Are You Sure? - Multi-Modal Human Decision Uncertainty Detection in Human-Robot Interaction

Lisa Scherf*
lisa_katharina.scherf@tu-darmstadt.de
Technische Universität Darmstadt
Darmstadt, Germany

Lisa Alina Gasche
Technische Universität Darmstadt
Darmstadt, Germany
lisaalina.gasche@stud.tu-darmstadt.de

Eya Chemangui
Technische Universität Darmstadt
Darmstadt, Germany
eya.chemangui@stud.tu-darmstadt.de

Dorothea Koert
Technische Universität Darmstadt
Darmstadt, Germany
dorothea.koert@tu-darmstadt.de

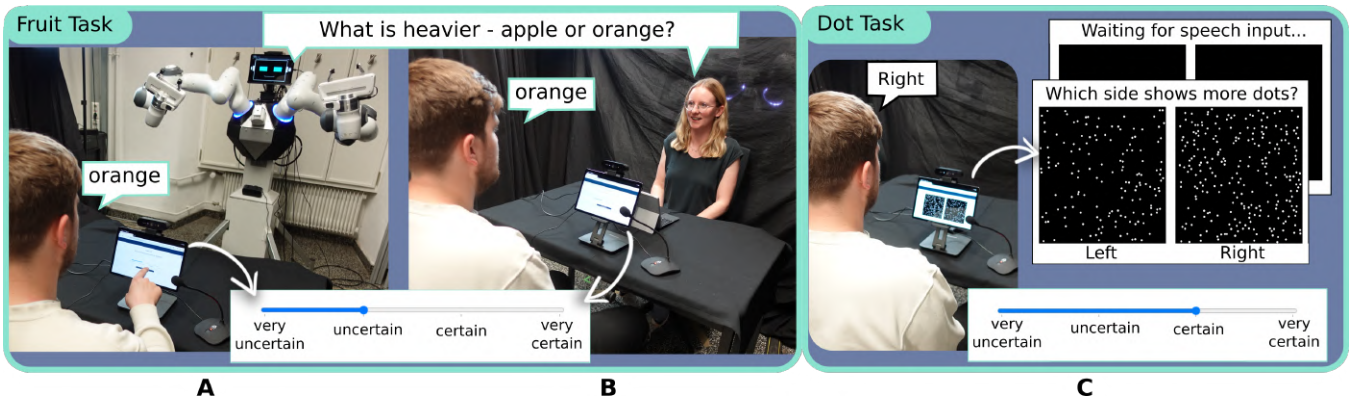

Figure 1: For our uncertainty detection model we collected video and audio data of 27 participants performing two decision tasks, i.e. a Fruit Task and Dot Task. In the Fruit Task either a human (A) or robot (B) asks the participant which of two fruits is heavier. In the Dot Task (C) the participant has to decide which of two images shown for one second contains more white dots.

ABSTRACT

In a question-and-answer setting, the respondent is often not only communicating the requested information but also indicating their confidence in the answer through various behavioral cues. Humans excel at interpreting these cues and monitoring the uncertainty of other persons. Being able to detect human uncertainty in human-robot interactions in a similar way can enable future robotic systems to better recognize uncertain and error-prone human input. Additionally, automatic human uncertainty detection can enhance the responsiveness of robots to the user in moments of uncertainty by providing help or clarification. While there is some work on uncertainty detection based on a single modality, only a few works focus on multi-modal uncertainty detection. Even fewer works explore how human uncertainty manifests through behavioral cues in human-robot interactions. In this work, we analyze occurrences of behavioral cues related to self-reported uncertainty on experimental data from 27 participants across two decision-making tasks. Additionally, in the first task, we varied if participants interacted with a human or a robot. On the recorded data, we extract features accessible via a webcam and a microphone and train a multi-modal classifier. Experimental evaluation of our developed classifier shows that it significantly outperforms third-person annotators in accuracy and F1 score. Humans report feeling less observed when responding to a robot compared to a human. Nevertheless, we found that the behavioral differences did not significantly affect the performance of our proposed uncertainty classification.

CCS CONCEPTS

• **Human-centered computing** → **User models**.

KEYWORDS

user study; human uncertainty detection; human-robot interaction

ACM Reference Format:
Lisa Scherf, Lisa Alina Gasche, Eya Chemangui, and Dorothea Koert. 2024. Are You Sure? - Multi-Modal Human Decision Uncertainty Detection in Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3610977.3634926

HRI '24, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0322-5/24/03.
https://doi.org/10.1145/3610977.3634926

1 INTRODUCTION

In a conversational setting, the main goal of asking questions is the exchange of information. However, the respondent is often not only communicating the requested information but also indicates the confidence in their answer [40]. Humans excel at monitoring another person's uncertainty conveyed through various behavioral cues, including visual cues (facial expressions, gaze, gestures), as well as auditory cues (intonation, fillers, pauses) [9, 22, 43]. Once robots enter more real-world settings, they will inevitably face situations where they require not only the ability to process human input on a factual level but additionally need to monitor their interaction partners' uncertainty about provided answers [24]. Specifically, the estimated uncertainty can serve as an indication for the correctness of human input, which can increase reliability in interactive robotic systems and enable learning from sub-optimal human input [18, 34, 35, 46]. Furthermore, the detection of human uncertainty can improve the responsiveness of assistive systems such as in student-tutor-frameworks by providing help or clarification to human users in moments of uncertainty [13, 29].

Studies indicate that humans transfer their behavior in human-human interactions to human-machine interactions, suggesting that humans also communicate their uncertainty in human-robot interactions [23]. However, some studies indicate differences in social reactions based on the presence and appearance of an embodied agent [20]. Overall, we found a lack of studies that compare how uncertainty manifests in behavioral cues in human-human vs. human-robot interactions. This raises the two main research questions of our work, i.e. how human uncertainty reflects in multi-modal behavioral cues in a question-answer setting with a robotic interaction partner (**RQ1**) and if a robot can learn to detect answer-related human uncertainty at a human level of accuracy (**RQ2**).

While there is some existing work on uncertainty detection in a non-robotic setting based on a single modality, such as acoustic cues and lexical features [13, 26, 28–30, 37, 47], facial expressions [5, 41], eye tracking data [10, 46], or brain activity [25, 38], only few works focus on multi-modal uncertainty detection [14]. Unlike related approaches that detect uncertainty in human-robot interaction [11], we focus on human decision uncertainty corresponding to a specific decision between options rather than uncertainty in a conversational setting. In particular, monitoring human decision uncertainty may help future robotic systems to asses an interaction partner's knowledge state [9] and increase reliability in processing and evaluating human input.

