PERSONALIZED EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR DECENTRALIZED AGENTS WITH HETEROGENEOUS KNOWLEDGE

D4.2
Scientific paper about model and mechanisms for the conflict detection, impact of different communication modalities employed in the explanation personalization process. [M34]

Version: 1.0
Last Update: 18/11/2024
Distribution Level: CO

Distribution levels*
PU = Public;
RE = Restricted to a group within the given consortium;
PP = Restricted to other program participants (including commission services);
CO = Confidential, only for the members of the EXPECTATION Consortium (including commission services);

Project supported by

EXPECTATION

The EXPECTATION Project Consortium groups Organizations involved:

Partner Name	Short name	Country
University of Applied Sciences and Arts Western Switzerland HES-SO Institute of Information Systems	HES-SO	Switzerland
University of Bologna / Dep. Of Computer Science and Engineering	UNIBO	Italy
University of Luxembourg, Department of Computer Science	UNILU	Luxembourg
Özyeğin University	OZU	Turkey

Document Identity

Creation Date: Last Update:	1/05/2021 18/11/2024

Revision History

Version	Edition	Author(s)	Date
1	1	All PIs	18/11/2024
Comments:			

Contents

This deliverable consolidates multiple scientific dissemination efforts, which have been synthesized and integrated into a comprehensive document. The merging of these various contributions ensures a cohesive representation of the scientific findings and discussions. Below is the table of contents included.

Paper Title	Partners	DOI
Evaluation of the User-Centric Explanation Strategies for Interactive Recommenders	OZU, HES-SO	https://doi.org/10.1007/978-3-031-70074-3_2
Effects of Communication Channels on Explainable Food Recommendation Systems	OZU, HES-SO, UNILU	Accepted, awaiting publication

Project supported by

Evaluation of the User-centric Explanation Strategies for Interactive Recommenders

Berk Buzcu[1,2][0000−0003−1320−8006], Emre Kuru[1][0009−0007−6130−6272], Davide Calvaresi[2][0000−0001−9816−7439], and Reyhan Aydoğan[1,3][0000−0002−5260−9999]

[1] Computer Science, Özyeğin University, Turkey
[2] University of Applied Sciences and Arts Western Switzerland, Switzerland
[3] Interactive Intelligence, Delft University of Technology, Delft, Netherlands

Abstract. As recommendation systems become increasingly prevalent in numerous fields, the need for clear and persuasive interactions with users is rising. Integrating explainability into these systems is emerging as an effective approach to enhance user trust and sociability. This research focuses on recommendation systems that utilize a range of explainability techniques to foster trust by providing understandable personalized explanations for the recommendations made. In line with this, we study three distinct explanation methods that correspond with three basic recommendation strategies and assess their efficacy through user experiments. The findings from the experiments indicate that the majority of participants value the suggested explanation styles and favor straightforward, concise explanations over comparative ones.

Keywords: Explainable Recommendations · Explanation Strategies · User Studies

1 Introduction

In the rapidly evolving technological environment, our dependence on algorithm-powered recommendation systems for a variety of decision-making processes is growing. These systems are used in a wide range of applications, from suggesting content on movie streaming services to recommending products on e-commerce platforms. While these systems prioritize the selection and presentation of recommendations, they often overlook the user's curiosity about the rationale behind the recommendations. To address this, it is essential to engage users in an interactive communication setting. This allows users to delve deeper into the reasoning behind the recommendations, fostering a stronger understanding of the domain and satisfying their curiosity about the "why" behind the recommendations. This interactive setting necessitates methods for the system and users to express themselves, akin to a conversation between a sales assistant and a customer. Enhancing the recommender system's ability to express itself can make it more user-friendly, potentially leading to more effective outcomes.

In this context, our study aims to illuminate the workings of food recommendation systems, with a particular emphasis on their use in providing healthconscious dietary recommendations. Our primary goal is to enhance the trustworthiness and credibility of these food recommendation systems by equipping them with the ability to offer informative explanations for their recipe recommendations. To achieve this, we investigate recommendation strategies and their corresponding explanation generation strategies in two main categories: (i) model-agnostic explanations and (ii) model-intrinsic explanations. The former involves generating explanations by examining the results of the recommendation strategy using a separate model, a process known as "post-hoc" explanation generation. The latter uses a single model to generate both recommendations and explanations, making them "intrinsically explainable". Numerous studies have experimented with model-agnostic explanations due to the increasing predictive power of black-box models [27, 1, 7, 28, 18]. However, other research argues that if the generated explanations are not connected to the model's decision-making process, the system cannot be considered transparent [17, 21], implying that explanations and recommendations should not be separated.

In light of this, our study employs and evaluates basic recommendation strategies from existing literature along with their corresponding explanation methods. Those explanation strategies can be categorized as *tree-based model agnostic*, *cluster-based model-agnostic* and *popularity-based model-intrinsic* explanation generation approaches. We scrutinize current approaches for explanation generation, incorporate them into our food recommender system, and compare the methods through user experiments, assessing user satisfaction and the effectiveness of the explanations. In the subsequent sections, we first outline the baseline strategy from the literature, followed by a detailed presentation of our proposed strategies.

