
	likeability	18	-0.60	.55
computer vision error	trustworthiness	18	-0.87	.39
text vs. audio	understandability	18	-0.50	.62
	performance	18	.74	.47
	likeability	18	-0.70	.49

Table 2: Significant differences in performance and trustworthiness between the two different error types (calibration vs. computer vision error).

Measurement	*df*	*t*	*p*	*d*
trustworthiness	19	-2.77	.012*	0.62
understandability	19	-0.89	.38	-
performance	19	2.18	.042*	0.49
likeability	19	1.78	.09	-

* $p < .05$

3.5 Discussion

From the pilot study, it became apparent that people rated the robot significantly worse in terms of its performance and likeability when it made a mistake. The general study design in terms of trust repair (comparing trust of a correct working robot and a robot who makes an error) was therefore maintained for the final study.

Based on the pilot study, it appeared that the explanation for the calibration error was too technical for end-users without experience in robotics. These resulted in significant lower trust rating and was mentioned by participants in the free-form feedback. This reflects the argument of Gerlings et al. [10] saying that there is no generalised user to address with explanations. Instead, explanations have to fit to the abilities and preferences of different stakeholders. To fit end-users needs, we therefore decided to use only the computer vision error in the final study and to generate explanations for it. The free-form revealed that users are not satisfied with just getting an explanation of the error that happened, but also want a solution to prevent the error in the future. This finding extends the work of Das et al. [6], who stated that explanations should include environmental context and a history of successful actions of the robot in the past to support non-expert users in robot-recovery assistance. Inspired by the free-form feedback, we decided to refine the problem statement for the study and compare 3 different levels of error explanation: (1) no explanation, (2) explanation of error source and (3) explanation of error source and a possible solution.

Since we did not find any significant differences regarding the modality of explanation (i.e., textual and auditory), we decided not to compare these factors in the final study. Due to better comparability, we decided to use only textual explanations.

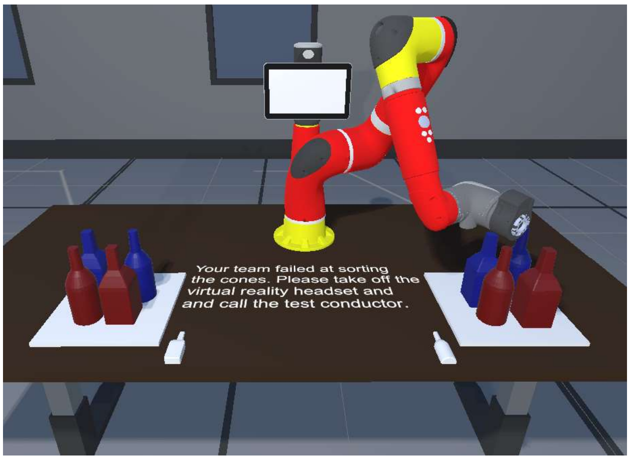
4 EXPERIMENT

To ensure high fidelity of system communication to the participants we opted to test HRC and mistake explanation using VR, rather than implementing and testing with a real robot and projection-based AR overlays. This also increased the test rate, as we could test with multiple participants at once, the only limit being the number of VR hardware setups. Based on the results from the pilot study, where the participants asked for more solution-oriented explanations rather than technical ones, we decided to define and test different explanation levels. The *first level* is an explanation to why the robot made the error, while the *second level*, in addition, explains how to solve the problem causing the error. We compare these two levels as trust-repairing actions after a robot mistake along with a control condition, where no explanation is provided, the user is only told that the robot failed the task. The trust-repair is evaluated in terms of both trust in the robot as well as perceived quality of the explanations. Our hypotheses are as follows:

- **H1:** Providing an explanation after a robot makes a mistake will yield higher levels of trust toward the robot than providing no explanation.
- **H2:** Providing different levels of explanation after a robot makes a mistake will yield different levels of trust toward the robot.
- **H3:** Adding solution-oriented details to robot mistake explanations will yield higher operator trust than explanations without them.

