Questioning the Robot: Using Human Non-verbal Cues to Estimate the Need for Explanations

Dimosthenis Kontogiorgos
Massachusetts Institute of Technology
Cambridge (MA), USA
dimos@csail.mit.edu

Julie Shah
Massachusetts Institute of Technology
Cambridge (MA), USA
julie_a_shah@csail.mit.edu

A robot arm explaining to a human how it detected an error using social signals, employing post-hoc explanations to communicate the predictions of an error detection deep-learning model. The study explored three explanation types: why-explanations (global feature importance), how-explanations (local feature contributions), and what-if-explanations (counterfactual scenarios). The figure shows a what-if explanation, illustrating how changes in the user's facial expressions could alter the prediction.

***Abstract*—As black-box AI systems become increasingly complex, understanding when and how to provide explanations to users is crucial. Multimodal signals, such as facial expressions, offer novel insights into how frequently explanations should be given. This paper explores whether users' facial features can help estimate the need for explanations in a collaborative robot task. We applied three state-of-the-art eXplainable AI (XAI) methods, addressing *how*, *why*, and *what-if* questions, explaining the robot's failure detection model. Each explanation type conveyed information differently: how-explanations described how the model functions, why-explanations provided personalised insights into input-feature-related cues, and what-if-explanations explored alternative scenarios. In a mixed-design study (N = 33), participants performed a robot-assisted pick-and-place task, receiving different explanation types. Our results show that users responded differently to these explanations, with why-explanations being the most preferred and prompting closer alignment in facial expressions with the robot. Contrary to expectations, what-if explanations led to the least alignment and required greater vocal effort. These findings demonstrate how non-verbal cues can guide the frequency and type of explanations (personalised or general) and further highlight the importance of model transparency in human-robot collaboration.**

***Index Terms*—explainability; failure detection; multimodality**

I. Introduction

In the 1950s, science fiction writers predicted that by 2000, robots would integrate into daily life. As we near the 2030s, general-purpose robots capable of performing complex tasks remain out of reach. Despite progress in AI, significant technical and ethical challenges persist, particularly with robots [139]. Issues such as limited dexterity, difficulty navigating unstructured environments and transparency remain problematic. Although AI systems can make decisions, their ability to explain them remains inadequate. Regulatory initiatives, like the EU's GDPR [143], aim to address concerns related to privacy and the "right to an explanation" [23, 47]. Explaining actions is crucial, especially when errors occur [4, 60, 82, 103, 115, 116, 118, 121]. This paper focuses on mechanistic interpretability of error detection and on how different explanation types influence user behaviour. While much of the work on robot errors centres on causality [85, 88], few works have examined error detection explanations, particularly when users provide social signals indicating an error.

Approach

In this study, we focus on user-centred transparency, proposing the use of human non-verbal signals to guide the explanation process. Robots should clearly show how these signals are used and automatically recognise when users need explanations [144]. We employed a deep-learning error detection model trained on Facial Action Units (FAUs) [35] from a public HRI dataset [133] to classify states of confusion [78, 79, 146]. This model was deployed on a robot arm for a pick-and-place task, where random perturbations induced errors. The robot detected these errors using users' social signals and provided explanations to make its model transparent.

We applied three state-of-the-art XAI techniques (within-subject factor) to deliver explanations of how errors were

This research was supported by the Knut and Alice Wallenberg Foundation.

detected. We also explored the frequency of explanations (between-subject factor), varying whether they were *explicitly requested* by the user or *proactively* generated by the system based on positive or negative classifications. This allowed us to assess how users responded to explanations, even when they did not request them. The three explanation types (how, why, and what-if) offered different ways to make the error detection model transparent. We analysed user behaviour through multimodal input, such as facial expressions and vocal effort, and we also asked participants to evaluate the explanations based on cognitive load, trust, and model understanding. Finally, we examined whether users adjusted their behaviour to align with the robot's model after receiving explanations.

Research questions

This study aimed to address the following research questions:

- *RQ1*: Which type of explanation (how, why, or what-if) is most suitable for explaining error detection in HRI, based on users' preferences and behaviours?
- *RQ2*: How do the prediction label and explanation frequency affect user behaviour and explanation perception?
- *RQ3*: After receiving explanations, do users adjust their behaviour and social signals to better align with the robot's error detection model?

Contributions of this article

To our knowledge, this is the first study to provide interactive explanations of an error detection model using non-verbal cues. Our study contributes to the HRI literature as follows:

- We developed a real-time robot explanation system, applying XAI methods interactively and using facial cues to explain an error detection model.
- We show that the type and frequency of explanation can alter users' behaviour and influence model understanding.
- We present findings suggesting that users may adjust their behaviour in response to robot explanations and that the type of explanation plays a role in shaping this process.

A robot's ability to detect errors through social signals and provide explanations has an impact on learning and care applications by addressing confusion, guiding users through training, and supporting therapy or care by identifying distress and providing personalised explanations.

II. XAI: eXplainable Artificial Intelligence

XAI has been applied across various fields [6, 52, 87, 92] to interpret the black-box nature of AI models, including areas like Affective Computing [48], HCI [2, 15, 63, 73, 74, 129, 149, 152, 159], NLP [7, 95, 117, 147, 156], recommender systems [32, 50, 53, 62], and social and cognitive sciences [18, 24, 104, 131, 138]. XAI has also been explored in philosophy [13, 67, 105, 130], folk-psychological approaches [28, 29, 140, 141], journalism [126], and across different modalities [137].

Explaining AI decisions has far-reaching implications, as computing is now central to human communication and information management [76]. Transparency has been studied in autonomous vehicles [150], especially in relation to intent communication for pedestrian trust [98], and in faulty robot scenarios [17, 31, 69, 77, 114, 127, 145]. The relationship between transparency, trust [37], and anthropomorphism [150, 153] has also been examined. Emerging research highlights a growing interest in user-adaptive explanations that leverage human input to tailor their content [34, 45, 86].

A. Explaining causality - explanations as justifications

Much of the work in HRI explainability has focused on causality. Research in this area aims to make robot behaviour transparent by providing justifications or clarifying intentions, such as through legibility [3, 33, 85], repeating failed actions [109], or selecting actions that lead to failures [110]. Other work focuses on robot planning [21, 44, 54, 55, 91, 101, 122]. Das et al. [26] used causal knowledge to generate context-aware explanations, while Han et al. [56, 57] explored hierarchical and failure-focused robot explanations. Other research has examined how to explain misalignment between human and robot expectations [1, 14, 135, 136]. Providing justifications has been shown to reduce negative perceptions of robots following errors [25] and mitigate decision-support challenges [96]. However, none of these works focus on explaining error detection models, especially those based on user signals.

B. Explaining the model - explanations for interpretability

In this paper, we adopt the mechanistic interpretability approach, using the SAFE-AI framework [119] to guide our study design and define explanation types for robot behaviour. This framework provides a high-level view of explanation methods in robotics, applicable to both causal explanations and model interpretability. It builds on situational awareness and transparency concepts [22, 36] aligning with Lim et al. [90] and the XAI Question Bank by Liao et al. [89]. Our within-subject factor uses the three SAFE-AI layers: (1) *XAI for Perception*, answering "how" questions; (2) *XAI for Comprehension*, answering "why" questions; and (3) *XAI for Projection*, addressing "what-if" questions [103, 115, 118].

