Automatic Early Detection of Explanation Needs in Human–Robot Interaction

Dimosthenis Kontogiorgos
dimos@csail.mit.edu
Massachusetts Institute of Technology
Cambridge, MA, USA
KTH Royal Institute of Technology
Stockholm, Sweden

Joakim Gustafson
jkgu@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

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

Abstract

Enabling robots to display the reasoning behind decisions requires them to detect when explanations are needed by users. A crucial driver of explanation need is that it often manifests implicitly: users exhibit behavioural signals indicating misalignment well before they explicitly request an explanation. Psychological studies show that in human interactions, such needs are sensed through multimodal cues and addressed through the co-construction of explanations in real time. Building on this, we introduce an approach for the early detection of explanation needs in HRI. Our method recognises when an explanation is likely to become necessary, enabling robots to act proactively. We evaluate the approach on an existing HRI dataset using features describing facial expressions, body movement, and vocal behaviour, combined with time-series classification techniques. Our results show that different classes of learning algorithms (unsupervised anomaly-based methods and supervised classification models) offer complementary strengths for detecting explanation needs. In particular, unsupervised methods enable early warning signals when labels are unavailable, while supervised models provide stronger discrimination (AUROC 0.7) when annotated data is available. We discuss the implications of these findings for the development of explanation-capable robots and outline future directions for proactive explanations in HRI.

Keywords

explainability, social signal processing, multimodality

1 Introduction

Modern computer systems are increasingly envisioned as capable of processing and understanding human social signals. Such systems can leverage information about users' social perceptions to adapt their behaviour, supporting predictions about user actions and experiences during interaction. Within this context, explanation needs can be understood as emergent user states, dynamic constructs that reflect properties of the user arising from their interaction with the system. Providing explanations has become a mainstream topic in machine learning and has gained substantial traction in human-computer interaction research [1]. However, despite this growing interest, fewer than 1% of explainable AI studies validate their approaches through empirical evaluation with human users [69]. As a result, our understanding of when explanations are necessary remains limited. In particular, relatively little research has examined when users exhibit behavioural signals indicating that explanations may be required. Consistent with findings from studies of human communication, explanations are not isolated events but are embedded within the ongoing dynamics of interaction [27, 41].

This paper focuses on the early detection of such interaction states, enabling robots to recognise implicit behavioural signals and provide explanations proactively, as humans do. To investigate how implicit explanation-need behaviour manifests, we analyse an existing multimodal HRI dataset [45], in which participants explicitly request explanations following robot errors. Using this dataset, we examine how early such explanation needs can be detected prior to explicit verbal requests. We consider two complementary learning strategies: an unsupervised anomaly detection approach evaluated against the provided labels, and a supervised learning method trained directly on ground-truth instances of explanation need. Our findings indicate that both approaches can detect explanation needs before they are explicitly verbalised. The unsupervised method demonstrates stronger generalisation capabilities but lower overall predictive performance, whereas the supervised model achieves higher detection accuracy but relies on labelled data, which may be difficult to obtain in real-world deployments.

Additionally, we provide an analysis of which human behavioural signals offer the most informative cues for detecting explanation needs. Given the high dimensionality of the multimodal feature space, we employ computationally efficient models based on Isolation Forests and Random Forests. The contributions of this paper are twofold: (1) a multimodal framework for predicting explanation needs in human-robot interaction, and (2) a participant-independent evaluation of the proposed framework on a dataset comprising 33 users interacting with a robot arm [45]. Overall, our results highlight the importance of integrating multimodal behavioural signals, including facial expressions and vocal behaviour, for the early detection of explanation needs in HRI.

2 Related Work

A fundamental challenge in human-robot interaction arises when robots are deployed in real-world settings, as human users often lack clear expectations regarding the robot's capabilities and behaviour [62]. Explanations can play a critical role in aligning user expectations with the robot's actual functionality. More broadly, explainable AI has been widely studied as a means of interpreting the black-box nature of AI models [4, 30, 50, 54]. Explanations have been investigated across a diverse range of domains in HCI and affective computing [2, 5, 26, 35, 39, 40, 66, 70, 74, 75, 78]. Yet, despite this extensive body of work, explanations are most commonly treated as static artefacts. Such approaches do not adequately reflect the dynamic nature of human interaction. In practice, explanation needs evolve over time: an explanation may no longer be necessary

once generated, or it may require refinement to accommodate the user's changing interactional state. Crucially, our understanding of when explanations are needed during interaction remains limited.

While prior work has rarely focused on the direct detection of explanation needs, several related approaches examine users' reactions to unintended or faulty robot behaviour. Such studies are relevant, as they model human behavioural responses that may give rise to explanation needs [8, 19, 38, 42, 59, 65, 71]. Additionally, existing research highlights a growing interest in user-adaptive explanations, which leverage human input to tailor explanatory content dynamically [21, 25, 49]. From an interactional perspective, Rohlfing et al. [61] emphasise the sequential organisation of explanations, arguing that explanations are co-constructed processes that evolve during interaction. In this view, explanations may be continuously refined in response to user signals of uncertainty[1]. Such adaptive monitoring processes are consistent with psycholinguistic accounts of communication, which describe how speakers collaboratively manage understanding through grounding and utterance reformulation when co-present [15].