4.1 Virtual Environment

The experiment was performed using HTC Vive VR headsets and Vive Wand 6 degrees-of-freedom controllers. The virtual environment consisted of an office environment with desks and office chairs with participants being situated in an isolated corner of the room. Within reach of the participant was a desk with the robot mounted on top. The robot was modeled after the Rethink Robotics Sawyer robot. On the table was also a white square platform at either side of the robot with a little copy of the bottles involved in the test shown next to them, indicating which shapes of bottles have to be put where. The task involved sorting bottles by whether they had a round base or a square base. At startup, there were four bottles on each of the platforms, two red and two blue on each, and both have one bottle of each color that does not match the shape. This means that when the test started both the participant and the robot had to switch two bottles between the platforms to complete the shared objective. Between the two platforms was room to display text to convey instructions and explanations to the participants. The text was displayed on the surface, similar to a projected AR overlay. The participants were able to pick up the red bottles by moving a controller within 20 cm of their center and pressing the trigger. Letting go of the trigger released the bottle, and they dropped straight down as they cannot be thrown. In the case that a bottle was dropped on the floor, rather than requiring the participant to pick it back up, it will be moved back to its initial position. The test setup and robot in the virtual environment is shown in Figure 3.



Your team failed at sorting
the cones. Please take off the
virtual reality headset and
and call the test conductor.

Figure 3: The virtual test setup featuring the robot, bottles, their platforms and indicators, and the display text on the desk surface.

4.2 Procedure

After reading the experiment information and signing a consent form, the participant was given instructions to how to complete the test by the test conductor. The participant was informed that they would perform a collaborative task with a virtual robot and that they would be given instructions via the text displayed on the table. It was emphasized that they should read the instructions carefully, before they were told to put on the VR headset. The participant was introduced to the task by the text display. They were told that robot was their teammate and that they were only supposed to move the red bottles while the robot moved the blue ones as they sorted the bottles according to the small white bottles shown next to their white platforms. The participants proceeded through the text instruction using the Menu button at the top of the Vive wands. Before starting the task the participants were told how to move the bottles and they were told to try it.

When the participant was told to press the Menu button to start the task and they proceeded to press it, the robot would start moving the blue bottles. If the participant moved the bottles before they started the task, the bottles were moved to their starting position when the task began. The task was completed when the participant had sorted their bottles and the robot was done moving its bottles. In the first task the robot moved the bottles successfully, and the participant was presented with this message on the table: *"Your team succeeded at sorting the bottles. Please take off the virtual reality headset and and call the test conductor"*. When they took off the headset, they were presented with the 14-item version of the Schaefer HRT questionnaire [33]. Once the participant had completed the questionnaire, they were instructed to put the headset back on and follow the instructions.

Once they had put the headset on again, the display told them to start the task again by pressing the Menu button. In the second test the robot would make a mistake. Rather than switching a round-base and square-base bottles between their platforms, sorting them correctly, it would switch two round-base bottles, leaving two blue bottles in their wrong positions. The task ended once the participant had completed their part of the task correctly and the robot had stopped moving. The participant was then presented with this message displayed on the table: *"Your team failed at sorting the bottles"*. If the participant was testing the condition with no explanation of the mistake they were immediately presented with the text, *"Please take off the virtual reality headset and and call the test conductor"*. If the participant was testing the condition where they were given an explanation, they were presented with the message, *"A computer vision error occurred. The system did not successfully distinguish the bottles"*, before being told to take the headset off. Lastly, if the participant was in the condition with solution-oriented details, in addition to the previously mentioned explanation they were presented with the message, *"Better lighting conditions will help with successful sorting"*, before being told to take the headset off. The lighting conditions were the same in the virtual environment between the two tasks. Once they had taken the headset off the participants was presented with another HRT questionnaire as well as additional post-test questionnaire, which they were told to fill out outside the laboratory. The approach of only doing two tasks was chosen due to the time required to answer the post-test questionnaires as well as to not have the participants put the VR headsets on and off too many times.

4.3 Evaluation Methods

To evaluate the participants' impression during and after the VR task, we used the following scales.

Trust. During and after the VR task, we presented the 14-item version of the Schaefer HRT questionnaire [33] at the end of each task. In the post-test questionnaire, we used the item "This explanation lets me judge when I should trust and not trust the robot" from the EES [17] to calculate an additional trust score reflecting the explanation quality.

Explanation Satisfaction. We used the ESS [17] to measure the participants' subjective satisfaction with the kind of information (no explanation, explanation, or explanation with solution) that we presented after the robot mistake.