How-explanations: XAI for Perception

This layer explains how decisions are made, covering both input and output of ML models. Common approaches include global feature importance [38, 61, 94, 99, 108, 125, 142], class-level information [111, 162], and summarising the model's learned patterns [11, 12, 19, 40, 70]. Techniques include prototypes [68, 71], linear model approximations [10, 66, 84], and rule extractions [27, 154, 163], with SHAP [97] being widely used. *For robot error detection, the robot would display its training data and the features important for its predictions.*

Why-explanations: XAI for Comprehension

These explanations link specific inputs to outputs, highlighting which features in the input are informative for the model's predictions [59, 72, 112]. Local feature importance and saliency methods [39, 42, 55, 65, 93, 97, 120, 124, 128, 134, 155, 157, 161] or rule extraction [51, 113, 158] are often used. LIME [112] is a popular technique. *In the case of a failing robot, this type of explanation would show users which features of their input led to the current prediction.*

What-if-explanations: XAI for Projection
This layer explores alternative outcomes by showing what changes in the input would lead to different results, using counterfactuals to guide future behaviour changes [9, 100]. Techniques include feature change influence [5, 43, 46, 83] and counterfactual features [30, 148, 160]. *For a failing robot, this explanation would show users what changes in their input could alter the prediction.*

C. The need for error detection explainability

While much research focuses on explaining causality and justifying robot behaviour, a gap remains in explaining how errors are detected through users' non-verbal signals. This presents an opportunity to improve transparency by showing users how their data is used in error detection. It is critical not only to use error detection as a recovery tool but also to reveal when the robot is wrong, as all ML models will eventually make incorrect predictions. Explaining why a robot error wasn't detected can be as valuable as explaining how it was, provided the user understands how the model operates. This paper does not aim to explain robot behaviour or address task-related errors. Instead, our approach, grounded in mechanistic interpretability, focuses on aligning user expectations with the system's interpretation of social signals.

III. Method

This study aims to investigate how different explanation types influence error detection empirically and whether proactively displaying explanations or allowing users to prompt them affects their utility. We used post-hoc, interactive, and model-agnostic explanations [144], presented uniformly across methods. We employed a 3x3 mixed-factor design, with one within-subject and one between-subject factor.

A. Explanation design

Within factor: explanation type
We examined three explanation types applying widely used XAI methods for how, why, and what-if explanations.

How-explanations. Global feature importance explains model predictions by showing how features impact the model's overall output. We use a SHAP algorithm to compute Shapley values [123] based on feature permutation, applying Shapley Value Sampling [20] from Kokhlikyan et al. [75].

Why-explanations. Local feature contribution shows how specific user inputs contributed to the model's prediction. We use Kernel-SHAP, which locally approximates the decision boundary and leverages the LIME framework [112] to compute local Shapley values [75, 97].

What-if-explanations. Counterfactuals display how minimal changes in two features could alter the prediction, keeping other features constant. This provides greater human intuition about the model's decision boundaries. We applied Mothilal et al.'s method [106] to generate feature-perturbed versions of the input to modify the model output.

Between factor: explanation elicitation
We aimed to assess how explanations are elicited based on information needs [119], either proactively displayed by the robot or prompted by the user while varying the level of detail. Participants were divided into three groups:

User-initiated. Explanations were provided only when explicitly requested by users. Participants informed the robot of an error to trigger the robot's prediction and an explanation.

Proactive - Low Frequency (Errors). The robot proactively displayed explanations only when errors were detected.

Proactive - High Frequency (Errors+NoErrors). Explanations were shown for both error and no-error predictions, with a maximum of four no-error explanations per interaction.

Participants could request the robot's prediction if they observed an error. This design ensured exposure to the model across various prediction classes, allowing for disagreement (e.g., a user reporting an error that the robot did not detect). As no-explanation baselines are well-established in HRI, we chose a user-initiated condition as the baseline to emphasise the importance of explanations without proactive provision. In the proactive conditions, the robot provided explanations based on positive and negative classifications (as interpreted from its perspective). Notably, there was no external ground truth beyond the robot's own predictions. Though widely applied in XAI research, there is little consensus on the relevance of these techniques in HRI, particularly for error detection.

Hypotheses
RQ1 - Hypothesis H1: We expected what-if explanations, which simplify information by only highlighting minimal changes, would be preferred due to cognitive load reduction (following Miller's 7±2 rule for information chunks [102]).
RQ1 - Hypothesis H2: We anticipated that how-explanations would be less engaging, as they are independent of user input.
RQ2 - Hypothesis H3: We hypothesised that participants in the Errors+NoErrors group would better understand the model as they are exposed to more diverse predictions.
RQ2 - Hypothesis H4: We hypothesised that the number of explanations would vary across conditions, with User-initiated having the fewest and Errors+NoErrors the most.

B. Introducing errors

Prior work shows that human reactions to robot errors are predictable [41, 81], and that failures in HRI are typically deviations from expected behaviour [82]. To stimulate the error detection model, we introduced random perturbations in the robot's input and output. We were interested in how users reacted during moments of uncertainty and how the robot's model interpreted these signals to provide explanations. Each interaction included four errors, each occurring once in a random order, and used confusion as a trigger for explanations. The failures were informed by Honig and Oron-Gilad's taxonomy [64] and Stiber et al.'s error dataset [133]. This study focuses on hardware and software errors, such as robot gripper and orientation issues, as well as ASR and computer vision errors (Table I).

C. Error detection deep-learning model

We used the robot error dataset from Stiber et al. [133], which contains 28 interactions and records 17 Facial Action

Hardware	**Gripper position error**: (send perturbed position value to ROS node) the robot drops an object before reaching the target **Griper orientation error**: (send perturbed orientation value to ROS node) the robot does not pick up an object because of gripper misconfiguration
Software	**ASR error**: (add noise to user's utterance) the robot incorrectly picks up an object due to ASR mis-understanding **Computer vision error**: (simulate CV failure) the robot is not able to find the object because of CV error

TABLE I: Technical errors used in every interaction.

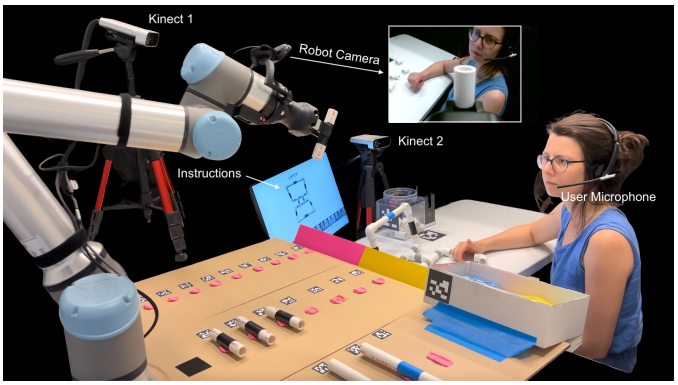

Fig. 1: Experimental setup with task objects and the UR5 robot positioned next to the user for the pick-and-place task.

Units at 3Hz. The error detection model (average detection delay: 3.2s) is trained to detect errors based on user behaviour and reduces false positives by employing weighted classification with Softmax probabilities (sliding window of 3.6s) [132]. The algorithm traces back to the estimated start of the error, which serves as the input for the explanation algorithms. In User-initiated conditions, predictions were displayed only when requested, while Errors and Errors+NoErrors conditions proactively displayed them when errors were detected or when programmed to display a negative prediction.

D. Experimental setup

The experiment involved a pick-and-place task where participants interacted with a robot arm. An external monitor displayed task instructions and explanations. Multiple sensors were used to capture participant behaviour (Figure 1).

Facial expression and body pose tracking. Two Azure Kinect captured participants' facial expressions via RGB cameras and body poses via depth cameras. This ensured continuous facial data capture, even if one camera was occluded [133]. An additional HD webcam (Logitech C615) mounted on the robot's gripper provided a robot's-eye view of the interaction.

Speech input and output. Participants wore close-talking microphone headsets to issue voice commands to the robot and hear responses through headphones. Microsoft's Azure ASR system transcribed their speech, prompting them to repeat if transcription confidence was low. If no errors were reported by the participant, the transcript was sent to a Large Language Model (OpenAI GPT-4) for intent parsing. The LLM disambiguated user commands based on a custom prompt designed to handle varied object descriptions and ASR misrecognitions (e.g., "the shot pipe" instead of "the short pipe", "the same one as before", "the one on the right").