Our main contributions here are threefold. First, we introduce a Bayesian fusion-based method for multi-modal detection of human decision uncertainty that significantly outperforms human annotators regarding accuracy and F1 Score. Specifically, the proposed classifier works solely on non-invasive features accessible over a webcam and a microphone. Second, we find that even though humans feel significantly less observed when interacting with a robot compared to a human, they overall show similar behavioral cues related to uncertainty. Third, we provide the research community a novel multi-modal dataset for human decision uncertainty detection[1] including self-reported uncertainty labels as well as third-person annotations.

[1]The recorded dataset is available at https://osf.io/48ksh/

2 RELATED WORK

While there is a large body of literature for methods to enable robots to recognize basic human emotions such as anger, happiness, sadness, or fear [1, 2, 27, 31], there are fewer works that explore human uncertainty recognition [11, 13, 15, 30]. Human uncertainty can occur in different forms [6], and when growing up, humans develop impressive abilities to detect behavioral cues for uncertainty in other humans [9, 22, 43]. In decision tasks, human uncertainty is often related to a high task difficulty, as well as inversely related to answer correctness [18, 34, 46]. Being able to monitor uncertainty is therefore helpful to assess the knowledge state of others and evaluate the corresponding response [9]. In interactive systems, automatic human uncertainty detection has the potential to improve reliability by being able to evaluate human input based on the corresponding human uncertainty [35].

In this paper, we focus on detecting decision uncertainty, which occurs when a person has to decide between multiple options. This includes question-and-answer settings where the respondent is often not only communicating the requested information but also indicates the corresponding confidence potentially as a form of self-presentation to save face in case of an incorrect response [40]. This reflects the ability of humans to evaluate one's own confidence internally. The ability to infer another person's confidence is important in decision-making involving other individuals [42]. In contrast to decision uncertainty, Cumbal et al. [11], for example, detect listener uncertainty in human-robot dyadic conversation based on facial expressions, gaze, head movements, and speech features. Specifically, they focus on detecting uncertainty caused by a failed understanding of spoken information from a conversational partner rather than a decision-making scenario with specific options.

A multi-modal approach that combines the information from multiple modalities can strengthen potential short-comings of each modality and can improve predictive power [2]. While there are only few works on multi-modal uncertainty detection [11, 14], there are several findings in the literature linking different modalities to human uncertainty [5, 41], as well as some approaches to human uncertainty detection based on a single modality [15, 17, 28, 37].

In dialogue systems or student-tutor frameworks, acoustic and linguistic features are often used to detect uncertainty [13, 26, 28–30, 37, 47]. Nevertheless, uncertainty is not only communicated through speech but also reflected in facial expressions, as the findings of Bitti et al. [5] and Stone and Oh [41] suggest. Furthermore, response time as the time taken to form a decision can serve as an indicator for uncertainty or the inversely related confidence in decision-tasks [8, 19]. Kontogiorgos et al. [21] use a combination of gaze and pointing modalities in order to detect listener uncertainty in human-human interactions. There seems to be a connection between uncertainty or the related concept of confusion and eye-tracking data such as gaze direction, pupil size, fixations, and saccades [10, 32, 39, 46]. There is some work on uncertainty detection or detection of the related Feeling-of-Knowing (FOK) based on multi-modal behavioral data. Swerts and Krahmer [23] reveal that low FOK answers tend to have a higher number of auditory and visual cues, such as funny faces, eyebrow movements, or high intonation. In addition, they show that human observers can distinguish between high and low FOK responses. However, they

do not learn a model to predict the FOK. Greis et al. [14] analyze the relation between response time, eye tracking data, and heart rate connected to uncertainty in a quiz task. While EEG signals, as well as the heart rate, are also related to uncertainty [14, 25, 38], in this work, we are focusing on human uncertainty detection on non-invasive behavioral signals that can be easily accessed in a human-robot interaction scenario using a camera or microphone.

There are different ways of combining multiple modalities. Multimodal models can either be trained using early fusion to combine multiple modalities on a feature-level [45], or by combining unimodal decision values on a decision level (late fusion) [1, 31, 36]. The Bayesian method Independent Opinion Pool (IOP) [4] is a probabilistic optimal fusion method according to Bayes rule and has already shown benefits for decision fusion in human intention recognition [44]. By combining multiple potentially inaccurate classifiers using IOP, the final decision uncertainty can be reduced.

3 HUMAN UNCERTAINTY DETECTION

We propose an approach to detect human decision uncertainty from multi-modal behavioral cues in human-robot interaction. In this section, we describe the experiment procedure and data collection (Section 3.1), the feature extraction process (Section 3.2), and how we trained our proposed multi-modal classifier for human decision uncertainty detection on the recorded data set (Section 3.3). An overview of our approach is illustrated in Figure 2.

3.1 Data Collection

In an experiment with 27 participants, we collected multi-modal behavioral data corresponding to human decision uncertainty. Within the experiment, the participants faced two different decision-making tasks, where they had to decide between two choices. During the first task (Fruit Task), we varied if a human or a robot interaction partner posed the questions. In the second task (Dot Task), the subjects interacted solely with a tablet. This results in three experiment conditions *fruits_human*, *fruits_robot*, and *dots*. The experiment was conducted in German, the participants' native language.

Experiment Setup. Figure 1 shows the three experiment conditions. In the *Fruit Task* the participants had to decide which of two fruits is heavier based on their prior knowledge. The questions were posed in the form "What is heavier - X or Y" and after each question, the participants answered with voice input, naming one of the fruits X or Y. In *fruits_human* (Figure 1 B), they were facing a human investigator. The investigator did not react to the participants' responses and kept a neutral facial expression. In *fruits_robot* (Figure 1 A), a robot instead of a human asked the questions. The robot consists of two Franka Emika Panda arms and a tablet displaying an animated face as a head, allowing the robot to move its mouth while talking. While posing the question, the robot moved first one arm and then the other arm up and down emphasizing two options. In the *dots* condition (Figure 1 C), the participants interacted solely with a tablet. They were tasked to select which of two images displayed for one second contained a higher amount of white dots by voice input, saying "left" or "right". Variations of the Dot Task have already been used in literature [33] as a decision-making task with perceptual uncertainty compared to the Fruit Task where participants have to query their internal knowledge.