Emotions. We used items for the sub-scales *anger*, *happiness*, *anxiety*, and *relaxation* of the Discrete Emotions Questionnaire (DEQ)[13] to evaluate the participants feelings after the VR task.

Self-efficacy. We used two items to measure the self-efficacy towards the robot. For this, we used a variation of the item proposed by Bernacki et al. [4] (i.e., "How confident are you that you would successfully interact with a robot like this one in the study in the future" and "How confident are you that you could solve a robot error like this one in the study in the future?").

4.4 Participants

30 participants between 21 and 31 years (M = 24.0, SD = 2.30) took part in our experiment. Of these 11 were female and 19 male. 29 of

the participants had heard of the term AI, but only 4 had heard of the term XAI.

5 RESULTS

5.1 Trust Scores

The participants answered an HRT questionnaire after completing each sorting task with the robot, the first one being successful, while in the second task the robot would make a mistake. With all data groups being parametric, performing a pair-wise t-test showed significant difference in HRT scores between the first and second task, whether no explanation ($t(15) = 5.3$, $p < .001$), the base explanation ($t(18) = 7.0$, $p < .001$) or solution-oriented explanations ($t(17) = 4.7$, $p < .001$) were provided. However, when comparing the levels of explanation provided to the participants, performing a one-way ANOVA showed no significant effects of the explanations nor the type of explanation on the HRT scores after the mistake ($F = (2, 27) = .23$, $p = .79$), nor on the delta of HRT scores between tasks ($F(2, 27) = .17$, $p = .84$). The average trust scores with confidence intervals are shown in Figure 4.

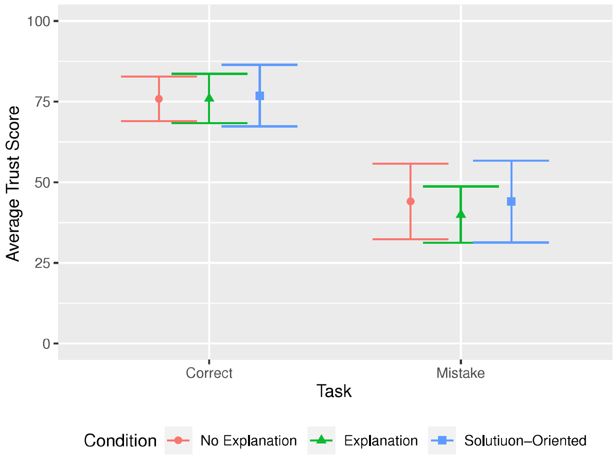


Figure 4: The average HRT scores and confidence intervals for the first and second HRC task between explanation conditions.

5.2 Post-Test Questionnaire

5.2.1 Explanation Satisfaction, Trust, and Self-efficacy. After the VR experiment, all participants answered the post questionnaire including questions about their explanation satisfaction and their trust in the explanation[5], their general impression of the robot and their self-efficacy towards the robot. To evaluate these variables between the three conditions, we conducted a one-way MANOVA. Here we found a significant statistical difference, Wilks' Lambda = 0.59, $F(10, 42) = 2.86$, $p = .008$. The following ANOVA revealed that only the variable trust showed a significant differences between the conditions, $F(2, 25) = 5.92$, $p = .008$.

To determine the direction of this difference between the three conditions, we used post-hoc comparisons[6]. We found the following differences:

- The participants' impression of helpfulness of the explanation to trust / distrust the system were significantly higher in the explanation & solution condition compared to the no explanation condition $t = -3.73$, $p = .002$, $d = 1.67$ (large effect).
- The participants' impression of helpfulness of the explanation to trust / distrust the system were significantly higher in the explanation condition compared to the no explanation condition $t = 2.49$, $p = .04$, $d = 1.13$ (large effect).

5.2.2 Emotional state. To evaluate possible differences in the emotional state of participants between the three conditions, we conducted a one-way MANOVA for the emotion categories happiness, anger, anxiety, and relaxation. Here we found no significant statistical difference, Wilks' Lambda = 0.84, $F(8, 46) = 0.50$, $p = .84$.

6 DISCUSSION

6.1 Main Findings

Based on the analyses of the trust scores we have to reject all three hypotheses. While all three conditions yielded significant decreases on reported HRT based on the scales, providing explanations to the error, with or without suggested solutions, showed no significant difference in trust, suggesting no trust-repairing effect.