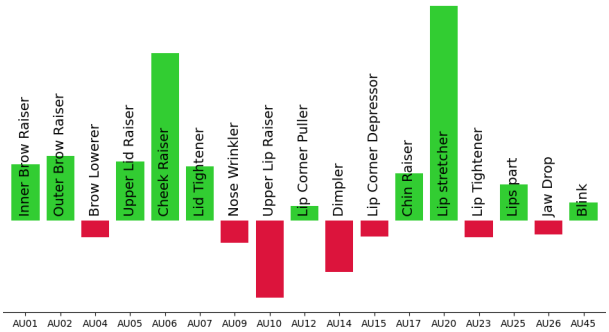

Fig. 2: Sample of global explanation bar chart in the study.

Robot motion and planning. Once the LLM identified the correct object, the robot's custom motion planner directed the robot's end effector to fetch the object. The robot, a Universal Robots (DK) UR5e with a Robotiq Hand-e gripper, was controlled by ROS1 Noetic nodes running on a Linux workstation. The robot always returned to a neutral position.

Error detection and multimodal signal processing. The robot's error detection model was integrated with custom LED lights on the gripper, which turned red when an error was detected. Simultaneously, the monitor displayed the robot's prediction and an explanation for 10 seconds. The ML and explanation system, synchronised through the Platform for Situated Intelligence (PSI) [16], ran on an NVIDIA RTX 3060 GPU. PSI synchronised multimodal data from facial expression tracking, voice input from the ASR, and visual input from the webcam, enabling real-time robot responses.

Explanation presentation and feedback. Explanations were visually displayed using bar charts (Figure 2), common in XAI for representing tabular data. Green bars indicated features positively contributing to the model's prediction, and red bars indicated negative contributions. The robot verbally introduced the explanation, explaining that it had analysed the participant's facial expressions to determine if an error had occurred, specifying the type of explanation (local, global, or counterfactual), however without naming the specific features.

Task. Participants sat at a 90-degree angle, with the robot on their left and their workspace in front of them, where pipe connectors were placed. The robot's workspace held PVC pipes for a collaborative pick-and-place task, in which participants used voice commands to request pipes to build three different structures. This assembly task, designed to require no prior skills, consisted of 12-14 pieces of varying colour and size. The robot communicated through Azure's text-to-speech (TTS) system, providing feedback such as "OK" or "Sure" to confirm user commands.

E. Behavioural measures

We extracted 23 features (Table II), focusing on participants' utterances, error detections, and explanations data points. *Facial expressions* were captured using OpenFace [8], extracting 17 facial Action Units as defined by the Facial Action Coding System (FACS) [35]. Using voice activity detection (VAD),

Facial Action Units	**AU1** (Inner Brow Raiser) **AU2** (Outer Brow Raiser) **AU4** (Brow Lowerer) **AU5** (Upper Lid Raiser) **AU6** (Cheek Raiser) **AU7** (Lid Tightener) **AU9** (Nose Wrinkler) **AU10** (Upper Lip Raiser) **AU12** (Lip Corner Puller) **AU14** (Dimpler) **AU15** (Lip Corner Depressor) **AU17** (Chin Raiser) **AU20** (Lip stretcher) **AU23** (Lip Tightener) **AU25** (Lips part) **AU26** (Jaw Drop) **AU45** (Blink)
Speech Features	**FrequencyDomainEnergy** Total energy in the signal's frequency spectrum **LogEnergy** Logarithm of the signal's total energy **LowFrequencyEnergy** Energy in the lower-frequency components **ZeroCrossingRate** Rate of zero-amplitude crossings
Gaze Features	**Gaze to explanation** Proportional gaze towards the explanation **Gaze to robot** Proportional gaze towards the robot

TABLE II: Behavioural measures captured in the interactions.

we identified all voiced segments during the interaction and extracted the *acoustic components* of participants' speech. *Head rotation* was also tracked to measure visual attention.

F. Subjective measures

Objective understanding. We adapted the model understanding measure from Wang and Yin [151] to assess global model understanding, local feature contributions, and counterfactual thinking. After each interaction, participants were given a list of all AU features used as inputs and asked to select the most influential features or identify the changes required to flip the prediction. For local and counterfactual explanations, participants were shown a prototype of facial features from the robot error dataset [132] along with a prediction.

Subjective understanding. Participants rated their perceived *model understanding* as in [151], the overall *model accuracy*, *cognitive load* (NASA-TLX [58]), *trust* (Muir's Trust scale [107]), perceived explanation *information amount* and *frequency* using a 5-point Likert scale. They also ranked their *most and least preferred explanations*.

G. Qualitative data

After each interaction, participants provided open-text feedback on the benefits and drawbacks of each explanation type. At the end of the study, they reflected on its overall purpose.

H. Procedure

After providing consent, participants were shown mock-ups of the explanations and informed that the robot used a deep-learning model to detect errors. Participants were informed that the robot was autonomous and that some errors might occur. They were also told they could ask for explanations if they noticed errors and that the robot would explain how it detected the error (without pointing to the cause). The robot was unaware that an error might occur and relied solely on users' social signals. Gaze was calibrated before interactions, and participants were instructed to pay attention to explanations and assess the model. Each participant engaged in three interactions with the robot, corresponding to the three explanation types. After each interaction, they completed questionnaires on their understanding of the model and other subjective metrics. Finally, participants completed a demographics questionnaire and ranked their explanation preferences, followed by debriefing.

I. Participants

We recruited 33 participants (11 per between-factor condition, reporting 14 female and 19 male) from the MIT campus, with an average age of 30.3 years ($\pm$9.7). A priori power analysis suggested that 36 participants would be needed for detecting medium-sized effects. Our sample size falls slightly below this threshold, with a weighted post hoc power analysis indicating an overall study power of approximately 73%, offering moderate sensitivity for exploratory analyses. We therefore advise caution when interpreting the results. Participants rated their English fluency at 4.6 ($\pm$0.5), technology experience at 4.6 ($\pm$0.7), and robot experience at 2.9 ($\pm$1.2) on a 1-5 scale. Ethics approval was obtained from the Institutional Review Board prior to the study (Protocol #2405001314). Each experiment lasted between 1 to 1.5 hours, and participants were compensated with a $30 Amazon voucher. The conditions were balanced using a Balanced Latin Square design.

IV. Results

We focused our analysis on the three RQs, using behavioural, subjective, and qualitative data. To account for the mixed-design structure, we employed Linear Mixed-Effects Models (LMMs), allowing us to test the main effects of within-subject factors (explanation type), between-subject factors (explanation elicitation), and their interactions. Random intercepts were included for participants to account for individual variability, for the order of interaction and the number of explanations exposed to, with the following notation: *DV* $\sim$ *ExplanationType * ExplanationElicitation + (1|Participant) + (1|InteractionOrder) + (1|ExplanationNum)*. Pairwise comparisons were performed using post-hoc Tukey's tests.

A. Behavioural measures

Facial action units. Table III shows the results from the LMMs on FAUs extracted during user utterances, comparing the fixed factors. Several facial expressions were linked to how participants responded to different explanation types (Figure 3), with why-explanations prompting greater expressiveness, as seen in higher activations of AU1, AU9, AU25, AU26, and AU45 (RQ1). Additionally, AU5 (upper lid raiser) was more active in the Errors+NoErrors condition (RQ2).