Experiment Procedure. First, the participants provided informed consent. At the beginning of each experiment condition, the corresponding task was explained in form of written instructions. The tasks were framed as a quiz. As an incentive, the participants were promised a prize for achieving a new high score in number of correct answers. We randomized the order of the Fruit Task and Dot Task, as well as the order of the two conditions *fruits_human* and *fruits_robot* within the Fruit Task. In addition, two sets of questions for the Fruit Task with different pairs of fruits were randomly assigned to the two Fruit Task conditions *fruits_human* and *fruits_robot*. In all three experiment runs, the participants could familiarize themselves with the task setting in two trial runs. Then, the participants had to choose 30 times between pairs of fruits or images, respectively. After selecting one option, a slider was shown on the tablet in front of them where they reported their certainty level regarding the choice on a 4-point Likert scale (very uncertain, uncertain, certain, very certain). For feature analysis, classifier training, and evaluation, we summarize the self-reported categories "very uncertain" and "uncertain" into *uncertain* and the self-reported "certain" and "very certain" into *certain*.

Data Recording. We collected data from 27 participants (18 female, 9 male), aged between 18 and 35. The recruitment process was through university online platforms and word of mouth. The experiments were approved by the ethics committee of TU Darmstadt on November 28, 2022 (EK 80/2022). During the experiment, a Logitech Brio Stream Webcam recorded the participants' faces with 30fps and 1280x720 resolution. In addition, a KLIM microphone on the table in front of the participants recorded audio files. We synchronously started the data recording using ROS (Robot Operating System) and saved ROS timestamps for all recordings. We manually labeled the end of each posed question and the beginning of the corresponding response of the participant. For the *Fruit Task*, the end of the question marks the point in time when the robot or investigator fully voiced the question and the participant has to name the heavier fruit. In the dot task, the end of the question marks the point in time when the two pictures disappear and the participant has to choose the picture with more dots. Even though the participants were instructed to not ask questions during the experiment, some participants asked clarifying questions, e.g. whether tomatoes or cherry tomatoes were meant. We excluded the corresponding six responses. In addition, the data recording failed for one *dots* and two *fruits_human* conditions due to technical problems.

Third-person annotations. We asked ten persons (6 male, 4 female) to manually annotate all responses of all participants, resulting in ten third-person annotations per response. First, we provided context about the data recording by showing the annotators the experiment instructions for all three experiment conditions. Then, the annotators replayed the recorded audio and video for each response from the end of the posed question until one second after the start of the participant's response. This duration was chosen since an inspection of the data revealed facial expressions corresponding to uncertainty even shortly after the response. The same time window is also used for the model training as described in Section 3.3. Note here that the annotators only observe the participant's response without knowing the posed question. This prevents biased annotations based on the question's difficulty. The annotators were

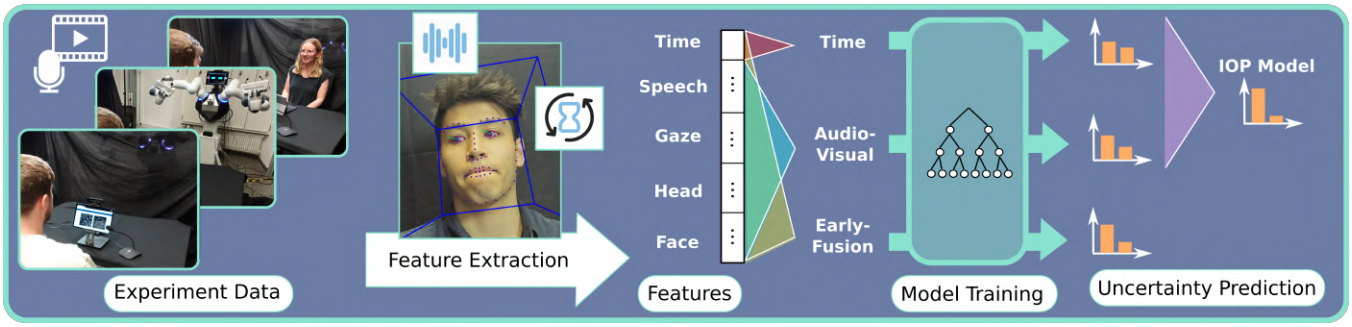

Figure 2: Overview of the model training pipeline. Features are extracted from all experiment recordings and used to train models on each modality individually, all modalities combined (early-fusion), and all modalities except response time (audio-visual). In addition, Independent Opinion Pool is used to combine the resulting response time model and audio-visual model.

then asked to decide whether the participant seemed uncertain or certain. They entered their uncertainty annotations via key presses.

3.2 Feature Extraction

To analyze behavioral cues related to human decision uncertainty and train a classifier on the collected data, we extracted several features for each of the participants' responses. Here, we consider the time window from the end of the posed question until one second after the participant's response.

Response Time. We calculate the response time as the difference between the end of the posed question and the participant's response. This feature corresponds to the time the participant takes to think about the question and respond.

Facial Behavior. We use OpenFace [3] to extract facial action units, head pose, and gaze direction. The system detects the intensity between zero and five of 18 action units corresponding to individual components of facial muscle movements. More information on the Facial Action Coding System can be found in [12]. We calculate the minimum, maximum, mean, standard deviation, and range based on the intensity of those action units for each frame in the response window.

Gaze. OpenFace estimates the 3D eye gaze direction for both eyes. We calculate the position and orientation change in x, y, and z direction between two frames for both eyes in the response window. We then take the minimum, maximum, mean, sum, and standard deviation of the position and orientation changes as features. In addition, we calculate the gaze velocity as degrees per second and take the minimum, maximum, and mean over the response window.

Head Orientation. Similar to calculating the gaze features, we calculate the changes in x, y, and z rotation of the head pose estimation detected by OpenFace and take the minimum, maximum, mean, sum, and standard deviation as features. For the head pose position, we calculate the change between two frames using the Euclidean distance and again calculate the minimum, maximum, mean, sum, and standard deviation.

Speech. All speech features are extracted based on the recorded audio data, using the Parselmouth library [7, 16]. Considering the described time window, we calculate the minimum, maximum, mean, and standard deviation of the pitch, intensity, and Harmonics-to-Noice Ration (HNR), respectively. We also calculate the upper and lower percentile for the intensity and pitch.