Users often request explanations in response to explicit questions ("Why did you do that?" or "Why did you make this decision?") [22, 29, 51, 52]. But *what happens when the question is never asked, yet an explanation is implicitly expected*? Given that a substantial portion of human communication is non-verbal [10], such expectations may frequently be conveyed through non-verbal behaviour rather than through direct verbal requests. In many AI systems, explanations are presented automatically alongside model predictions. However, users may also value access to explanations on demand, suggesting that the question of when AI systems should be explainable fundamentally concerns the balance between automatic and user-initiated explanations [31, 37].

Across domains, explanation requirements vary depending on the nature of the interaction. In medical AI, systems commonly provide automatic explanations following decision-making processes [7, 23, 76]. Similar patterns are observed in entertainment and recommendation systems [17, 21, 48, 53, 55, 63]. In contrast, task-driven domains often favour on-demand explanations, as users prefer to retain primary control over the interaction [9, 56]. Continuous decision-making contexts, such as financial systems and autonomous driving, predominantly employ automatic explanations [6, 11, 13, 14, 60, 64]. Other domains, including human resource management and e-commerce, typically require a combination of automatic and on-demand explanations [9, 16, 21, 57, 79]. Overall, prior work suggests that explanations are most commonly delivered automatically, while supplementary contextual information is frequently expected on demand. Yet, *when* explanations should be proactively delivered remains insufficiently specified.

Despite substantial progress in generating robot explanations and detecting robot errors, it remains unclear whether:

- users explicitly communicate their need for explanations
- explanation needs can be inferred from implicit signals

While much prior research focuses on generating robot explanations or detecting robot errors and confusion-related behaviours, comparatively less attention has been given to determining *when* explanations are needed. Existing work has established robust paradigms for robot error detection and confusion detection [43, 44, 47, 67, 68, 73]. However, these approaches primarily model system failures rather than users' emerging explanation needs, or rely on third-person annotations of explanation need [72]. In practice, explanation needs do not arise solely from errors, and not all errors or confusion warrant an explanation. *Our dataset and analysis therefore complement prior work by modelling the implicit behavioural signals that precede explicit explanation requests.*

This paper does not aim to explain robot behaviour or address task-related errors. Instead, we focus on human behavioural signals that arise in response to interactional irregularities and may precede explicit explanation requests. Concretely, we investigate whether explanation needs can be detected from multimodal human signals before they are verbally articulated, through a machine learning analysis. More broadly, our work emphasises the benefits of robots equipped with predictive forward models capable of anticipating interactional outcomes. Within this perspective, explanations can be understood not merely as reactive outputs, but as interactional resources that support the establishment of common ground [18].

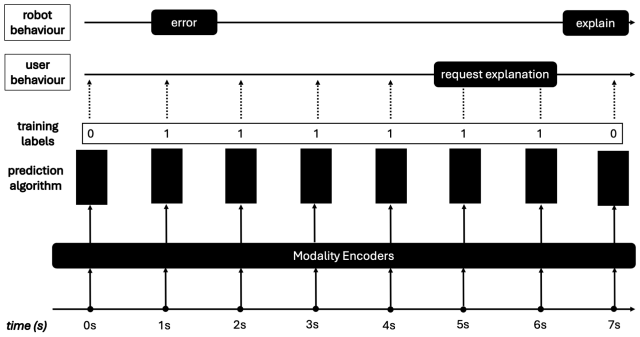

Figure 1: Formulation of early explanation-need detection. Positive labels denote explanation-need segments preceding the request. Early detection is operationalised as a threshold occurring within a lead window before the explicit request.

3 Problem Formulation: Explanation Need

Let $\{\mathbf{x}_t\}_{t=1}^{T}$ denote a temporally aligned multimodal time series, where $\mathbf{x}_t \in \mathbb{R}^d$ contains the features extracted at time t. For each interaction episode i, we observe an *explicit explanation request* time $t_r^{(i)}$, operationalised by the onset of the `explanation_request` (a $0 \rightarrow 1$ transition). A model produces a frame-level score $s_t = f(\mathbf{x}_{1:t})$ (or probability $\hat{p}_t = \sigma(s_t)$) indicating the likelihood that an explanation will be requested imminently. Given a decision threshold θ, we define the *detection time* for episode i as the first threshold crossing within a lead window of length $W > 0$ seconds prior to the request:

$$\hat{t}^{(i)} = \min \left\{ t \in [t_r^{(i)} - W,\ t_r^{(i)}) \;\middle|\; \hat{p}_t^{(i)} \geq \theta \right\}. \tag{1}$$

If no such t exists, we set $\hat{t}^{(i)} = \varnothing$.

[1] Uncertainty here refers to its interactional and participatory role in dialogue. While explanation generation is also closely related to model uncertainty [4], our focus is on observable behavioural signals arising during interaction.

We quantify *earliness* via the lead time (in seconds):

$$\ell^{(i)} = \begin{cases} t_r^{(i)} - \hat{t}^{(i)}, & \hat{t}^{(i)} \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

An episode is considered *successfully detected early* if a threshold crossing occurs within the lead window:

$$\mathbb{I}_{\text{det}}^{(i)}(W) = \begin{cases} 1, & \hat{t}^{(i)} \neq \varnothing, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

and the event-level detection rate is:

$$\text{DR}(W) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\text{det}}^{(i)}(W), \tag{4}$$

where N is the number of request events. We report $\text{DR}(W)$ together with the distribution of $\ell^{(i)}$ (e.g., median and mean lead time) over detected events. We set θ per fold as the q-th quantile of scores on training negatives (with $q = 0.99$), to control false alarms. Robot log entries with non-interaction system states were excluded, as these correspond to robot initialisation, termination, and idle behaviour. Similarly, intervals associated with explanation presentation were removed to avoid label leakage.