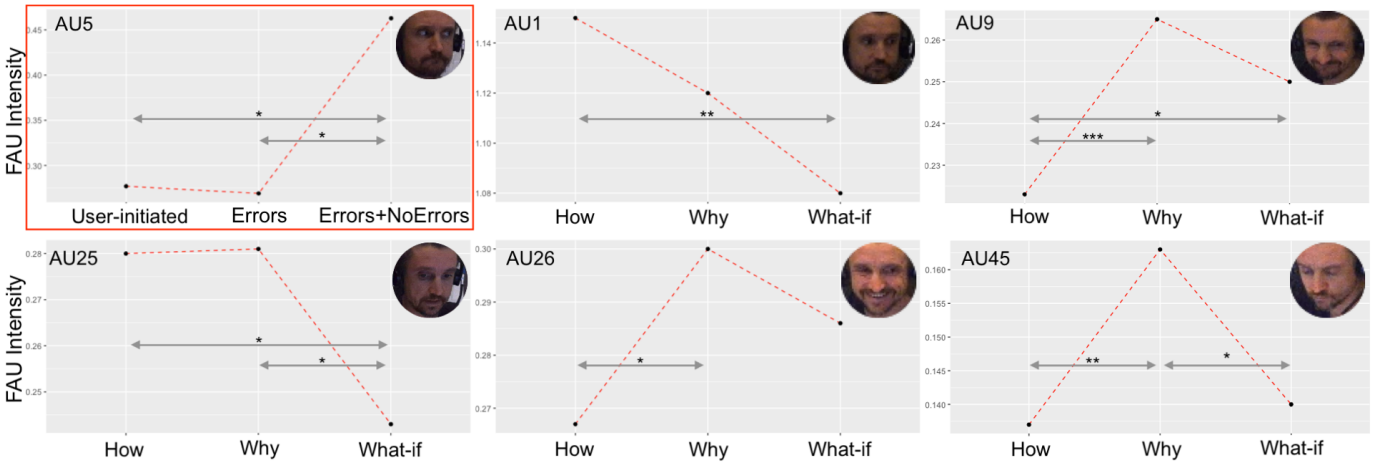

Fig. 3: Effects of explanation type and elicitation on FAUs, indicating greater expressiveness on why-explanations. All figures depict within-factors, except when marked in red (between-factor). P-value indicators: $*p \leq 0.05, **p \leq 0.01, ***p \leq 0.001$.

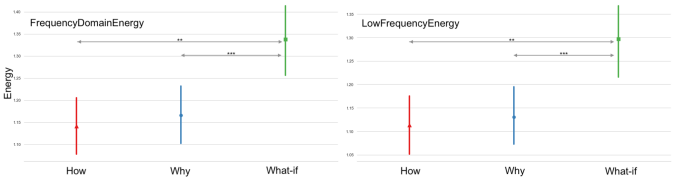

Fig. 4: Participants exerted greater vocal effort during interactions with what-if explanations.

Multimodal Features	Explanation Type	Explanation Elicitation	Explanation Type*Elicitation
AU1 (Inner Brow Raiser)	F=5.203, p=0.005	-	F=8.1668, p<0.001
AU2 (Outer Brow Raiser)	-	-	F=5.543, p<0.001
AU5 (Upper Lid Raiser)	-	F=4.196, p=0.024	F=4.010, p=0.002
AU6 (Cheek Raiser)	-	-	F=4.279, p=0.001
AU7 (Lid Tightener)	-	-	F=9.790, p<0.001
AU9 (Nose Wrinkler)	F=7.840, p<0.001	-	F=4.316, p=0.001
AU10 (Upper Lip Raiser)	-	-	F=2.706, p=0.028
AU14 (Dimpler)	-	-	F=2.633, p=0.032
AU15 (Lip Corner Depressor)	-	-	F=7.791, p<0.001
AU17 (Chin Raiser)	-	-	F=13.493, p<0.001
AU20 (Lip stretcher)	-	-	F=10.249, p<0.001
AU23 (Lip Tightener)	-	-	F=9.423, p<0.001
AU25 (Lips part)	F=5.352, p=0.004	-	F=6.932 p<0.001
AU26 (Jaw Drop)	F=3.253, p=0.038	-	-
AU45 (Blink)	F=6.267, p=0.001	-	F=2.934, p=0.019
FrequencyDomainEnergy	F=7.669, p<0.001	-	F=4.796, p<0.001
LowFrequencyEnergy	F=7.463, p<0.001	-	F=4.793, p<0.001
ZeroCrossingRate	-	-	F=4.242, p=0.001
Gaze to explanation	-	-	F=4.763, p<0.001
Gaze to robot	-	-	F=2.414, p=0.047

TABLE III: Multimodal features across within and between factors, overall indicating greater expressiveness with why-explanations and greater vocal effort with what-if explanations.

Acoustic features. Table III presents the statistical tests on the acoustic components from users' voiced segments, showing greater vocal effort during interactions with what-if explanations (RQ1), indicated by higher values in FrequencyDomainEnergy and LowFrequencyEnergy (Figure 4). A similar effect was observed for ZeroCrossingRate in the Errors (p<0.05) and Errors+NoErrors (p<0.05) conditions (RQ2).

Gaze features. We measured participants' head direction during calibration, using cosine similarity on angular differences to calculate the gaze direction. Focusing on participants' visual attention during explanations, revealed that

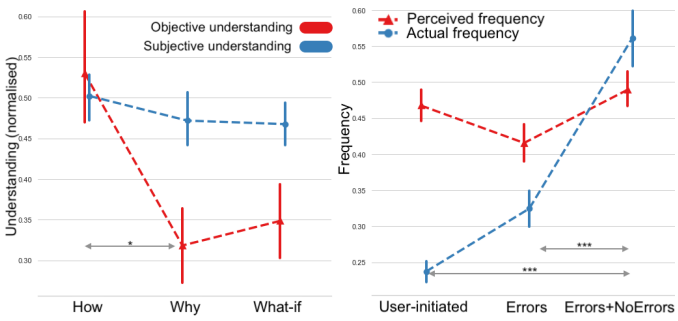

Fig. 5: a) Normalised objective and subjective understanding for each explanation type. b) Perceived frequency of explanations by between-factor condition, alongside the actual number of explanations participants were exposed to.

Explanation type	Most preferred	Least preferred
How	18%	42%
Why	**58%**	6%
What-if	24%	**52%**

TABLE IV: Participants' preferences for explanation types.

participants spent less time engaging with the explanations in the Errors+NoErrors condition. An interaction effect between explanation type and elicitation (Table III) showed that participants in the Errors group engaged more with how-explanations (p<0.01), while those in the Errors+NoErrors group engaged the least with how-explanations (p<0.05). The opposite effect was observed for gaze towards the robot.

B. Subjective measures

We extracted participants' responses for both **objective** and **subjective** model understanding. For each explanation type, we assessed whether participants correctly identified the relevant features (How: AU06 & AU20, Why: AU15 & AU26, What-if: AU20 & AU45). Following Wang and Yin [151], we normalised the objective and subjective understanding scores

	Subjective Measures								
Measure	**W1: How**	**W2: Why**	**W3: What-if**	**B1: User-initiated**	**B2: Errors**	**B3: Errors+NoErrors**	**Explanation Type**	**Explanation Elicitation**	**Explanation Type*Elicitation**
Objective Understanding	m=0.53, se=0.41	m=0.31, se=0.27	m=0.34, se=0.26	m=0.43, se=0.32	m=0.36, se=0.38	m=0.39, se=0.29	F=4.216, p=0.019	-	-
Trust	m=2.98, se=0.91	m=3.10, se=0.94	m=3.06, se=0.88	m=3.15, se=0.93	m=2.90, se=0.90	m=3.10, se=0.88	-	-	F=2.918, p=0.028

TABLE V: Mean values for subjective measures across within and between factors, along with statistical test results.

by dividing them by the maximum possible values (Figure 5). A significant main effect of explanation type was found for objective understanding (Table V), with better understanding observed for how-explanations, which referenced global feature importance independent of user input (RQ1). No significant effects were found for explanation elicitation (RQ2) or its interaction with explanation type. Similarly, no significant differences were observed in subjective understanding.

No significant effects were found in the perceived **model accuracy** dimension, which was expected since participants were exposed to the same detection model across all conditions (with explanations being post-hoc and not influencing predictions). Likewise, no statistical differences were observed in **cognitive load** (NASA-TLX) across explanation types or elicitation forms. However, an interaction effect on **trust** was identified (Table V), with users in the Errors+NoErrors group showing lower trust in how-explanations compared to why- and what-if explanations, possibly indicating overwhelm due to the repetitive nature of how-explanations.

No significant effects were found on the **amount of information**, across explanation types or elicitation. However, we observed differences in the **number of explanations** between the groups (RQ2). LMMs indicated that explanation elicitation significantly influenced how many explanations participants received (Figure 5), confirming the success of the between-group manipulation: F=14.711, p<0.001. No significant effects were found within the explanation type condition, as anticipated. Interestingly, participants' perceived **frequency of explanations** did not show a significant difference, but a high correlation was found between the actual number of explanations and perceived frequency: r=0.325, p=0.001.