3.3 Multi-modal Uncertainty Classification

Multi-modal Classifiers. Let $\mathcal{X} \in \mathbb{R}^{N \times D}$ denote the input data, where N is the number of responses for all experiment conditions and D the number of extracted features. We want to learn a classifier that maps this input data to a probability for human uncertainty u

$$C : \mathcal{X} \mapsto p(u|\mathcal{X}). \tag{1}$$

Since, in particular, the self-reported labels "very uncertain" and "very certain" appeared less often (Figure 3), we randomly upsample less frequent labels such that the training data set is balanced for each participant. For model training, we select all features described in Section 3.2 that show a highly significant difference between uncertain and certain responses according to a Wilcoxon signed-rank test with significance level $\alpha = .001$. This feature selection based on statistical testing is interpretable and showed better results in pre-tests compared to other feature selection methods such as PCA or feature importances. We normalize all features using the minimum and maximum feature value of all responses of one participant and experiment condition

$$\hat{x}^{s,c}_{n,d} = \frac{x^{s,c}_{n,d} - \max\limits_{0 \le n \le N^c} x^{s,c}_{n,d}}{\max\limits_{0 \le n \le N^c} x^{s,c}_{n,d} - \min\limits_{0 \le n \le N^c,} x^{s,c}_{n,d}}, \tag{2}$$

where $x^{s,c}_{n,d}$, denotes the feature d corresponding to response n of participant s for experiment condition c and N^c is the number of responses for experiment condition c. Then, we standardize all features by subtracting the mean and scaling them to unit variance. We evaluate three different classifiers: Support Vector Machine (SVM), Random Forest (RF), and Multilayer Perceptron (MLP). All classifiers are implemented using the sklearn Python library. We train and evaluate the models using leave-one-out cross-validation by training the model on the data of all but one participant and evaluating it on the remaining one participant. We report the average macro F1-score, accuracy, precision, and recall over these validation splits. All model hyper-parameters are tuned first using a broad random search over the parameter space and afterwards an

exhaustive grid search using coarse-to-fine tuning. For the SVM, we vary the kernel $k \in$ {rbf, poly, sigmoid}, C, and γ parameter. For the RF model, we vary the number of estimators, maximum depth, and maximum number of features. Lastly, for the MLP we choose the best values for the number of hidden layers, maximum number of iterations, activation function $\Phi \in$ {tanh, relu}, solver $\in$ {stochastic gradient descent (sgd), adam} and learning rate.

Feature Fusion. We compare early and late fusion to combine features of different modalities. For early fusion, we combine features of different modalities to one feature input vector X and train the classifier as described above. For late fusion, we train separate probabilistic classifiers for each modality individually or a subset of all modalities and then combine the resulting M categorical probability distributions $p(u|X_1), .., p(u|X_M)$ in a Bayesian optimal way using Independent Opinion Pool (IOP) [4, 44]

$$p(u|X_1, .., X_M) \propto \prod_{n=1}^{M} p(u|X_n). \quad (3)$$

We test combinations of different classifiers trained on each modality, as well as subsets of all modalities by combining the resulting probability distributions using IOP. Out of these combinations, we report the results of the best-performing IOP model.

4 DATA ANALYSIS AND CLASSIFIER EVALUATION

On our recorded data set, we first analyze behavioral feature occurrences in relation to self-reported decision uncertainty (Section 4.1). Subsequently, we compare different classifier models trained on identified relevant features with human annotator accuracy (Section 4.2) and investigate differences between human-human and human-robot interactions (Section 4.3).

4.1 Feature Analysis

Our data set consists of video and audio recordings, third-person human annotations, and self-reported uncertainty labels of 27 participants with in total 780 responses for the *dots* condition, 745 for the *fruits_human*, and 809 for the *fruits_robot* condition. The distribution of the self-reported uncertainty values for each task is shown in Figure 3.

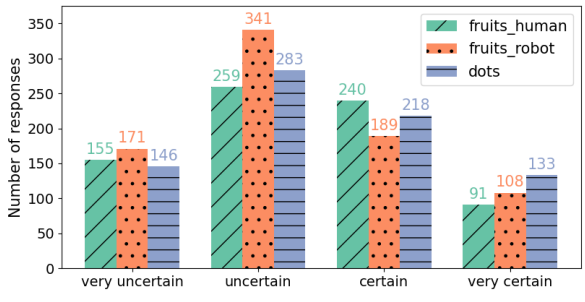

Figure 3: Self-reported uncertainty over experiment conditions. For model training we combine 'very uncertain' / 'uncertain' (*uncertain*) and 'very certain' / 'certain' (*certain*).

Table 1: Features with a highly significant difference ($\alpha = .001$) between uncertain and certain questions.

	Face	Gaze	Head	Speech	Time	Total
dots	9/90	0/43	1/20	2/19	1/1	13/173
fruits_human	21/90	7/43	6/20	1/19	1/1	36/173
fruits_robot	10/90	1/43	3/20	6/19	1/1	21/173
all tasks	27/90	8/43	6/20	3/19	1/1	45/173

We extracted 173 features in total, as described in Section 3.2. A Wilcoxon signed-rank test, comparing the participants' average feature values for uncertain and certain responses, shows a statistically highly significant difference in 45 features (significance level $\alpha = .001$). Table 1 shows the share of these 45 features for the different modalities. In addition, we analyzed feature differences between uncertain and certain for the data of each task individually. For the *fruits_human* data, the Wilcoxon signed-rank test finds a difference between uncertain and certain for 36 features compared to 21 for the *fruits_human* and 13 for the *dots* experiment condition. For the *fruits_robot* data, speech seems to play an important role compared to the other tasks. In contrast, for the *fruits_human* data, a higher number of facial behavior, head, and gaze features show a significant difference between uncertain and certain responses.

The response time shows a significant difference for all tasks individually, as well as for the combined data (all $p < .001$). The unnormalized response time in seconds for all participants and uncertain vs. certain responses, as well as for the different tasks, are visualized in Figure 5 (A). Here, the average response time over all experiment conditions is higher for uncertain responses (Mean=2.98, Mdn=2.30) than certain responses (Mean=1.69, Mdn=1.50).

When looking at the facial behavior in detail, at least three features computed based on the intensity of action units AU07, AU09, AU10, and AU17 show a highly significant difference (all $p < .001$) for uncertain and certain responses. These action units are described as Lid Tightener (AU07), Nose Wrinkler (AU09), Upper Lip Raiser (AU10), and Chin Raiser (AU17) [12]. Examples of facial expressions for uncertain responses with high intensity for some of these action units (> 2.0) are shown in Figure 4. Figure 5 (B) visualizes the unnormalized mean AU10 intensity for all participants and each task and certain vs. uncertain responses. The mean AU10 intensity is significantly higher (Wilxocon, $\alpha = .01$) for uncertain responses compared to certain responses for all tasks individually (*fruits_human*: $p = .005$, *fruits_robot*: $p = .002$, *dots*: $p = .009$).

For the speech features, three out of seven features based on the speech intensity (mean, standard deviation, upper percentile: all $p < .001$) show a highly significant difference over all experiment conditions (Wilcoxon, $\alpha = .001$). The unnormalized mean intensity over all experiment conditions is slightly lower for certain (Mean=26.28, Mdn=25.98) compared to uncertain responses (Mean=27.26, Mdn=26.81), suggesting that participants talked louder when being certain of the answer. However, when looking at the mean intensity feature for all three experiment conditions individually, there is a significant difference for *fruits_robot* and *dots* ($p < .001$) but no significant difference for *fruits_human* ($p = .022$) between certain and uncertain responses.