4 Method

4.1 Dataset

To evaluate explanation need in an interactional setting capable of eliciting a broad range of user reactions and behaviours associated with such needs, we used a dataset from Kontogiorgos and Shah [45, 46]. This study was approved by the Institutional Review Board (IRB) at Massachusetts Institute of Technology (MIT). The dataset comprises 33 participants (14 female, 19 male), with an average age of 30.3 years (±9.7). Participants reported high English fluency (4.6 ±0.5), technology experience (4.6 ±0.7), and comparatively lower robot experience (2.9 ±1.2) on a 1-5 scale. In the dataset, each experimental session lasted between 1 and 1.5 hours. Participants completed three robot-assisted assembly tasks, yielding 99 human-robot interaction sessions and 28.1 hours of multimodal recordings. The dataset was selected due to its interactional structure, including explicit explanation requests from users and task-related irregularities (robot errors [47]), which provide observable conditions under which explanation needs may emerge.

In the study (Figure 2), participants assembled structures made of PVC pipes with robotic assistance. The pipes were positioned between the participant and the robot, and the robot handed over components upon request. The interactions were recorded using two depth cameras capturing participants' body pose and facial action units, along with a close-talking microphone capturing speech and acoustic features. The interactions included moments in which the robot failed, naturally creating situations of uncertainty. Participants were informed that, in such moments, they could verbally request the robot to provide explanations, which triggered the robot to display an explanation on a screen. We extracted instances where a robot error occurred and a participant explicitly requested

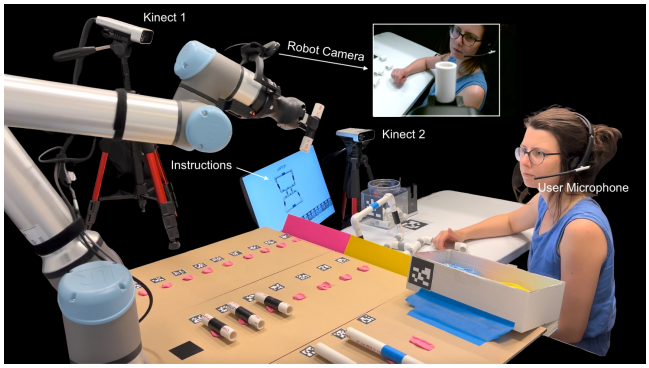

Figure 2: Experimental setup. Multimodal sensor streams captured user behaviour, including body pose, facial expressions, gaze, and speech. Reprinted with permission from [45].

an explanation. The interactions were temporally aligned and timestamped across multiple multimodal streams captured through several sensors. In the following, we describe the multimodal machine learning pipeline developed to model these signals.

The interaction framework leveraged Microsoft Azure Automatic Speech Recognition (ASR) for speech-to-text processing, GPT-4 for dialogue management, and Azure Text-to-Speech (TTS) for vocal output. The extraction of labels was performed as follows (Figure 1). All instances in which a robot error occurred and a participant explicitly requested an explanation were identified. For each instance, we labelled the time-series data spanning from the moment the error occurred until the participant's explanation request as *explanation-need* segments. Our modelling objective is to determine how early the need for explanation can be detected prior to explicit verbalisation by the participant.

4.2 Feature Extraction

All multimodal signals were temporally aligned at the recording rate of 3 Hz. Sensor data from depth cameras and microphones were used for feature extraction. Voice activity detection was applied to isolate voiced segments, enabling the computation of acoustic features from speech behaviour. Our representation combines high-level behavioural descriptors with low-level video-based embeddings, allowing the models to capture both interpretable interaction cues and rich latent multimodal patterns (Figure 3, Table 6).

Speech Acoustics. From the audio signal, we extract features based on the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [24], which comprises descriptors related to frequency, energy, spectral balance, pitch, jitter, shimmer, harmonic properties, and temporal characteristics of the voice. GeMAPS was selected due to its established use in modelling affective and paralinguistic speech characteristics. We also extract audio features capturing speaker activity (robot vs human). The extracted features capture both interaction structure (speaker activity) and paralinguistic properties.

Linguistic and Interactional Features. Automatic Speech Recognition (ASR) transcripts are used to derive linguistic and temporal descriptors capturing interaction dynamics. These include features encoding temporal context, disfluencies, speech rate, robot activity, and recent-error context. We further extract part-of-speech (POS)

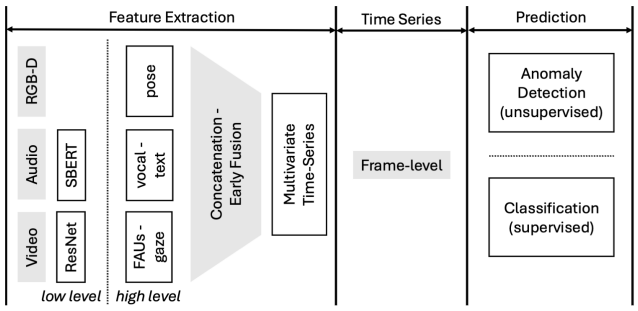

Figure 3: Overview of the multimodal ML pipeline. Low-level representations are combined with high-level behavioural descriptors, fused via early concatenation at the frame level.