Overall, we found that participants had a strong **preference** for why-explanations (Table IV): $\chi^2 = 8.909, p = 0.011$, whereas what-if-explanations were the least preferred: $\chi^2 = 11.454, p = 0.003$ (RQ1).

C. Effects of facial AU alignment with explanations

We calculated the alignment between users' facial expressions and the explanations they were exposed to. Using Dynamic Time Warping (DTW), we compared the continuous stream of explanations shown to users during the interaction with the time series of their facial expressions on error detections. This allowed us to measure the distance similarity between the two temporal sequences and calculate how aligned they were. For each feature, we computed the mean DTW distance (ignoring zero cases), then aggregated these to create a single alignment measure per feature, accounting for explanation type. This allowed us to observe whether facial expressions aligned with the "peaks" and "valleys" of the explanations. Features such as AU7, AU12, and AU45 exhibited high alignment overall, while AU15, AU17, and AU20 showed less alignment. We found a statistically significant main effect of explanation type (Figure 6): F=17.071, p<0.001, with why-explanations producing the closest alignment and what-if-explanations the least (RQ3). No significant effects were found for explanation elicitation groups or interactions.

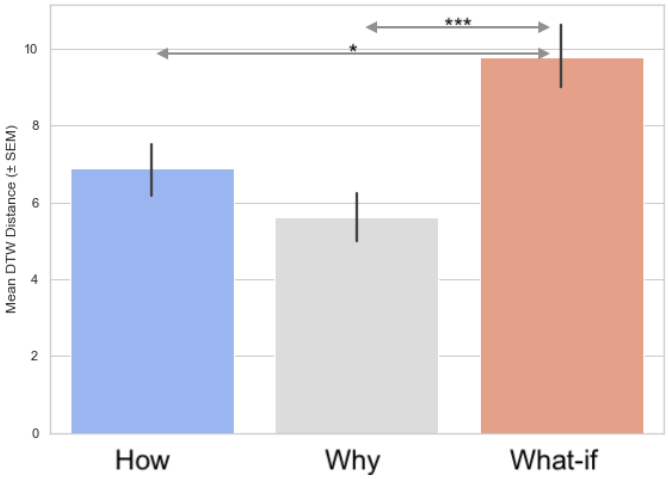

Fig. 6: DTW across explanations (lower values indicate higher alignment), indicating lower alignment in what-if explanations.

Explanation	**Benefits**	**Drawbacks**
How	Generalised (45%) Intuitive / understandable (36%) Consistent / concise (6%)	Not personalised (42%) Unclear / unspecific (18%) Frustrating (9%) Non-adaptive (6%) Additional information (6%)
Why	Personalised / user-specific (45%) Understandable (18%) Easy / concise (12%) Helpful (6%)	Unclear - difficult (18%) Non-adaptive (15%) Limiting (9%) Making me self-conscious of my reactions (9%)
What-if	Understanding input changes (39%) Understandable (18%) Easier (12%) Clear / concise (9%)	Confusing / counter-intuitive (21%) Unclear (18%) Making me self-conscious of my reactions (18%) Non-adaptive (9%) Repeating (6%)

TABLE VI: Benefits and drawbacks expressed by participants for each explanation type. Percentages indicate the proportion of participants who commented on each theme.

D. Qualitative data

Table VI summarises participants' perceptions of the explanations, highlighting recurring themes and the percentage of participants who mentioned them among the listed benefits and drawbacks. Open-ended questions were presented between tasks to gather participants' views on the explanations. Thematic analysis revealed that how-explanations were generally seen as clear and easy to understand, but lacked personalisation. In contrast, why-explanations were perceived

Explanation	Behavioural measures	Subjective measures
How	high FAU expressiveness low vocal effort high FAU alignment	high model understanding low preference
Why	high FAU expressiveness low vocal effort high FAU alignment	low model understanding high preference
What-if	low FAU expressiveness high vocal effort low FAU alignment	low model understanding low preference

TABLE VII: Summary of results from behavioural and subjective measures for each explanation type.

as personalised and user-specific, though some participants found them difficult to interpret. Counterfactual thinking in the what-if explanations was understood by many participants, but others found it confusing or unclear, as also reflected in the overall explanation preferences (RQ1).

V. Discussion & Conclusion

Key findings

1. How-explanations were considered intuitive but impersonal. Users in the *Errors+NoErrors* group trusted how-explanations less, increased exposure did not lead to better understanding. *2. Why-explanations* were preferred, perceived as personalised and led to the highest facial alignment, although how-explanations were understood the best.

3. What-if-explanations, despite being expected to be simpler, resulted in the least alignment. They required more vocal effort and were less engaging overall, with lower expressiveness despite participants reporting no additional cognitive load.

RQ1 - Effects of explanation type

Building upon the work of Stiber et al. [133] and Kontogiorgos et al. [81], our study offered a user-centred perspective on the types of explanations expected from an error detection model. Why-explanations consistently outperformed how- and what-if-explanations (Table VII) aligning with prior research [73, 145]. Unexpectedly, what-if-explanations were the least effective, requiring greater vocal effort and leading to lower expressiveness, potentially indicating disengagement (AU25) or attentiveness loss (AU1). In contrast, why-explanations prompted higher expressiveness, possibly as users tried to align their facial expressions with the model, and were preferred by users, *rejecting H1*. This may be because counterfactual explanations provide a high level of granularity, whereas users might prefer lower granularity and a broader overview of features rather than engaging in counterfactual reasoning. Despite the overall performance of why-explanations, how-explanations were less engaging, even though they were the most easily understood, *partially confirming H2*.

RQ2 - Effects of explanation elicitation

We found few effects related to how explanations were elicited (user-initiated or proactive), aside from trust measures showing that the Errors+NoErrors group trusted how-explanations less. That group also showed a significant increase in AU5 (Upper Lid Raiser), potentially indicating frustration, rather than understanding the model better, therefore *rejecting H3*. Some Errors participants noted that seeing true negatives could be helpful (P21: "*It would have been nice to see a comparison of facial features when the robot also guessed no error*"). However, others felt self-conscious (P15: "*Showing me what flips the model's predictions can bias me to suppress certain facial expressions*"). This suggests that explanations might sometimes hinder interactions. Finally, explanation frequency was not perceived as expected, despite the manipulation check confirming its success (*confirming H4*).

RQ3 - Effects of behaviour alignment

We observed a stronger behavioural alignment with why-explanations, as indicated by DTW distance, compared to other conditions. Participants also consciously adjusted their behaviour (P27: "*Understanding how the robot considers facial features, it's helpful to change your own expressions*"; P5: "*I get feedback on my expressions, so I can adapt to the system*"). Additionally, some participants appreciated the transparency in how their data was used (P27: "*It keeps me aware of how my data is being gathered and used*").

Limitations and impact

The visualisation format has its limitations, as not all users may easily comprehend bar plots and may benefit from alternative modalities, such as natural language. Furthermore, these results should be interpreted with caution due to power limitations. While they may generalise to error detection models, they may not be directly applicable to causality explanations. Additionally, readers should consider implementation-specific results when attempting to replicate this study, as different implementations could lead to alternative outcomes. We encourage replications across diverse interaction settings and detection models to further validate these findings. The purpose of error detection models is to inform the robot of an error that the user already knows occurred. If the model does not perform well, it may disrupt the user's task, leading them to seek an explanation. While users may engage with explanations, they primarily want to continue their task. Explanations must be concise and useful, avoiding information overload. Like human conversations, they should be integrated into the interaction, following audience design principles [49, 80].

Conclusion

In summary, our study provides empirical evidence on the types of explanations that are most effective for error detection models based on facial expressions. Explanations not only help users understand the robot's decisions but may also affect behavioural alignment with the user. We hope our findings will guide the design of future studies on robot error detection models, encouraging the integration of multimodal signals and the combination of error detection with error causality, placing multimodal HRI at the forefront of XAI research.