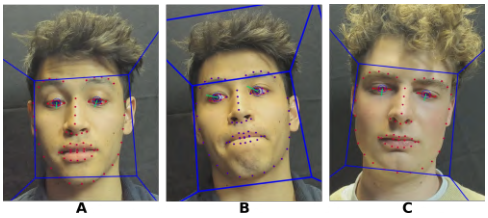

Figure 4: Example facial expression for uncertain responses with detected facial landmarks, gaze direction, and head pose by OpenFace. Here, participant ERMF18 shows a high AU02, AU17, AU26 intensity in (A) and a high AU02 intensity in (B). Participant XEAF02 shows a high intensity for AU07 (C).

We observed that some participants leaned down to the microphone in the *fruits_robot* and *dots* conditions. They might suspected a speech recognition system and therefore tried to articulate their response loud and clear, leading to differences in speech features compared to *fruits_human*.

For the head movement features, there is a significant difference (Wilcoxon, $\alpha = .001$) for the mean rotation change in x direction or pitch between uncertain and certain responses for the *fruits_human* condition ($p < .001$), as well as for the combined data ($p < .001$). For *fruits_robot* ($p = .006$) and *dots* ($p = .002$) there is no significant difference. In addition, the minimum position change shows a significant difference for all data combined, as well as all experiment conditions individually with $p < .001$. For the mean rotation change in x direction, the participants tend to show a higher change for certain responses (Mean=0.40, Mdn=0.26) compared to uncertain responses (Mean=0.32, Mdn=0.27), which might reflect a nodding behavior. The minimum position change shows lower values for uncertain responses (Mean= 0.28, Mdn=0.23) than certain responses (Mean=0.38, Mdn=0.33), so the participants seemed to have moved less if they were uncertain.

For the gaze features, the normalized minimum position change shows a significant difference (Wilcoxon, $\alpha = .001$) between uncertain and certain responses for left ($p < .001$, uncertain: Mean=0.30, Mdn=0.23, certain: Mean=0.37, Mdn=0.32) and right eye ($p < .001$, uncertain: Mean=0.29, Mdn=0.24, certain: Mean=0.37, Mdn=0.31).

4.2 Multi-modal Uncertainty Detection

We compare different models trained on identified relevant features with the third-person annotations (**RQ1**). While self-reported uncertainty and perceived uncertainty are not to be equated, we consider this a valuable baseline that was also used before [30]. The human annotators (Section 3.1) achieve an average accuracy of 0.695 and an F1 score of 0.658. Here, the lowest accuracy and F1 score is achieved for the *dots* condition with 0.666 and 0.610, respectively, compared to *fruits_robot* (Acc=0.723, F1=0.678) and *fruits_human* (Acc=0.709, F1=0.673). There was a moderate agreement between the annotators with an average kappa inter-annotator agreement of 0.546 and standard deviation of 0.158.

For early feature fusion a RF model with a maximum depth of 4, 48 maximum features, and 850 estimators achieves the best performance (Acc=0.722, F1=0.711, precision=0.662, recall=0.728) compared to SVM (Acc=0.716, F1=0.694, precision=0.699, recall=0.654.) and MLP (Acc=0.707, F1=0.703, precision=0.664, recall=0.662).

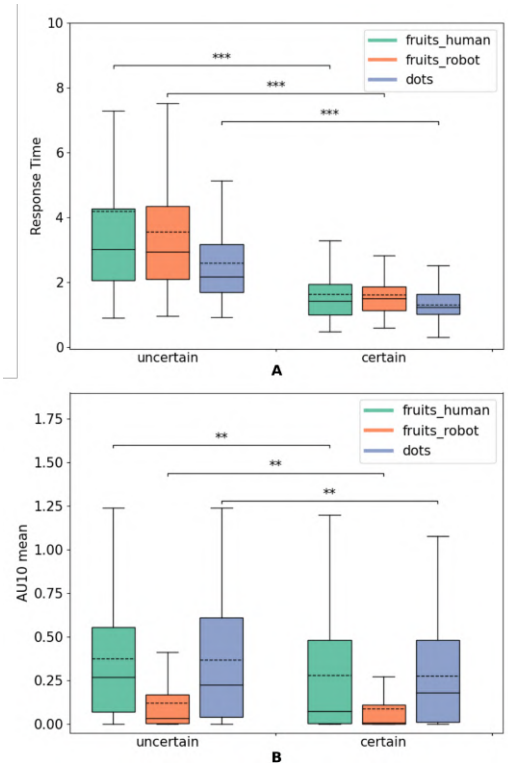

Figure 5: Average response time in sec. (A) and AU10 intensity (B) for each participant shown for each task and uncertain vs. certain responses according to self-reported labels. * marks significant difference with $\alpha = .001$, ** marks significant difference with $\alpha = .01$, Median is solid, mean is dashed line.**

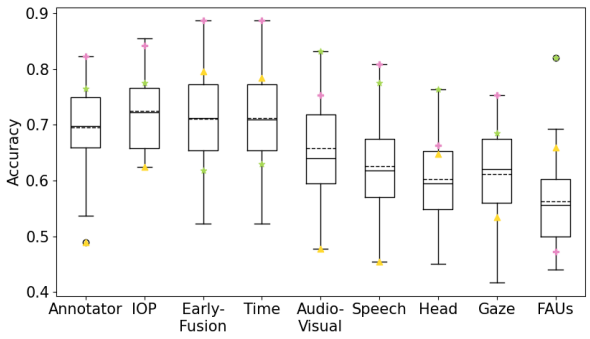

Figure 6: Accuracy of all RF models, human annotators, and IOP model for each participant. Highlighted: ZAMM05 (yellow triangle), SHTB31 (pink cross), XEAF02 (green star), RHMZ14 (orange star). Median is solid and mean dashed line.

Figure 6 visualizes the accuracies for each participant for the human annotations and the best RF model trained on each modality separately (response time, speech, head, gaze, facial expressions), as well as trained on all modalities combined (early-fusion). For late fusion we compared IOP combinations of different classifiers trained on each modality individually, as well as IOP combinations

of subsets of all modalities. Out of these combinations, the IOP model that fuses the response time model with the model trained on all remaining modalities (audio-visual) performed best. The results of this model are also visualized in Figure 6. Table 2 reports the average accuracy, balanced accuracy, macro F1 score, precision, and recall over all participant cross-validation splits for all RF models and the IOP model. The IOP model (Acc=0.725, F1=0.726) outperforms human annotations (Acc=0.696, F1=0.662). A Wilcoxon signed-rank test over all participants shows a significant difference ($\alpha = .01$) for F1 score ($p < .001$) and accuracy ($p = .005$).