Hyperparameter	Isolation Forest	Random Forest
Configurations evaluated	54	48
$n_{\text{estimators}}$	100	200
max_samples	0.5	–
max_features	0.5	0.5
max_depth	–	20
min_samples_split	–	5
min_samples_leaf	–	2
contamination	0.005	–
anomaly threshold	0.99 quantile	–

Table 1: Best-performing hyperparameters obtained through grid search for the Isolation Forests and Random Forests.

statistics, sentiment-related features, and sentence-level embeddings (Sentence-BERT embeddings [58]), enabling the models to capture both structural and semantic properties of user speech.

Facial features. Facial Action Units (FAUs) are extracted using OpenFace [3]. For each FAU, we retain a binary feature indicating activation and a continuous feature representing the intensity of the activation, capturing the presence/magnitude of facial behaviour.

Pose features. Using the Kinect SDK, we extract pose features derived from 3D body keypoints. These features capture users' body configuration and movement throughout the interaction, enabling the modelling of spatial and kinematic properties of user behaviour. In addition to static skeletal descriptors, we compute kinematic features capturing motion dynamics, including velocity, acceleration, and jerk of major body joints. We further derive relational features encoding spatial configuration and user–robot geometry (joint distances, posture alignment), together with event-based behavioural indicators (hand raises, pointing gestures).

Gaze features. Gaze behaviour is modelled using probability estimates describing users' visual attention throughout the interaction. These features encode whether participants orient their gaze towards the robot, the workspace, task-relevant objects, or the screen.

Visual Embeddings. We extract low-level visual representations using a ResNet-based convolutional neural network [32]. Frame-level embeddings derived from the visual stream capture rich latent structure in users' movement, providing a representation of visual dynamics, complementing the high-level behavioural features.

4.3 Multimodal Representation and Prediction

All extracted features are fused via feature-level concatenation and processed in three stages (Figure 3): (1) feature extraction, (2) time-series representation, and (3) classification. The prediction task aims to estimate whether a user will produce an explicit explanation request in the near future. The model is designed for real-time deployment and leverages features describing both the user's behavioural state and the system's own behaviour. System-level features include indicators of whether the robot is currently speaking, when it last spoke, and the elapsed time since the most recent explanation. These variables allow the model to account for interactional context while preventing redundant or temporally inappropriate explanation behaviour. Given multimodal time-series data capturing user behaviour during interactions involving robot errors, the objective is to detect moments when users require explanations, as well as how early such needs can be predicted. This formulation enables real-time inference, supporting adaptive explanation strategies in interactive robotic systems.

This formulation also allows predictions to be updated continuously; given the sampling rate of 3Hz, the model generates updated estimates three times per second. The model outputs a score for each window, and for each true event onset we evaluate whether the score exceeds a predefined threshold within a lead window preceding the explanation request. Event-level metrics are computed based on the temporal overlap between predicted above-threshold segments and labelled pre-event intervals. Thus, early detection is defined by threshold crossings within pre-event windows. Missing values are handled by computing per-feature medians on the training data, with NaN values imputed using these medians (falling back to zero for features with entirely missing values).

4.4 Models and Experimental Setup

We conduct experiments using two complementary learning approaches. First, we employ an unsupervised anomaly detection method based on Isolation Forests (IF), which is evaluated against the ground-truth labels. Second, we use a supervised ensemble classification model based on Random Forests (RF). Both IF and RF are ensemble tree models; however, IF operate in an unsupervised manner by detecting anomalies, while RF rely on labelled data to perform supervised classification. IF was trained only on negative frames, then scored on all frames. Hyperparameter tuning was performed for both models using grid search. A total of 54 configurations were evaluated for Isolation Forest and 48 for Random Forest. The best-performing configurations are reported in Table 1. These hyperparameters were used in the Leave-One-Participant-Out (LOPO) cross-validation evaluation for both models.

Both models are implemented in Python and trained on an NVIDIA Spark 20-core CPU and 128GB of memory. The frame-level label distribution further highlights the strong class imbalance inherent in the dataset. Across all aligned frames, the explanation-need label comprised 6,304 positive sample frames and 165,519 negative sample frames. This imbalance motivates the use of evaluation metrics robust to skewed class distributions. We also report feature importance alongside AUROC and AUPRC metrics to account for the imbalanced class distributions. Feature importance

Metric	Baseline	Isolation Forest	Random Forest
AUROC (mean)	0.50	0.52 ± 0.11	0.70 ± 0.10
AUPRC (mean)	0.04	0.07 ± 0.05	0.14 ± 0.09

Table 2: Frame-level classification and ranking performance. The baseline corresponds to the majority ranking.

Metric	Isolation Forest	Random Forest
Detection Rate	0.90	0.52
Median Lead Time (s)	3.33	6.00
Mean Lead Time (s)	3.85	6.51

Table 3: Early detection performance using a 10s lead window.

was computed using impurity-based importance (Mean Decrease in Impurity with Gini criterion). In LOPO, importances were computed per fold and then aggregated as mean ± std across folds. Early-detection performance is quantified using event-level detection rate and lead-time statistics (median and mean), computed relative to explicit explanation-request onsets from the users.