6.1.1 Explanations alone are not sufficient to recover trust after robot-mistakes. While the ESS trust score showed that participants found the given explanations helpful to decide whether to trust or distrust the robot, this subjective impression of the participants was not reflected in their trust ratings during the VR task. Nevertheless, the ESS trust score can be seen as a first indicator that explanations might support trust-recovery in HRC, but that an explanation alone is not enough to recover trust after a robot-mistake, even when participants retrospectively rate the explanation as helpful. Despite the effect of the helpfulness of the explanations to trust or distrust the robot, this trust can not be assumed to be transferable to trust in the robot itself, especially as scales for trust in automation and HRT are not interchangeable [21].

The effectiveness of explanations seems to depend on various aspects. Researchers like Gerlings et al. [10] state that explanations have to fit different stakeholders and not to "the user" in general. We extend this view by saying that it is important to differentiate between the perception of an explanation given in an actual HRC situation and the rating of an explanation afterward. Our work contributes to the operationalization of the taxonomy of interpretability proposed by Doshi-Velez and Kim [8]. Here the authors state that the evaluation of explanations should not be done by using only proxy-tasks (i.e., studies without humans) but also include users by conducting human-grounded evaluations (for simple tasks) as we did in our study. The next step in their taxonomy is to use the

[5] We calculated an overall explanation satisfaction value and used in addition the item for the helpfulness of explanation to trust or distrust the robot as a single variable. For details, see section 4.3

[6] We used the Holm correction for multiple testing to adjust the p-values for all post-hoc tests we calculated.

insights from these simple-task experiments to conduct application-grounded evaluations using real-world tasks with domain experts. Our results, therefore, build a baseline for real-world applications. Another important variable is the scenario of the task. Compared to our VR task, Nikolaidis et al. [27] found out that in their study (i.e., a physical human-robot collaboration task), explanations greatly increased human trust to take robot's suggestions. Another important variable is the emotional presentation of the explanation. The affect in how an explanation is presented to the user plays a role in the effectiveness of the explanation [22, 29]. Affective feedback given by a robot leads to a more positive user impression [14, 24]. The work of Robinette et al. [31] propose that the apology of a robot after an error increases trust in the user.

To make the explanations for HRC more effective and improve robot trustworthiness, the recommendations of Kunkel et al. [23] and Weld and Bansal [40], among others, should be considered for further studies. Kunkel et al. [23] point out that richer explanations are preferred by users. In addition, Weld and Bansal [40] recommend interactive explanations. Here, the robot could be provide answers to follow-up questions and actions (e.g., giving more details, changing the vocabulary, attempting to correct the error), leading to a more social process of explanation.

6.1.2 Include variables such as emotions and self-efficacy to get a complete view of explanations' impact. The explanations in our study did not increase participants' self-efficacy, meaning that they did not feel more confident to interact with the robot in the future. In addition, the emotional state of the participants in the three conditions did not differ. As Mertes et al. [26] stated, it is important to measure the emotional state and the self-efficacy of users during human-computer interactions as they are relevant to get a complete view of the impact of XAI. They found that successful explanations (i.e. helping the user to perform better in a task and to understand the AI better) leading to more positive and less negative emotions and increase self-efficacy and trust. In our study, we showed that participants were not emotionally affected by the explanations neither did the explanations change the self-efficacy of users. This is in line with the fact that the explanations did not increase trust in the robot after a mistake and indicates that there could be a connection between the emotional state of users and the trust they have in robots.

For future studies, it would be valuable to explicitly ask participants about how their perception of the system communication affects their perception of the robot. In addition, investigating whether there is a separation between the robot and its operating system and communications in the participant's mental model could gain deeper insights into how users perceive the given explanations of a robot. Considering participants showed higher trust toward the explanations relative to the robot, they may consider the robot and the communication system as two separate entities.

6.2 Limitations

We conducted a VR-based instead of a real-world HRC task. This, in fact, likewise represents a limitation of the current work, but as Petrak et al. [28] stated, VR can be a helpful tool for prototyping scenarios where humans and robots interact. We are convinced that our setup used and the associated results may prove helpful in designing real-life interaction HRC studies and might be developed further and in more detail in future work.