The early-fusion model trained on all modalities achieves a similar balanced accuracy of 0.725 compared to the IOP model but slightly lower values for Acc=0.711 and F1=0.702. There is no significant difference in accuracy ($p = .196$) and F1 score ($p = .348$) between the two models (Wilcoxon, $\alpha = .001$). The early-fusion model does not significantly outperform human annotators in F1 score ($p = .026$) and accuracy ($p = .645$).

The RF model trained on only the response times (Acc=0.713, F1=0.704) performs slightly worse than the IOP model. However, a Wilcoxon signed-rank test ($\alpha = .01$) does not show a statistically significant difference for both accuracy ($p = .241$) and F1 score ($p = .441$). The performance comparison of the response time model to the human annotations reveals no significant difference in accuracy ($p = .645$) or F1 score ($p = .019$).

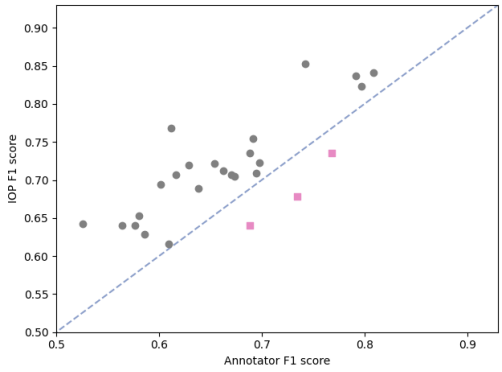

Figure 7: High correlation between IOP and annotator F1 score. For most participants, IOP is better (grey dots). For 3 participants the annotator F1 score is higher (pink squares).

In general, we see person-dependent variations in model performance. To illustrate this person-dependence, in Figure 6, the performance of some participants is highlighted in color across all models. For participant XEAF02, the models trained on only facial expressions (Acc=0.820), head movements (Acc=0.764), or speech data (Acc=0.775) perform well, resulting in an even higher performance for the audio-visual model (Acc=0.831). The response time model (Acc=0.629), however, performs poorly compared to the audio-visual model. In contrast, for participant SHTB31, response time is an important indicator of uncertainty. Here, the response time model achieves a high accuracy of 0.888. In addition, the model trained on only speech data (Acc=0.809) performs well, whereas the model using facial expressions as input performs poorly (Acc=0.472). For participant ZAMM05, both human annotators (Acc=0.489) and the audio-visual model (Acc=0.477) perform poorly with below-chance accuracies. However, the response time model performs well with an accuracy of 0.784. Both accuracy and F1 score of the best-performing IOP model shows a strong positive Pearson correlation r between model and annotator performance over all participants (Acc: $r = .795$, $p < .001$, F1: $r = .784$, $p < .001$). Figure 7 shows the annotators' F1 score vs. the IOP model.

4.3 Behavioral Differences between Conditions

We analyze differences in behavioral cues related to uncertainty for human-human vs. human-robot interactions (**RQ2**). We compare the average feature values for each participant between *fruits_human* and *fruits_robot* using a Wilcoxon signed-rank test with significance level $\alpha = .01$. Note here that we compare unnormalized features values and focus on features that showed a significant difference between uncertain and certain responses (Section 4.1). For the majority of these features, there is no significant difference between *fruits_human* and *fruits_robot*. This includes the response time, mean change in head pitch, and most features related to action units AU07, AU09, and AU17, which are linked to uncertainty. However, all features related to action units AU12 and AU10 (except minimum intensity), as well as the average AU07 intensity, and minimum head position change show a significant difference for these two experiment conditions (all $p < .001$). AU12 (Lip Corner Puller) shows a higher average intensity for *fruits_human* (Mean=0.53, Mdn=0.33) compared to *fruits_robot* (Mean=0.24, Mdn=0.05) which suggests that the participants smiled more when interacting with the human. Similarly, AU10 shows a higher average intensity for *fruits_human* (Mean=0.28, Mdn=0.10) compared to *fruits_robot* (Mean=0.11, Mdn=0.02) as shown in Figure 5 (B). In addition, four out of seven speech intensity features show a difference between *fruits_human* (Mean=26.87, Mdn=26.38) compared to *fruits_robot* (Mean=27.62, Mdn=27.23) (Wilcoxon, $\alpha = .01$). The participants might have talked louder to the robot to improve a suspected speech recognition or since the robot produced some background noise. Furthermore, the minimum gaze position change for both eyes is significantly higher ($p < .001$) for *fruits_human* (left/right eye: Mean=0.17/0.17, Mdn=0.15/0.15) compared to *fruits_robot* (left/right eye: Mean=0.14/0.13, Mdn=0.11/0.11). One participant specifically stated after the experiment that he tried to read the face of the opposite person, resulting in multiple in-between gazes at the experimenter. This behavior might have occurred less when interacting with the robot, leading to differences in gaze behavior. Similarly, the minimum head position change is significantly higher ($p < .001$) for *fruits_human* (Mean=0.93, Mdn=0.87) compared to *fruits_robot* (Mean=0.79, Mdn=0.71).

When testing the best-performing IOP model on only the data of the *fruits_robot* (Acc=0.719, F1=0.687) and *fruits_human* (Acc=0.687, F1=0.716) condition for each participant, we see no statistical difference in accuracy ($p = .493$) and F1 score ($p = .361$) (Wilcoxon, $\alpha = .01$) even though the model performs slightly better for *fruits_human*. The difference in performance is even higher for the response time only model with Acc=0.743, F1=0.732 for *fruits_human* and Acc=0.705, F1=0.684 for *fruits_robot*. However, a Wilcoxon signed-rank test with significance level $\alpha = .01$ shows no statistically significant difference in accuracy ($p = .197$) and F1 score ($p = .136$).

	Annotator	IOP	Early-Fusion	Time	Audio-Visual	Speech	Head	Gaze	Face
Acc	0.696	**0.725**	0.711	0.713	0.658	0.625	0.603	0.612	0.563
Balanced Acc	0.678	**0.726**	0.725	**0.726**	0.649	0.598	0.590	0.607	0.559
F1 score	0.662	**0.726**	0.702	0.704	0.630	0.570	0.580	0.592	0.539
Precision	0.709	**0.691**	0.635	0.637	0.642	0.574	0.537	0.553	0.514
Recall	0.514	0.684	**0.786**	**0.786**	0.546	0.429	0.494	0.522	0.486

Table 2: Performance of the annotators, IOP model, and all Random Forest models. We report the average accuracy, balanced accuracy, macro F1 score, precision, and recall over all participant cross-validation splits. The highest values are highlighted.