4.5 Motivation for Explanation Requests

While the present work does not explicitly model users' motivations for requesting explanations, the dataset includes self-reported responses from the Curiosity Checklist [36]. Participants indicated their motivation for requesting explanations by reporting one or more predefined reasons related to curiosity and explanation-seeking behaviour. These responses are analysed descriptively to contextualise the observed explanation requests in the interaction.

5 Results

The performance of the unsupervised Isolation Forest and supervised Random Forest models is summarised in Tables 2 and 3. To characterise user behaviour, we analysed the temporal relationship between robot errors and explicit explanation requests. On average, participants requested explanations 11.87*s* after an error occurred (±6.20*s*), indicating substantial variability in response timing.

5.1 Classification and Ranking Performance

At the frame level, RF consistently outperforms IF across classification and ranking metrics, indicating that RF provides better discrimination under class imbalance, suggesting improved separability between explanation-need and non-need frames.

5.2 Behavioural Evidence of Explanation Need

Early detection analysis reveals complementary model behaviours. Isolation Forest achieves markedly higher detection rates, identifying most explanation request events within the lead window. However, RF produces substantially longer lead times when detections occur, suggesting more temporally stable predictions. This difference reflects the inherent trade-off between sensitivity and selectivity: IF frequently signals anomalies, whereas RF generates fewer but earlier predictions relative to explicit requests. Detection performance is quantified at the event level, ensuring that evaluation reflects the models' ability to anticipate explanation-request episodes rather than individual frame classifications.

Modality	Importance (%)
Visual Embeddings (ResNet)	63.61
Pose / Kinematics	20.39
Engineered NLP	10.91
Interaction Context / State	2.79
Gaze	1.18
Speech Acoustics	0.70
Facial Action Units	0.34
Text Embeddings (SBERT)	0.08

Table 4: Aggregated feature importance by modality (RF).

Motivation	Responses (%)
I want to know what the AI just did	24.2
I want to know that I understand this AI system correctly	36.4
I want to understand what the AI will do next	18.2
I want to know why the AI did not make some other decision	51.5
I want to know what the AI would have done if something had been different	36.4
I was surprised by the AI's actions and want to know what I missed	60.6

Table 5: Self-reported motivations for requesting explanations, revealing that explanation requests are predominantly associated with surprise and counterfactual reasoning mechanisms. *Responses were not mutually exclusive.

5.3 Feature Importance Analysis

To better understand which information sources drive explanation-need prediction, we aggregate feature importance scores by modality. Table 4 summarises the proportion of total importance attributed to each feature group. Visual embeddings derived from ResNet features dominate the model, accounting for approximately 63.6% of the total importance. This indicates that rich visual representations capture substantial behavioural information relevant to emerging explanation needs. Pose and kinematic features constitute the second most influential modality (20.4%), suggesting that users' body configuration and movement dynamics provide meaningful cues. Engineered NLP features contribute 10.9% of the total importance, highlighting that linguistic structure, disfluencies, and sentiment-based descriptors encode complementary signals. Interaction context features (temporal variables describing recent system activity) provide a smaller but consistent contribution (2.8%), reflecting the relevance of conversational and task-state dynamics. Gaze, speech acoustics, and facial action units collectively account for a modest proportion of importance. While individually weaker, these modalities likely provide stabilising or disambiguating information when combined with stronger visual and pose-based representations. Overall, the results suggest that explanation-need prediction is primarily driven by multimodal behavioural representations, with visual and kinematic signals serving as the dominant sources of predictive information.

5.4 Reported Motivation for Explanation Need

Responses indicate that explanation requests were predominantly driven by surprise and uncertainty (Table 5). A chi-square goodness-of-fit test showed that the distribution of motivations deviated significantly from a uniform distribution ($\chi^2 = 11.16$, $p = 0.048$), suggesting that explanation requests were not randomly motivated but systematically associated with distinct cognitive mechanisms. Reported motivations included expectation violation, uncertainty, mental model monitoring, and counterfactual reasoning.

5.5 Summary of Findings

Overall, these results highlight that the supervised RF model achieves superior discriminative performance, particularly in ranking metrics, indicating improved modelling of explanation-need structure. In contrast, the unsupervised IF model exhibits higher detection sensitivity, highlighting its ability to capture deviations in user behaviour without reliance on labels. These results underscore the complementary strengths of anomaly detection and supervised classification for modelling explanation-related interaction dynamics.

6 Discussion

The results indicate that explanation needs can be detected prior to explicit requests using multimodal signals. Although frame-level classification performance remains modest due to class imbalance, both models exhibit reliable early-detection behaviour, suggesting that explanation-seeking behaviour presents measurable precursors. Specifically, the models detect instability in users' non-verbal signals several seconds before an explicit explanation request, with mean lead times of 3.8s for the unsupervised model and 6.5s for the supervised model. The importance of non-verbal cues is reflected in the models' reliance on visual and kinematic signals. The observed performance patterns further suggest a complementary relationship between anomaly detection and supervised classification, depending on label availability and the temporal requirements of early detection. When multimodal signals are jointly modelled, the proposed pipeline captures distinct dimensions of interaction, including kinematic variation, visual representations, and the linguistic structure of user speech preceding explicit explanation requests. The integration of these signals enables the model to develop a more robust representation of the user's interactional context and to anticipate uncertainty arising from robot errors. Such models can be deployed across a range of HRI scenarios, including settings that complement existing error detection frameworks [44, 67, 68, 73].