The results of our study may have been affected by the participants' understanding of the collaborative task. Some participants seemed to have difficulty with the task, as they would often move a bottle matching the shape of the bottle moved by the robot, rather than following the instructions and sorting bottles according to the indicators on the table. The difficulty understanding the task may affect the participants' perception of the robot's mistake and the explanation by extension. If the participants do not understand the task, when told that the team failed the task, they may not think to inspect the robot's work and recognize its mistake, which can affect their perception of the explanations. Lastly, having the participants perform tasks simultaneously with the robot may affect how attentive they can be toward the robot and whether they can critically inspect the robot's work during the task. In future experiments the instructions should be clearer or the bottles should be distinguishable by more factors than their shapes while still indicating which should be moved by the robot or the participant. In addition, future studies could include physiological measures (1) as emotional indicators (see [3] for an overview) and (2) for a more reliable measurement of trust (e.g., eye tracking [25]).

7 CONCLUSION

We set out to investigate whether system explanations as a trust-repairing action after a robot makes a mistake in a collaborative task is helpful. In our conducted pilot study we found that end-users preferred less technical explanations with a greater emphasis on how to solve the error. Using a VR testing environment for our main study, we evaluated three levels of explanations after the robot made a mistake in executing a shared objective (i.e., sorting a set of bottles by shape) in collaboration with our participants. After comparing the conditions (no explanation, explanation of robot error, and explanation of error with solution-oriented details) with 30 participants we found no significant effects on their trust toward the robot. While participants found the explanations helpful to trust or distrust the system, we can not assume this trust to be transferable to the robot. Future studies should consider the participants' understanding of the shared task with the robot, ensuring that they recognize the nature of the robot's mistake and gain the most from the explanations. In addition, special consideration should be put into investigating the participants' mental model, emotional state, and self-efficacy when interacting with a robot supported by an explanation system to gain understanding regarding which construct the trust is placed in.

ACKNOWLEDGMENTS

This research is supported in part by Innovation Fund Denmark. This work was also partially funded by the European Union's Horizon 2020 research and innovation program Mind Bot (grant agreement No 847926).

REFERENCES

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140. https://doi.org/10.1371/journal.pone.0130140

[2] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, IEEE, Munich, Germany, 701–706. https://doi.org/10.1109/ROMAN.2008.4600749

[3] Stephanie Balters and Martin Steinert. 2017. Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices. *Journal of Intelligent Manufacturing* 28, 7 (2017), 1585–1607.

[4] Matthew L Bernacki, Timothy J Nokes-Malach, and Vincent Aleven. 2015. Examining self-efficacy during learning: variability and relations to behavior, performance, and learning. *Metacognition and Learning* 10, 1 (2015), 99–117.

[5] Michael W Boyce, Jessie YC Chen, Anthony R Selkowitz, and Shan G Lakhmani. 2015. Effects of agent transparency on operator trust. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. ACM, New York, NY, USA, 179–180. https://doi.org/10.1145/2701973.2702059

[6] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, USA, 351–360. https://doi.org/10.1145/3434073.3444657

[7] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.

[8] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608

[9] Mihai Duguleana, Florin Grigorie Barbuceanu, and Gheorghe Mogan. 2011. Evaluating human-robot interaction during a manipulation experiment conducted in immersive virtual reality. In *Virtual and Mixed Reality - New Trends*. Springer, Springer, Berlin, Heidelberg, 164–173. https://doi.org/10.1007/978-3-642-22021-0_19

[10] Julie Gerlings, Millie Søndergaard Jensen, and Arisa Shollo. 2021. Explainable AI, but explainable to whom? arXiv:2106.05568

[11] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI Magazine* 40, 2 (2019), 44–58. https://doi.org/10.1126/scirobotics.aay7120

[12] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 5 (oct 2011), 517–527. https://doi.org/10.1177/0018720811417254

[13] Cindy Harmon-Jones, Brock Bastian, and Eddie Harmon-Jones. 2016. The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PloS one* 11, 8 (2016), e0159915.