Acknowledgements

We thank Mike Hagenow, Mycal Tucker, and Lindsay Sanneman for their contributions to the discussions on this study. We also thank Maia Stiber, Matt Boyd, and Ane San Martin for their assistance in the development of the system, as well as the reviewers for their constructive comments.

REFERENCES

[1] M. Adamik, A. P. Madsen, and M. Rehm, "Explainability in collaborative robotics: The effect of informing the user on task performance and trust," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 1252–1257.

[2] F. Alizadeh, P. Tolmie, M. Lee, P. Wintersberger, D. Pins, and G. Stevens, "Voice assistants' accountability through explanatory dialogues," in *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 2024, pp. 1–12.

[3] G. Angelopoulos, A. Rossi, C. Di Napoli, and S. Rossi, "You are in my way: non-verbal social cues for legible robot navigation behaviors," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 657–662.

[4] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.

[5] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 82, no. 4, pp. 1059–1086, 2020.

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[7] N. Attari, M. Heckmann, and D. Schlangen, "From explainability to explanation: Using a dialogue setting to elicit annotations with justifications," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 2019, pp. 331–335.

[8] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[9] A. Bansal, A. Farhadi, and D. Parikh, "Towards transparent systems: Semantic characterization of failure modes," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 2014, pp. 366–381.

[10] O. Bastani, C. Kim, and H. Bastani, "Interpretability via model extraction," *arXiv preprint arXiv:1706.09773*, 2017.

[11] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6541–6549.

[12] D. Bau, J.-Y. Zhu, J. Wulff, W. Peebles, H. Strobelt, B. Zhou, and A. Torralba, "Seeing what a gan cannot generate," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4502–4511.

[13] K. Baum, S. Mantel, E. Schmidt, and T. Speith, "From responsibility to reason-giving explainable artificial intelligence," *Philosophy & Technology*, vol. 35, no. 1, p. 12, 2022.

[14] S. Bensch and A. Eriksson, "Mining multi-modal communication patterns in interaction with explainable and non-explainable robots," *arXiv preprint arXiv:2312.14634*, 2023.

[15] A. Bertrand, T. Viard, R. Belloum, J. R. Eagan, and W. Maxwell, "On selective, mutable and dialogic xai: A review of what users say about different types of interactive explanations," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–21.

[16] D. Bohus, S. Andrist, A. Feniello, N. Saw, M. Jalobeanu, P. Sweeney, A. L. Thompson, and E. Horvitz, "Platform for situated intelligence," 2021.

[17] A. Bremers, A. Pabst, M. T. Parreira, and W. Ju, "Using social cues to recognize task failures for hri: A review of current research and future directions," *arXiv preprint arXiv:2301.11972*, 2023.

[18] D. A. Broniatowski and D. A. Broniatowski, *Psychological foundations of explainability and interpretability in artificial intelligence*. US Department of Commerce, National Institute of Standards and Technology, 2021.

[19] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, "Understanding the origins of bias in word embeddings," in *International conference on machine learning*. PMLR, 2019, pp. 803–811.

[20] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the shapley value based on sampling," *Computers & operations research*, vol. 36, no. 5, pp. 1726–1730, 2009.

[21] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati, "Plan explanations as model reconciliation–an empirical study," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Ieee, 2019, pp. 258–266.

[22] J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes, "Situation awareness-based agent transparency," *US Army Research Laboratory*, no. April, pp. 1–29, 2014.

[23] E. Committee, "A european approach to artificial intelligence," https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence, accessed: 2024-09-24.

[24] R. Confalonieri, T. R. Besold, T. Weyde, K. Creel, T. Lombrozo, S. T. Mueller, P. Shafto *et al.*, "What makes a good explanation? cognitive dimensions of explaining intelligent machines." in *CogSci*, 2019, pp. 25–26.

[25] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva, "Exploring the impact of fault justification in human-robot trust," in *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, 2018, pp. 507–513.

[26] D. Das, S. Banerjee, and S. Chernova, "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*, 2021, pp. 351–360.

[27] S. Dash, O. Gunluk, and D. Wei, "Boolean decision rules via column generation," *Advances in neural information processing systems*, vol. 31, 2018.

[28] M. M. De Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *2017 AAAI Fall Symposium Series*, 2017.

[29] ——, "People's explanations of robot behavior subtly reveal mental state inferences," in *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 2019, pp. 239–248.

[30] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," *Advances in neural information processing systems*, vol. 31, 2018.

[31] M. Diehl and K. Ramirez-Amaro, "Why did i fail? a causal-based method to find explanations for robot failures," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8925–8932, 2022.

[32] I. Donoso-Guzmán, J. Ooge, D. Parra, and K. Verbert, "Towards a comprehensive human-centred evaluation framework for explainable ai," in *World Conference on Explainable Artificial Intelligence*. Springer, 2023, pp. 183–204.

[33] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2013, pp. 301–308.

[34] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: a technique for explainable ai and its effects on human perceptions," in *Proceedings of the 24th international conference on intelligent user interfaces*, 2019, pp. 263–274.

[35] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[36] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[37] A. Ezenyilimba, M. Wong, A. Hehr, M. Demir, A. Wolff, E. Chiou, and N. Cooke, "Impact of transparency and explanations on trust and situation awareness in human–robot teams," *Journal of cognitive engineering and decision making*, vol. 17, no. 1, pp. 75–93, 2023.

[38] A. Fisher, C. Rudin, and F. Dominici, "Model class reliance: Variable importance measures for any machine learning model class, from the "rashomon" perspective," *arXiv preprint arXiv:1801.01489*, vol. 68, pp. 331–346, 2018.

[39] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2950–2958.

[40] R. Fong and A. Vedaldi, "Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8730–8738.

[41] F. Förster, M. Romeo, P. Holthaus, B. Nesset, M. J. Galvez Trigo, C. Dondrup, and J. E. Fischer, "Working with troubles and failures in conversation between humans and robots," in *Proceedings of the 5th International Conference on Conversational User Interfaces*, 2023, pp. 1–4.

[42] M. Fraile, C. Fawcett, J. Lindblad, N. Sladoje, and G. Castellano, "End-to-end learning and analysis of infant engagement during guided play: Prediction and explainability," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 444–454.

[43] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[44] K. Gavriilidis, A. Munafo, W. Pang, and H. Hastie, "A surrogate model framework for explainable autonomous behaviour," *arXiv preprint arXiv:2305.19724*, 2023.

[45] B. Ghai, Q. V. Liao, Y. Zhang, R. Bellamy, and K. Mueller, "Explainable active learning (xal): An empirical study of how local explanations impact annotator experience," *arXiv preprint arXiv:2001.09219*, 2020.

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

[47] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[48] J. M. Górriz, I. Álvarez-Illán, A. Álvarez-Marquina, J. E. Arco, M. Atzmueller, F. Ballarini, E. Barakova, G. Bologna, P. Bonomini, G. Castellanos-Dominguez *et al.*, "Computational approaches to explainable artificial intelligence: advances in theory, applications and trends," *Information Fusion*, vol. 100, p. 101945, 2023.

[49] J. Götze and D. Schlangen, ""why do you say so?" dialogical classification explanations in the wild and elicited through classification games," in *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*, 2023.

[50] M. Guesmi, M. A. Chatti, S. Joarder, Q. U. Ain, R. Alatrash, C. Siepmann, and T. Vahidi, "Interactive explanation with varying level of details in an explainable scientific literature recommender system," *International Journal of Human–Computer Interaction*, pp. 1–22, 2023.

[51] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," *arXiv preprint arXiv:1805.10820*, 2018.

[52] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[53] S. Guo, S. Zhang, W. Sun, P. Ren, Z. Chen, and Z. Ren, "Towards explainable conversational recommender systems," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2786–2795.

[54] A. Halilovic, V. Chandrayan, and S. Krivic, "Exploring the impact of explanation representation on user satisfaction in robot navigation," in *Proceedings of the 2024 International Symposium on Technological Advances in Human-Robot Interaction*, 2024, pp. 1–9.