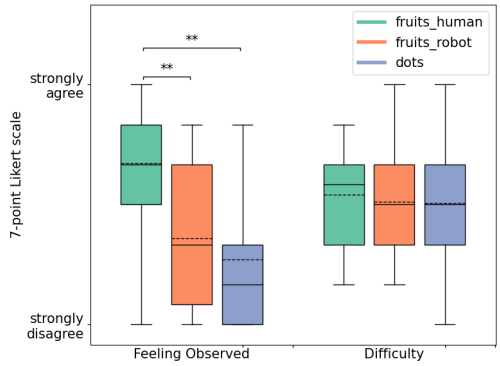

Figure 8: Questionnaire results for the two items: "I felt observed during the task" and "I had a hard time answering the questions" for each condition. ** marks significant difference with $\alpha = .01$. Median is the solid and mean the dashed line.

In a questionnaire, we asked the participants after each experiment condition if they felt observed and if they found it difficult to answer the questions on a 7-point Likert scale. The results are shown in Figure 8. There is no significant difference between the experiment conditions regarding how difficult it felt for the participants to answer the questions ($p = .308$) according to a Friedman test with a significance level of $\alpha = .01$. However, there is a significant difference in how observed they felt during each experiment run ($p < .001$). A Nemenyi-Friedman posthoc test reveals a significant difference between *fruits_human* and *fruits_robot* ($p = .001$), as well as between *fruits_human* and *dots* ($p = .001$). The participants felt more observed when interacting with a human (Mean=5.04, Mdn=5.0) compared to interacting with a robot (Mean=3.15, Mdn=3.0) or during the *dots* task (Mean=2.63, Mdn=2.0). Between the *dots* and *fruits_robot* condition, there was no statistically significant difference. One participant explicitly commented that she tended to show her uncertainty in order to avoid embarrassment. This is in line with Smith and Clark [40] who hypothesize that humans signal their uncertainty to maintain self-esteem.

4.4 Implications and Limitations

While we contribute a valuable dataset and a first multi-modal approach to detect human decision uncertainty in HRI, the size and diversity with respect to different tasks, persons, and environmental conditions is still limited and might influence model performance in different scenarios. Individual variations in how uncertainty manifests itself in behavioral cues are challenging and a person-dependent model calibration should be considered to increase robustness. Furthermore, bad lighting or environmental noise might lead to a decrease in model performance. Here, late-fusion methods with situation-dependent weighting of different modalities are an interesting line of future research. While human uncertainty is often related to answer correctness [18, 34, 46], the two are not to be equated and in some cases humans might not even be able to assess their own uncertainty correctly. Moreover, while we used a 4-point Likert scale in our experiments, the best way of letting humans rate their own uncertainty is still an open research question.

5 CONCLUSION AND FUTURE WORK

In this work, we proposed an experimental setup to collect behavioral data related to human decision uncertainty. The resulting dataset includes video and audio data of 27 participants facing two decision-making tasks in which they interacted with another human, a robot, or a tablet. From 2334 responses, we extracted multi-modal features, including response time, facial behavior, gaze, head movements, and speech features. The evaluation of classifiers trained on the extracted feature shows that a late Bayesian Fusion approach that combines a response time classifier with a classifier based on audio-visual features outperforms single modality classifiers and early feature fusion classifiers in terms of precision. The proposed classifier also significantly outperforms human annotators in terms of accuracy and F1 score. While there are some behavioral differences between human-robot and human-human interaction, and participants report feeling more observed when interacting with a human compared to a robot, most features show no significant difference, and the classifier performance is unaffected.

However, we saw variations of magnitude in behavioral features related to uncertainty across participants. One line of future work is, therefore, to investigate such differences further and develop methods for how a robot can learn to adapt its uncertainty detection and automatically re-calibrate across persons and tasks. Furthermore, Long Short-Term Memory (LSTMs) networks might be beneficial to exploit potential sequential patterns in the data. Lastly, we see human uncertainty detection as an important feature to integrate into interactive learning paradigms, such as interactive reinforcement learning, where it can enable the robot to weigh human feedback or advice based on its estimated certainty.

ACKNOWLEDGMENTS

This Work was funded by German Federal Ministry of Education and Research (project IKIDA 01IS20045). We thank all participants and annotators for their valuable contributions.

REFERENCES

[1] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends* 2, 02 (2021), 52–58.

[2] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. 2023. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications* 17 (2023), 200171.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[4] James O Berger. 1989. Statistical decision theory. In *Game Theory*. Springer, 217–224.

[5] Pio E Ricci Bitti, Luisa Bonfiglioli, Paolo Melani, Roberto Caterina, and Pierluigi Garotti. 2014. Expression and communication of doubt/uncertainty through facial expression. *Ricerche di Pedagogia e Didattica. Journal of Theories and Research in Education* 9, 1 (2014), 159–177.

[6] Amy R Bland and Alexandre Schaefer. 2012. Different varieties of uncertainty in human decision-making. *Frontiers in neuroscience* 6 (2012), 85.

[7] Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Vers. 6.1.38, retrieved 2 Jan. 2023 http://www.praat.org/.

[8] Claude Bonnet, Jordi Fauquet, and Santiago Ferrer. 2008. Reaction times as a measure of uncertainty. *Psicothema* 20 (03 2008), 43–8.

[9] S. E. Brennan and M. Williams. 1995. The Feeling of Another's Knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers. *Journal of Memory and Language* 34, 3 (1995), 383–398. https://doi.org/10.1006/jmla.1995.1017

[10] Tad T. Brunyé and Aaron L. Gardony. 2017. Eye tracking measures of uncertainty during perceptual decision making. *International Journal of Psychophysiology* 120 (10 2017), 60–68. https://doi.org/10.1016/J.IJPSYCHO.2017.07.008

[11] Ronald Cumbal, José Lopes, and Olov Engwall. 2020. Detection of listener uncertainty in robot-led second language conversation practice. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 625–629.

[12] Paul Ekman, W v Friesen, and J Hager. 2002. Facial action coding system: Research Nexus. *Network Research Information, Salt Lake City, UT* 1 (2002).