[14] Helen Hastie, Pasquale Dente, Dennis Küster, and Arvid Kappas. 2016. Sound emblems for affective multimodal output of a robotic tutor: A perception study. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 256–260. https://doi.org/10.1145/2993148.2993169

[15] A. Heimerl, K. Weitz, T. Baur, and E. Andre. 2020. Unraveling ML Models of Emotion with NOVA: Multi-Level Explainable AI for Non-Experts. *IEEE Transactions on Affective Computing* (2020), 1–1. https://doi.org/10.1109/TAFFC.2020.3043603

[16] Joshua D. Hoffman, Michael J. Patterson, John D. Lee, Zachariah B. Crittendon, Heather A. Stoner, Bobbie D. Seppelt, and Michael P. Linegang. 2006. Human-Automation Collaboration in Dynamic Mission Planning: A Challenge Requiring an Ecological Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 23 (2006), 2482–2486. https://doi.org/10.1177/154193120605002304

[17] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608 [cs.AI]

[18] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 164–168. https://doi.org/10.1145/2856767.2856811

[19] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. 2021. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence* 301 (2021), 103571. https://doi.org/10.1016/j.artint.2021.103571

[20] Poornima Kaniarasu and Aaron M Steinfeld. 2014. Effects of blame on trust in human robot interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, IEEE, Edinburgh, UK, 850–855.

[21] Theresa T Kessler, Cintya Larios, Tiffani Walker, Valarie Yerdon, and PA Hancock. 2017. A comparison of trust measures in human–robot interaction scenarios. In *Advances in human factors in robots and unmanned systems*. Springer, 353–364. https://doi.org/10.1007/978-3-319-41959-6_29

[22] Jonathan Klein, Youngme Moon, and Rosalind W Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with computers* 14, 2 (2002), 119–140.

[23] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[24] Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. 2012. Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, New York, NY, USA, 367–374. https://doi.org/10.1145/2157689.2157811

[25] Yidu Lu and Nadine Sarter. 2019. Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability. *IEEE Transactions on Human-Machine Systems* 49, 6 (2019), 560–568. https://doi.org/10.1109/THMS.2019.2930980

[26] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. 2020. This is not the texture you are looking for! Introducing novel counterfactual explanations for non-experts using generative adversarial learning. arXiv:2012.11905

[27] Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha Srinivasa. 2018. Planning with verbal communication for human-robot collaboration. *ACM Transactions on Human-Robot Interaction (THRI)* 7, 3 (2018), 1–21.

[28] Björn Petrak, Katharina Weitz, Ilhan Aslan, and Elisabeth André. 2019. Let me show you your new home: studying the effect of proxemic-awareness of robots on users' first impressions. In *2019 28th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, IEEE, ew Delhi, India, 1–7. https://doi.org/10.1109/RO-MAN46459.2019.8956463

[29] Rosalind W Picard and Jonathan Klein. 2002. Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with computers* 14, 2 (2002), 141–169.

[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[31] Paul Robinette, Ayanna M Howard, and Alan R Wagner. 2015. Timing is key for robot trust repair. In *International conference on social robotics*. Springer, Springer, Cham, 574–583. https://doi.org/10.1007/978-3-319-25554-5_57

[32] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/2696454.2696497

[33] Kristin Schaefer. 2013. *The perception and measurement of human-robot trust.(2013).* Ph.D. Dissertation. Doctoral dissertation, University of Central Florida Orlando, Florida.

[34] Raymond Sheh. 2017. "Why did you do that?" Explainable intelligent robots. In *AAAI Workshop-Technical Report*. AAAI Press, San Francisco, California,USA, 628–634.

[35] K. Stubbs, P. J. Hinds, and D. Wettergreen. 2007. Autonomy and Common Ground in Human-Robot Interaction: A Field Study. *IEEE Intelligent Systems* 22, 2 (2007), 42–50. https://doi.org/10.1109/MIS.2007.21

[36] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2006. The role of physical embodiment in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, IEEE, Hatfield, UK, 117–122. https://doi.org/10.1109/ROMAN.2006.314404

[37] Ning Wang, David V Pynadath, and Susan G Hill. 2016. The impact of pomdp-generated explanations on trust and performance in human-robot teams. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*. ifaamas, Richland, SC, 997–1005.

[38] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Christchurch, New Zealand, 109–116. https://doi.org/10.1109/HRI.2016.7451741

[39] Katharina Weitz, Teena Hassan, Ute Schmid, and Jens-Uwe Garbas. 2019. Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods. *tm-Technisches Messen* 86, 7-8 (2019), 404–412. https://doi.org/10.1515/teme-2019-0024

[40] Daniel S Weld and Gagan Bansal. 2018. Intelligible artificial intelligence. arXiv:1803.04263