Notably, the predominance of surprise-driven requests suggests that explanations primarily function as mechanisms for repairing expectation violations and restoring alignment between users' mental models and robot behaviour. This observation complements the main modelling results by indicating that explanation requests often arise from cognitive states associated with uncertainty and expectation mismatch, which may manifest through the behavioural signals captured by the multimodal features used in the models.

Limitations and Future Work

The unsupervised approach, while conceptually appealing for real-world deployments, exhibited comparatively low predictive performance. This outcome is likely influenced by the pronounced class imbalance present in the dataset. Explanation requests represent relatively rare events in HRI, as the majority of interaction does not involve an explicit need for explanation. However, when explanation needs do arise, the associated behavioural signals tend to be strong, which partly motivates the relevance of anomaly detection frameworks. Future work should explore modelling approaches that explicitly capture temporal dynamics, including sequence-aware models and alternative stride-based representations [71]. While the present study focused on frame-level prediction, richer temporal modelling may better characterise the evolution of explanation need signals over time. Another limitation of the present evaluation is that no additional hold-out test set or nested cross-validation was used beyond the LOPO cross-validation procedure. This may introduce a modest risk of overfitting hyperparameters to the validation data. Furthermore, the models were trained and evaluated on a dataset involving collaborative, low-stakes decisions. It remains unclear how these findings extend to other interaction contexts, particularly those involving non-collaborative settings or higher-stakes decisions. In safety-critical applications, explanation needs may manifest differently, exhibiting stronger or more polarised behavioural patterns. We therefore advise caution when attempting to generalise these results beyond the studied interaction setting.

Additionally, the models developed in this work provide the opportunity to proactively construct explanations based on user signals. While the present study examines the timing of such signals, determining when to provide an explanation requires empirical investigation. Explanations delivered at inappropriate moments or without sufficient contextual grounding may produce unintended negative effects [34]. The central premise of the proposed models is to determine whether deviations emerge in users' behaviour or whether the interaction progresses as expected. Such deviations can be interpreted relative to interactional norms, including principles of cooperative communication [28]. Within this perspective, misalignment or breakdowns in coordination may naturally invite explanation-oriented behaviour. These findings imply that explanation need constitutes an emergent interactional state rather than a purely reactive event. At the same time, performance remains constrained by label scarcity and dataset-specific interaction dynamics.

Conclusion

In summary, determining when explanations are necessary is central to responsible AI deployment. Explanations promote transparency, support the identification of potential biases [12, 33], and shape how users interpret, trust, and act upon algorithmic decisions [20]. Beyond fostering trust, explanations in human-robot interactions provide insight into how robot decisions are formed, including the communication of uncertainties and potential biases. We hope this work encourages further research on proactive explanation models for real-world applications [77], supporting transparency, interpretability, improved understanding, and socially aware robot behaviour in both domestic and industrial settings.

Acknowledgments

We thank Mike Hagenow and Andre Pereira for insightful discussions and for their support throughout this work. We also thank the anonymous reviewers for their constructive comments.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[2] Fatemeh Alizadeh, Peter Tolmie, Minha Lee, Philipp Wintersberger, Dominik Pins, and Gunnar Stevens. 2024. Voice Assistants' Accountability through Explanatory Dialogues. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–12.

[3] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* 6, 2 (2016), 20.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[5] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R Eagan, and Winston Maxwell. 2023. 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*. 1–21.

[6] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[7] Dawn Branley-Bell, Rebecca Whitworth, and Lynne Coventry. 2020. User trust and understanding of explainable AI: exploring algorithm visualisations and user biases. In *International Conference on Human-Computer Interaction*. Springer, 382–399.

[8] Alexandra Bremers, Alexandria Pabst, Maria Teresa Parreira, and Wendy Ju. 2023. Using social cues to recognize task failures for hri: A review of current research and future directions. *arXiv preprint arXiv:2301.11972* (2023).

[9] Joost Broekens, Maaike Harbers, Koen Hindriks, Karel Van Den Bosch, Catholijn Jonker, and John-Jules Meyer. 2010. Do you get it? User-evaluated explainable BDI agents. In *German Conference on Multiagent System Technologies*. Springer, 28–39.

[10] Judee K Burgoon, David B Buller, Jerold L Hale, and Mark A de Turck. 1984. Relational messages associated with nonverbal behaviors. *Human communication research* 10, 3 (1984), 351–378.

[11] Larissa Chazette and Kurt Schneider. 2020. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering* 25, 4 (2020), 493–514.

[12] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[13] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I think i get your point, AI! the illusion of explanatory depth in explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 307–317.

[14] Douglas Cirqueira, Dietmar Nedbal, Markus Helfert, and Marija Bezbradica. 2020. Scenario-based requirements elicitation for user-centric explainable AI: A case in fraud detection. In *International cross-domain conference for machine learning and knowledge extraction*. Springer, 321–341.

[15] Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. (1991).

[16] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial intelligence* 298 (2021), 103503.

[17] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction* 18, 5 (2008), 455–496.