[13] Kate Forbes-Riley and Diane Litman. 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication* 53, 9-10 (2011), 1115–1136. https://doi.org/10.1016/j.specom.2011.02.006

[14] Miriam Greis, Jakob Karolus, Hendrik Schuff, Paweł W. Woźniak, and Niels Henze. 2017. Detecting uncertain input using physiological sensing and behavioral measurements. In *Proceedings of the 16th International Conf. on Mobile and Ubiquitous Multimedia*. ACM, NY, USA, 299–304. https://doi.org/10.1145/3152832.3152859

[15] Alexander E. Hramov, Nikita S. Frolov, Vladimir A. Maksimenko, Vladimir V. Makarov, Alexey A. Koronovskii, Juan Garcia-Prieto, Luis Fernando Antón-Toro, Fernando Maestú, and Alexander N. Pisarchik. 2018. Artificial neural network detects human uncertainty. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28, 3, 033607. https://doi.org/10.1063/1.5002892

[16] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71 (2018), 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

[17] Pavel Jahoda, Antonin Vobecky, Jan Cech, and Jiri Matas. 2018. Detecting decision ambiguity from facial images. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 499–503.

[18] Adam Kepecs and Zachary F. Mainen. 2012. A computational framework for the study of confidence in humans and animals. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 367, 1594 (2012), 1322–1337. https://doi.org/10.1098/rstb.2012.0037

[19] Roozbeh Kiani, Leah Corthell, and Michael N. Shadlen. 2014. Choice certainty is informed by both evidence and decision time. *Neuron* 84, 6 (2014), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

[20] Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson. 2019. The effects of anthropomorphism and non-verbal social behaviour in virtual assistants. In *Proceedings of the 19th ACM Int'l Conf. on Intelligent Virtual Agents*. 133–140.

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

[22] Emiel Krahmer and Marc Swerts. 2005. How children and adults produce and perceive uncertainty in audiovisual speech. *Language and speech* 48, Pt 1 (2005), 29–53. https://doi.org/10.1177/00238309050480010201

[23] Nicole C Krämer, Astrid von der Pütten, and Sabrina Eimler. 2012. Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. *Human-computer interaction: The agency perspective* (2012), 215–240.

[24] Jan Leusmann, Chao Wang, Michael Gienger, Albrecht Schmidt, and Sven Mayer. 2023. Understanding the Uncertainty Loop of Human-Robot Interaction. *arXiv preprint arXiv:2303.07889* (2023).

[25] Sheng Li and Feitong Yang. 2012. Task-dependent uncertainty modulation of perceptual decisions in the human brain. *The European journal of neuroscience* 36, 12 (2012), 3732–3739. https://doi.org/10.1111/ejn.12006

[26] Jackson Liscombe, Julia Bell Hirschberg, and Jennifer J Venditti. 2005. Detecting certainness in spoken tutorial dialogues. In *Proceedings of Interspeech*.

[27] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. 2021. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems* 14, 2 (2021), 715–729.

[28] Tim Paek and Yun-Cheng Ju. 2008. Accommodating explicit user expressions of uncertainty in voice search or something like that. In *Ninth Annual Conference of the International Speech Communication Association*.

[29] Heather Pon-Barry, Karl Schultz, Elizabeth Owen Bratt, Brady Clark, and Stanley Peters. 2006. Responding to student uncertainty in spoken tutorial dialogue systems. *Int'l Journal of Artificial Intelligence in Education* 16, 2 (2006), 171–194.

[30] Heather Pon-Barry and Stuart M Shieber. 2011. Recognizing Uncertainty in Speech. *EURASIP Journal on Advances in Signal Processing* 2011 (2011), 11. https://doi.org/10.1155/2011/251753

[31] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9. https://doi.org/10.1109/WACV.2016.7477679

[32] Johanna Renker and Gerhard Rinkenauer. 2016. The acquisition of mental representations under uncertainty: an eye movement study. *Kognitive Systeme* 2016, 1 (2016).

[33] Marion Rouault, Tricia Seow, Claire Gillan, and Stephen Fleming. 2018. Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry* 84. https://doi.org/10.1016/j.biopsych.2017.12.017

[34] Joshua I. Sanders, Balázs Hangya, and Adam Kepecs. 2016. Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron* 90, 3 (2016), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

[35] Lisa Scherf, Cigdem Turan, and Dorothea Koert. 2022. Learning from Unreliable Human Action Advice in Interactive Reinforcement Learning. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 895–902.

[36] Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. 2021. Leveraging recent advances in deep learning for audio-Visual emotion recognition. *Pattern Recognition Letters* 146 (2021), 1–7. https://doi.org/10.1016/j.patrec.2021.03.007

[37] Tobias Schrank and Barbara Schuppler. 2015. Automatic detection of uncertainty in spontaneous german dialogue. In *Sixteenth annual conference of the international speech communication association*.

[38] A. Selimbeyoglu, Y. Keskin-Ergen, and T. Demiralp. 2012. What if you are not sure? Electroencephalographic correlates of subjective confidence level about a decision. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* 123, 6 (2012), 1158–1167. https://doi.org/10.1016/j.clinph.2011.10.037

[39] Shane Sims and Cristina Conati. 2020. A Neural Architecture for Detecting Confusion in Eye-tracking Data. *arXiv preprint arXiv:2003.06434* (2020).

[40] Vicki L. Smith and Herbert H. Clark. 1993. On the Course of Answering Questions. *Journal of Memory and Language* 32, 1 (1993), 25–38. https://doi.org/10.1006/jmla.1993.1002

[41] Matthew Stone and Insuk Oh. 2008. Modeling Facial Expression of Uncertainty in Conversational Animation. In *Modeling Communication with Robots and Virtual Humans*, Ipke Wachsmuth and Günther Knoblich (Eds.). Lecture Notes in Computer Science, Vol. 4930. Springer Berlin Heidelberg, Berlin, Heidelberg, 57–76. https://doi.org/10.1007/978-3-540-79037-2_4

[42] Shinsuke Suzuki. 2022. Inferences regarding oneself and others in the human brain. *PLoS Biology* 20, 5 (2022), e3001662.

[43] Marc Swerts and Emiel Krahmer. 2005. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language* 53, 1 (2005), 81–94.

[44] Susanne Trick, Dorothea Koert, Jan Peters, and Constantin A Rothkopf. 2019. Multimodal uncertainty reduction for intention recognition in human-robot interaction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7009–7016.

[45] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Bjorn W. Schuller, and Stefanos Zafeiriou. 2017. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309. https://doi.org/10.1109/JSTSP.2017.2764438

[46] Anne Urai, Anke Braun, and Tobias Donner. 2017. Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications* 8 (03 2017), 14637. https://doi.org/10.1038/ncomms14637

[47] Charlotte Wollermann, Bernhard Schröder, and Ulrich Schade Fraunhofer. 2014. Audiovisual prosody of uncertainty: An overview. *Ricerche di Pedagogia e Didattica. Journal of Theories and Research in Education* 9, 1 (2014), 137–157.