[55] A. Halilovic and F. Lindner, "Visuo-textual explanations of a robot's navigational choices," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 531–535.

[56] Z. Han, D. Giger, J. Allspaw, M. S. Lee, H. Admoni, and H. A. Yanco, "Building the foundation of robot explanation generation using behavior trees," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 3, pp. 1–31, 2021.

[57] Z. Han and H. Yanco, "Communicating missing causal information to explain a robot's past behavior," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 1, pp. 1–45, 2023.

[58] S. Hart, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Human mental workload/Elsevier*, 1988.

[59] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 2017, pp. 303–312.

[60] T. Hellström, "The relevance of causation in robotics: A review, categorization, and analysis," *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 238–255, 2021.

[61] A. Henelius, K. Puolamäki, H. Boström, L. Asker, and P. Papapetrou, "A peek into the black box: exploring classifiers by randomization," *Data mining and knowledge discovery*, vol. 28, pp. 1503–1529, 2014.

[62] D. C. Hernandez-Bocanegra and J. Ziegler, "Explaining recommendations through conversations: Dialog model and the effects of interface type and degree of interactivity," *ACM Transactions on Interactive Intelligent Systems*, vol. 13, no. 2, pp. 1–47, 2023.

[63] R. R. Hoffman, G. Klein, and S. T. Mueller, "Explaining explanation for "explainable ai"," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 62, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2018, pp. 197–201.

[64] S. Honig and T. Oron-Gilad, "Understanding and resolving failures in human-robot interaction: Literature review and model development," *Frontiers in psychology*, vol. 9, p. 861, 2018.

[65] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[66] U. Johansson and L. Niklasson, "Evolving decision trees using oracle guides," in *2009 IEEE Symposium on Computational Intelligence and Data Mining*. IEEE, 2009, pp. 238–244.

[67] A. Kasirzadeh, "Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence," *arXiv preprint arXiv:2103.00752*, 2021.

[68] E. M. Kenny, M. Tucker, and J. Shah, "Towards interpretable deep reinforcement learning with human-friendly prototypes," in *The Eleventh International Conference on Learning Representations*, 2023.

[69] P. Khanna, E. Yadollahi, M. Björkman, I. Leite, and C. Smith, "Effects of explanation strategies to resolve failures in human-robot collaboration," in *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 1829–1836.

[70] B. Kim, E. Reif, M. Wattenberg, S. Bengio, and M. C. Mozer, "Neural networks trained on natural scenes exhibit gestalt closure," *Computational Brain & Behavior*, vol. 4, no. 3, pp. 251–263, 2021.

[71] B. Kim, C. Rudin, and J. A. Shah, "The bayesian case model: A generative approach for case-based reasoning and prototype classification," *Advances in neural information processing systems*, vol. 27, 2014.

[72] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[73] S. S. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández, "" help me help the ai": Understanding how explainability can support human-ai interaction," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–17.

[74] R. Kocielnik, S. Amershi, and P. N. Bennett, "Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[75] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan *et al.*, "Captum: A unified and generic model interpretability library for pytorch," *arXiv preprint arXiv:2009.07896*, 2020.

[76] D. Kontogiorgos, "Explanations as communicative acts in human-robot miscommunication," in *ROMAN 2023 - Workshops*, 2023.

[77] ——, "Utilising explanations to mitigate robot conversational failures," *arXiv preprint arXiv:2307.04462*, 2023.

[78] D. Kontogiorgos, A. Pereira, and J. Gustafson, "Estimating uncertainty in task-oriented dialogue," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 414–418.

[79] D. Kontogiorgos, A. Pereira, B. Sahindal, S. van Waveren, and J. Gustafson, "Behavioural responses to robot conversational failures," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 53–62.

[80] D. Kontogiorgos and D. Schlangen, "Explainable embodied intelligence for collaborative robots: An interactive approach," in *IPrA 2023*, 2023.

[81] D. Kontogiorgos, M. Tran, J. Gustafson, and M. Soleymani, "A systematic cross-corpus analysis of human reactions to robot conversational failures," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 112–120.

[82] D. Kontogiorgos, S. Van Waveren, O. Wallberg, A. Pereira, I. Leite, and J. Gustafson, "Embodiment effects in interactions with failing robots," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–14.

[83] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 5686–5697.

[84] R. Krishnan, G. Sivakumar, and P. Bhattacharya, "Extracting decision trees from trained neural networks," *Pattern recognition*, vol. 32, no. 12, 1999.

[85] M. Kwon, S. H. Huang, and A. D. Dragan, "Expressing robot incapability," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 87–95.

[86] V. Lai, Y. Zhang, C. Chen, Q. V. Liao, and C. Tan, "Selective explanations: Leveraging human input to align explainable ai," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW2, pp. 1–35, 2023.

[87] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 2, 2017, pp. 4762–4763.

[88] D. Lewis, "Causal explanation," *Philosophical Papers/Oxford University Press*, 1986.

[89] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–15.

[90] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *Proceedings of the 11th international conference on Ubiquitous computing*, 2009, pp. 195–204.

[91] F. Lindner and C. Olz, "Step-by-step task plan explanations beyond causal links," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 45–51.

[92] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[93] G. Liu, J. Zhang, A. B. Chan, and J. Hsiao, "Human attention-guided explainable ai for object detection," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 45, no. 45, 2023.

[94] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 623–631.

[95] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022.

[96] M. B. Luebbers, A. Tabrez, K. Ruvane, and B. Hayes, "Autonomous justification for enabling explainable decision support in human-robot teaming." in *Robotics: Science and Systems*, 2023.

[97] S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[98] S. M. Faas, J. Kraus, A. Schoenhals, and M. Baumann, "Calibrating pedestrians' trust in automated vehicles: does an intent display in an external hmi support trust calibration and safe crossing behavior?" in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–17.

[99] S. Madan, M. Gahalawat, T. Guha, R. Goecke, and R. Subramanian, "Explainable human-centered traits from head motion and facial expression dynamics," *arXiv preprint arXiv:2302.09817*, 2023.

[100] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2018, pp. 3237–3243.

[101] H. Mellmann, P. Arbuzova, D. Kontogiorgos, M. Yordanova, J. X. Haensel, V. V. Hafner, and J. J. Bryson, "Effects of transparency in humanoid robots-a pilot study," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 750–754.

[102] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological review*, vol. 63, no. 2, p. 81, 1956.

[103] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[104] T. Miller, P. Howe, and L. Sonenberg, "Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences," *arXiv preprint arXiv:1712.00547*, 2017.

[105] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 279–288.

[106] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 607–617.

[107] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *International journal of man-machine studies*, vol. 27, no. 5-6, pp. 527–539, 1987.

[108] A. Nguyen, J. Yosinski, and J. Clune, "Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks," *arXiv preprint arXiv:1602.03616*, 2016.

[109] M. N. Nicolescu and M. J. Mataric, "Learning and interacting in human-robot domains," *IEEE Transactions on Systems, man, and Cybernetics-part A: Systems and Humans*, vol. 31, no. 5, pp. 419–430, 2001.

[110] S. Nikolaidis, S. Nath, A. D. Procaccia, and S. Srinivasa, "Game-theoretic modeling of human adaptation in human-robot collaboration," in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 2017, pp. 323–331.

[111] V. V. Ramaswamy, S. S. Kim, N. Meister, R. Fong, and O. Russakovsky, "Elude: Generating interpretable explanations via a decomposition into labelled and unlabelled features," *arXiv preprint arXiv:2206.07690*, 2022.

[112] M. T. Ribeiro, S. Singh, and C. Guestrin, "" 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*, 2016, pp. 1135–1144.

[113] ——, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[114] D. A. Robb, X. Liu, and H. Hastie, "Explanation styles for trustworthy autonomous systems," in *22nd International Conference on Autonomous Agents and Multiagent Systems 2023*. Association for Computing Machinery, 2023, pp. 2298–2300.

[115] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Autonomous agents and multi-agent systems*, vol. 33, pp. 673–705, 2019.