[18] Maartje MA De Graaf and Bertram F Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

[19] Maximilian Diehl and Karinne Ramirez-Amaro. 2022. Why did i fail? a causal-based method to find explanations for robot failures. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8925–8932.

[20] Mark Edmonds, Feng Gao, Hangxin Liu, Xu Xie, Siyuan Qi, Brandon Rothrock, Yixin Zhu, Ying Nian Wu, Hongjing Lu, and Song-Chun Zhu. 2019. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics* 4, 37 (2019), eaay4663.

[21] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. 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*. 263–274.

[22] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. 2021. Operationalizing human-centered perspectives in explainable AI. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–6.

[23] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *Proceedings of the 23rd international conference on intelligent user interfaces*. 211–223.

[24] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[25] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2020. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. *arXiv preprint arXiv:2001.09219* (2020).

[26] Juan M Górriz, Ignacio Álvarez-Illán, Agustín Álvarez-Marquina, Juan Eloy Arco, Martin Atzmueller, F Ballarini, Emilia Barakova, Guido Bologna, P Bonomini, Germán Castellanos-Dominguez, et al. 2023. Computational approaches to explainable artificial intelligence: advances in theory, applications and trends. *Information Fusion* 100 (2023), 101945.

[27] Jana Götze and David Schlangen. 2023. "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*.

[28] Herbert P Grice. 1975. Logic and conversation. In *Speech acts*. Brill, 41–58.

[29] Mouadh Guesmi, Mohamed Amine Chatti, Shoeb Joarder, Qurat Ul Ain, Rawaa Alatrash, Clara Siepmann, and Tannaz Vahidi. 2023. Interactive explanation with varying level of details in an explainable scientific literature recommender system. *International Journal of Human–Computer Interaction* (2023), 1–22.

[30] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI magazine* 40, 2 (2019), 44–58.

[31] AKM Bahalul Haque, AKM Najmul Islam, and Patrick Mikalef. 2023. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change* 186 (2023), 122120.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[33] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 771–787.

[34] Guy Hoffman, Maya Cakmak, and Crystal Chao. 2014. Timing in human-robot interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 509–510.

[35] Robert R Hoffman, Gary Klein, and Shane T Mueller. 2018. Explaining explanation for "explainable AI". In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 197–201.

[36] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[37] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[38] Parag Khanna, Elmira Yadollahi, Mårten Björkman, Iolanda Leite, and Christian Smith. 2023. 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, 1829–1836.

[39] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. " 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*. 1–17.

[40] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. 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*. 1–14.

[41] Dimosthenis Kontogiorgos. 2023. Explanations as communicative acts in human-robot miscommunication. In *ROMAN 2023-Workshops*.

[42] Dimosthenis Kontogiorgos. 2023. Utilising Explanations to Mitigate Robot Conversational Failures. *arXiv preprint arXiv:2307.04462* (2023).

[43] Dimosthenis Kontogiorgos, Andre Pereira, and Joakim Gustafson. 2019. Estimating uncertainty in task-oriented dialogue. In *2019 International Conference on Multimodal Interaction*. 414–418.

[44] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural responses to robot conversational failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 53–62.

[45] Dimosthenis Kontogiorgos and Julie Shah. 2025. Questioning the Robot: Using Human Non-verbal Cues to Estimate the Need for Explanations. In *2025 20th*

ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 717–728.

[46] Dimosthenis Kontogiorgos and Julie Shah. 2025. Socially-Aware Robot Explanations: Inferring Needs from Human Facial Expressions. In *Proceedings of the 3rd TRR Conference Contextualizing Explanations (ContEx)*.

[47] Dimosthenis Kontogiorgos, Minh Tran, Joakim Gustafson, and Mohammad Soleymani. 2021. A systematic cross-corpus analysis of human reactions to robot conversational failures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 112–120.

[48] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th international conference on intelligent user interfaces*. 379–390.

[49] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective explanations: Leveraging human input to align explainable ai. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.

[50] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable agency for intelligent autonomous systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31. 4762–4763.

[51] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[52] Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*. 195–204.

[53] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.

[54] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[55] Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. 2020. Exploring mental models for transparent and controllable recommender systems: a qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 183–191.

[56] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[57] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2021. Human-AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–15.

[58] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 3982–3992.

[59] David A Robb, Xingkun Liu, and Helen Hastie. 2023. Explanation styles for trustworthy autonomous systems. In *22nd International Conference on Autonomous Agents and Multiagent Systems 2023*. Association for Computing Machinery, 2298–2300.

[60] M Rodriguez-Sampaio, Mariano Rincón, Sonia Valladares-Rodríguez, and Margarita Bachiller-Mayoral. 2022. Explainable artificial intelligence to detect breast cancer: a qualitative case-based visual interpretability approach. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 557–566.

[61] Katharina J Rohlfing, Philipp Cimiano, Ingrid Scharlau, Tobias Matzner, Heike M Buhl, Hendrik Buschmeier, Elena Esposito, Angela Grimminger, Barbara Hammer, Reinhold Häb-Umbach, et al. 2020. Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Transactions on Cognitive and Developmental Systems* 13, 3 (2020), 717–728.

[62] Stephanie Rosenthal, Peerat Vichivanives, and Elizabeth Carter. 2022. The impact of route descriptions on human expectations for robot navigation. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 4 (2022), 1–19.

[63] Philipp Schmidt, Felix Biessmann, and Timm Teubner. 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29, 4 (2020), 260–278.

[64] Tobias Schneider, Sabiha Ghellal, Steve Love, and Ansgar RS Gerlicher. 2021. Increasing the user experience in autonomous driving through different feedback modalities. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 7–10.

[65] Loïck Simon, Clément Guérin, Philippe Rauffet, Christine Chauvin, and Éric Martin. 2023. How humans comply with a (potentially) faulty robot: Effects of multidimensional transparency. *IEEE Transactions on Human-Machine Systems* 53, 4 (2023), 751–760.

[66] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. 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*. 1–13.

[67] Micol Spitale, Minja Axelsson, Neval Kara, and Hatice Gunes. 2023. Longitudinal evolution of coachees' behavioural responses to interaction ruptures in robotic positive psychology coaching. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 315–322.

[68] Maia Stiber, Russell H Taylor, and Chien-Ming Huang. 2023. On using social signals to enable flexible error-aware hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 222–230.

[69] Ashley Suh, Isabelle Hurley, Nora Smith, and Ho Chit Siu. 2025. Fewer than 1% of explainable ai papers validate explainability with humans. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[70] Pradyumna Tambwekar and Matthew Gombolay. 2023. Towards reconciling usability and usefulness of explainable ai methodologies. *arXiv preprint arXiv:2301.05347* (2023).

[71] Lennart Wachowiak, Andrew Fenn, Haris Kamran, Andrew Coles, Oya Celiktutan, and Gerard Canal. 2024. When Do People Want an Explanation from a Robot?. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 752–761.

[72] Lennart Wachowiak, Peter Tisnikar, Gerard Canal, Andrew Coles, Matteo Leonetti, and Oya Celiktutan. 2024. Predicting When and What to Explain From Multimodal Eye Tracking and Task Signals. *IEEE Transactions on Affective Computing* (2024).

[73] Lennart Wachowiak, Peter Tisnikar, Andrew Coles, Gerard Canal, and Oya Celiktutan. 2024. A Time Series Classification Pipeline for Detecting Interaction Ruptures in HRI Based on User Reactions. In *Proceedings of the 26th International Conference on Multimodal Interaction*. 657–665.

[74] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[75] Xinru Wang and Ming Yin. 2023. 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*. 1–19.

[76] Yao Xie, Ge Gao, and Xiang'Anthony' Chen. 2019. Outlining the design space of explainable intelligent systems for medical diagnosis. *arXiv preprint arXiv:1902.06019* (2019).

[77] Elmira Yadollahi, Fethiye Irmak Dogan, Marta Romeo, Dimosthenis Kontogiorgos, Peizhu Qian, and Yan Zhang. 2025. 3rd Workshop on Explainability in Human-Robot Collaboration: Real-World Concerns. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1994–1996.

[78] Wencan Zhang and Brian Y Lim. 2022. Towards relatable explainable AI with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[79] Robert Zimmermann, Daniel Mora, Douglas Cirqueira, Markus Helfert, Marija Bezbradica, Dirk Werth, Wolfgang Jonas Weitzl, René Riedl, and Andreas Auinger. 2023. Enhancing brick-and-mortar store shopping experience with an augmented reality shopping assistant application using personalized recommendations and explainable artificial intelligence. *Journal of Research in Interactive Marketing* 17, 2 (2023), 273–298.

A Appendix: Model Features by Modality

Modality	Features	Type
Visual Embeddings	x2048 (ResNet visual embedding)	continuous
Pose / Kinematics	All 3D joint coordinates (Elbow, Wrist, Shoulder, Spine, Head, Nose, etc.; X/Y/Z)	continuous
	Derived kinematics (speed, acceleration, jerk, distances, joint angles, torso lean/twist, head orientation, alignment measures)	continuous
	Asymmetry and difference metrics	continuous
	Hand state indicators (left/right_hand_raised, pointing)	binary
	Hand event indicators (raise_event, point_event)	binary
Engineered NLP	Disfluency counts (tokens, fillers, repetitions, corrections; raw + per 100 words)	continuous
	Word count, speech rate (wps, wpm)	continuous
	POS tag counts (raw + per 100 words)	continuous
	Sentiment scores (neg, neu, pos, compound)	continuous
	Robot speaking indicator	binary
Interaction Context	Time since user speech	continuous
	Time since robot speaking	continuous
	Time since explanation	continuous
	Time since request	continuous
Gaze	robot_gaze, table_gaze, robottable_gaze, basket_gaze, screen_gaze	continuous
Speech Acoustics	Energy features (frequencydomainenergy, logenergy, lowfrequencyenergy, Loudness_sma3)	continuous
	Spectral features (spectralentropy, spectralFlux_sma3, alphaRatio_sma3, hammarbergIndex_sma3, slope features)	continuous
	MFCCs (mfcc1–4_sma3)	continuous
	F0 and voicing features	continuous
	Formant frequencies (F1–F3 freq/bandwidth/amplitude)	continuous
	Voice quality (jitter, shimmer, HNR)	continuous
Facial Action Units	AU01–AU45 (intensity)	continuous
	AU01–AU45 (occurrence)	binary
Text Embeddings (SBERT)	x384 (SBERT embedding)	continuous

Table 6: Features used in the models, grouped by modality. Feature type indicates whether the variable is continuous or binary.