[116] F. Sado, C. K. Loo, W. S. Liew, M. Kerzel, and S. Wermter, "Explainable goal-driven agents and robots-a comprehensive review," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–41, 2023.

[117] S. Saha, P. Hase, N. Rajani, and M. Bansal, "Are hard examples also harder to explain? a study with human and model-generated explanations," *arXiv preprint arXiv:2211.07517*, 2022.

[118] T. Sakai and T. Nagai, "Explainable autonomous robots: a survey and perspective," *Advanced Robotics*, vol. 36, no. 5-6, pp. 219–238, 2022.

[119] L. Sanneman and J. A. Shah, "The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems," *International Journal of Human–Computer Interaction*, vol. 38, no. 18-20, pp. 1772–1788, 2022.

[120] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," *International journal of computer vision*, vol. 128, pp. 336–359, 2020.

[121] R. Setchi, M. B. Dehkordi, and J. S. Khan, "Explainable robotics in human-robot interactions," *Procedia Computer Science*, vol. 176, pp. 3057–3066, 2020.

[122] J. Shah and C. Breazeal, "An empirical analysis of team coordination behaviors and action planning with application to human–robot teaming," *Human factors*, vol. 52, no. 2, pp. 234–245, 2010.

[123] L. S. Shapley, "A value for n-person games," *Contribution to the Theory of Games*, vol. 2, 1953.

[124] V. Shitole, F. Li, M. Kahng, P. Tadepalli, and A. Fern, "One explanation is not enough: structured attention graphs for image classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 352–11 363, 2021.

[125] A. Silva, P. Tambwekar, M. Schrum, and M. Gombolay, "Towards balancing preference and performance through adaptive personalized explainability," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 658–668.

[126] A. Simkute, E. Luger, B. Jones, M. Evans, and R. Jones, "Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable," *Journal of Responsible Technology*, vol. 7, p. 100017, 2021.

[127] L. Simon, C. Guérin, P. Rauffet, C. Chauvin, and É. Martin, "How humans comply with a (potentially) faulty robot: Effects of multidimensional transparency," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 4, pp. 751–760, 2023.

[128] K. Simonyan, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[129] A. Smith-Renner, R. Fan, M. Birchfield, T. Wu, J. Boyd-Graber, D. S. Weld, and L. Findlater, "No explainability without accountability: An empirical study of explanations and feedback in interactive ml," in *Proceedings of the 2020 chi conference on human factors in computing systems*, 2020, pp. 1–13.

[130] F. Sovrano and F. Vitali, "Generating user-centred explanations via illocutionary question answering: From philosophy to interfaces," *ACM Transactions on Interactive Intelligent Systems*, vol. 12, no. 4, pp. 1–32, 2022.

[131] R. Srinivasan and A. Chander, "Explanation perspectives from the cognitive sciences—a survey," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 4812–4818.

[132] M. Stiber, R. Taylor, and C.-M. Huang, "Modeling human response to robot errors for timely error detection," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 676–683.

[133] M. Stiber, R. H. Taylor, and C.-M. Huang, "On using social signals to enable flexible error-aware hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 222–230.

[134] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and information systems*, vol. 41, pp. 647–665, 2014.

[135] A. Tabrez, S. Agrawal, and B. Hayes, "Explanation-based reward coaching to improve human performance via reinforcement learning," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2019, pp. 249–257.

[136] A. Tabrez, M. B. Luebbers, and B. Hayes, "Descriptive and prescriptive visual guidance to improve shared situational awareness in human-robot teaming," in *Proceedings of the 21st international conference on autonomous agents and multiagent systems*, 2022, pp. 1256–1264.

[137] P. Tambwekar and M. Gombolay, "Towards reconciling usability and usefulness of explainable ai methodologies," *arXiv preprint arXiv:2301.05347*, 2023.

[138] J. E. T. Taylor and G. W. Taylor, "Artificial cognition: How experimental psychology can help generate explainable artificial intelligence," *Psychonomic Bulletin & Review*, vol. 28, no. 2, pp. 454–475, 2021.

[139] theguardian.com, ""it's the robot we were all expecting – like c3po": why aren't humanoids in our homes yet?" https://www.theguardian.com/science/2024/sep/22/its-the-robot-we-were-all-expecting-like-c3po-why-arent\-humanoids-in-our-homes-yet, accessed: 2024-09-22.

[140] S. Thellman and M. M. De Graaf, "The challenges of first-and second-order belief reasoning in explainable human-robot interaction," in *ICRA2023 Workshop on Explainable Robotics*, 2023.

[141] S. Thellman, A. Silvervarg, and T. Ziemke, "Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots," *Frontiers in psychology*, vol. 8, p. 1962, 2017.

[142] G. Tolomei, F. Silvestri, A. Haines, and M. Lalmas, "Interpretable predictions of tree-based ensembles via actionable feature tweaking," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 465–474.

[143] E. Union, "Gdpr," https://gdpr.eu/, accessed: 2024-09-24.

[144] L. Wachowiak, A. Coles, G. Canal, and O. Celiktutan, "A taxonomy of explanation types and need indicators in human–agent collaborations," *International Journal of Social Robotics*, pp. 1–12, 2024.

[145] L. Wachowiak, A. Fenn, H. Kamran, A. Coles, O. Celiktutan, and G. Canal, "When do people want an explanation from a robot?" in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 752–761.

[146] L. Wachowiak, P. Tisnikar, G. Canal, A. Coles, M. Leonetti, and O. Celiktutan, "Predicting when and what to explain from multimodal eye tracking and task signals," *IEEE Transactions on Affective Computing*, 2024.

[147] H. Wachsmuth and M. Alshomary, "" mama always had a way of explaining things so i could understand": A dialogue corpus for learning to construct explanations," *arXiv preprint arXiv:2209.02508*, 2022.

[148] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[149] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–15.

[150] J. Wang, Y. Liu, T. Yue, C. Wang, J. Mao, Y. Wang, and F. You, "Robot transparency and anthropomorphic attribute effects on human–robot interactions," *Sensors*, vol. 21, no. 17, p. 5722, 2021.

[151] X. Wang and M. Yin, "Effects of explanations in ai-assisted decision making: Principles and comparisons," *ACM Transactions on Interactive Intelligent Systems*, vol. 12, no. 4, pp. 1–36, 2022.

[152] ——, "Watch out for updates: Understanding the effects of model explanation updates in ai-assisted decision making," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–19.

[153] Y. Wang and S. You, "Enhancing robot explainability in human-robot collaboration," in *International Conference on Human-Computer Interaction*. Springer, 2023, pp. 236–247.

[154] D. Wei, S. Dash, T. Gao, and O. Gunluk, "Generalized linear rule models," in *International conference on machine learning*. PMLR, 2019, pp. 6687–6696.

[155] K. Weitz, D. Schiller, R. Schlagowski, T. Huber, and E. André, ""let me explain!": exploring the potential of virtual agents in explainable ai interaction design," *Journal on Multimodal User Interfaces*, vol. 15, no. 2, pp. 87–98, 2021.

[156] B. Yao, P. Sen, L. Popa, J. Hendler, and D. Wang, "Are human explanations always helpful? towards objective evaluation of human natural language explanations," *arXiv preprint arXiv:2305.03117*, 2023.

[157] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[158] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting cnns via decision trees," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6261–6270.

[159] W. Zhang and B. Y. Lim, "Towards relatable explainable ai with the perceptual process," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–24.

[160] X. Zhang, A. Solar-Lezama, and R. Singh, "Interpreting neural network judgments via minimal, stable, and symbolic corrections," *Advances in neural information processing systems*, vol. 31, 2018.

[161] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[162] B. Zhou, Y. Sun, D. Bau, and A. Torralba, "Interpretable basis decomposition for visual explanation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 119–134.

[163] Z.-H. Zhou, Y. Jiang, and S.-F. Chen, "Extracting symbolic rules from trained neural network ensembles," *Ai Communications*, vol. 16, no. 1, pp. 3–15, 2003.