2 Related Work

This section briefly overviews the literature on explainable recommender systems and different explanation strategies to persuade and convince users about given recommendations. Recent studies have emphasized incorporating explanations into recommendations to enhance transparency, trust, and acceptability. For instance, Tintarev and Masthoff investigate various aspects of explanations' impact, such as transparency, scrutability, trustworthiness, persuasiveness, effectiveness, efficiency, and satisfaction. Experiment results showed that in various domains especially in low investment domains, providing explanations is likely to improve these aspects of a recommender system [25]. All of these attributes enhance the system's reliability as supported by Gedikli *et. al.* in which they assessed varying explanation attributes, including user satisfaction, efficiency, effectiveness, and trust, by evaluating different explanation styles in recommender systems via user study responses [9] (e.g., empirically they use response times to measure explanation effectiveness and subjectively user ratings for user satisfaction). Additionally, Herlocker *et. al.* investigate the effects of varying techniques used to explain collaborative-filtering recommendation methods. They follow the principle of collaborative filtering in their recommendation strategy and show ratings of similar users to the underlying user in the form of explanations [12]. They could not reach any concrete outcome according to their hypotheses that adding collaborative filtering based explanations to recommendation systems would improve the acceptance of that system and the performance of filtering decisions within the groups of ordinary users; however, they find out their system makes it more convenient for an expert to sympathize with a recommendation (e.g., the group of experts is more fond of the system with higher success in predictions of user acceptance).

Sharma and Cosley devised a framework to investigate the influence of social explanations (e.g., explanations that are similar to collaborative filtering in nature, where they are generated according to a grouping of users, and relating to other users contextually) within music recommenders [19]. They found that varying explanations might have different effects depending on the person. Similarly, Milliecamp explores the visual explanations within the music domain. They show that users react to explanations depending on their need for cognition, confidence, sophistication, and visualization literacy. Explanations boost confidence for those with a low need for cognition and speed up song judgments for those with higher musical sophistication. Users with lower visualization literacy tend to judge songs more quickly and precisely [15]. Furthermore, Pu and Chen develop an explanation interface to investigate the user experience advantages of using explanations for building trust and to assess whether system features can contribute to trust-related benefits [16]. They show that users prefer to re-use systems that offer explanations more often than those that do not, and users prefer a comparative explanation style where they get a broader view of available items and respective differences.

Besides, Balog, Radlinski, and Arakelyan present a set-based recommendation framework that utilizes interrelated features for generating explanations that account for conditional preferences [2]. Such as liking "Science Fiction" movies only when it's about "Space Exploration". Symeonidis *et. al.* introduces a prototype for a movie recommender system designed to gauge user satisfaction via various explanation styles [23]. They point out that providing an explanation along with movie recommendations will increase the likelihood of a user estimating its movie ranking while also increasing the number of correct estimations to predict a user's favorite movie by boosting the user's confidence in providing information to the system. Guesmi *et. al.* showed that users have different goals and may react differently to given explanations [11]. They claimed through their work that explanations are not a one-size-fits-all solution and that the explanations should be customized according to the characteristics of the users. In the following section, we survey the existing explanation mechanisms for recommenders.

In recent years, ample research has focused on developing model-agnostic (i.e., post-hoc) explanation generation techniques in machine learning, where explanations are generated after the predictions are made without requiring modifications to the underlying model's architecture or training process. The goal is to improve transparency and interpretability while not decreasing the accuracy of the predictions. Post-hoc explanation generation models often leverage feature importance analysis [18], rule-based reasoning [27], gradient-based attribution [1], or surrogate models [28] to generate meaningful explanations that can shed light on the factors influencing the model's predictions. These explanations help stakeholders gain insight into how the given model makes its decisions, build trust, and facilitate error analysis [7].

Unlike post-hoc explanations, model-intrinsic techniques focus on generating explanations directly extracted from the internal mechanisms of the recommendation models. Thus, the generated explanations offer a solid understanding of the model's decision-making process. Rago *et. al.* present a novel graphical framework, which establishes connections between items and their aspects within a recommendation system [17]. This framework utilizes Tripolar Argumentation Frameworks (TFs), an extension of the classical argumentation frameworks, to represent relationships among the items' features and the recommendations. TFs incorporate three distinct types of relations: positive, negative, and neutral, signifying whether an aspect of an item supports, attacks, or remains neutral to a recommendation. Through these relations, the users have the flexibility to customize their explanations based on their queries. If a user seeks an explanation as to why an item was recommended, the system can focus on highlighting the positive aspects of the item (supporters) within the Tripolar framework. Conversely, if a user wishes to understand why an item was not recommended, the system can emphasize the negative aspects (attackers). Shimizu, Matsutani, and Goto improve the state-of-the-art knowledge graph attention network (KGAT) by significantly decreasing its computation time, thus allowing more side information for generating explanations [21]. Here, KGAT represents the relationships between users, items, and their side information. Using attention weights to signify the importance of a node's or an edge's influence on a recommendation. When the model makes a recommendation, it can explain why it made that particular recommendation by highlighting the nodes or edges in the knowledge graph that received the highest attention weights.

Recent studies have succeeded in the realm of *counterfactual* explanations. These mechanisms aim to provide users with insightful explanations for predictions by generating counterfactual instances through "what-if" scenarios. It inquires whether a particular interaction or an attribute of the recommended item may influence any changes in the recommendation. Tan *et. al.* extract aspect-aware explanations by looking for the minimal change in the recommended items' features such that the item would have not been recommended anymore; thereby finding the most crucial features for explanations [24], whereas Tran,Ghazimatin and Roy generate explanations by observing how much the recommendation changes if certain interactions were missing from the training dataset [26]. Mainly, they focus on whether their appreciation of an item would change if they did not experience any particular product before. Table 1 summarizes the related explanation approaches in the literature.

Table 1: Comparison Matrix of Explanation Approaches

Study	Explanation Type	Visual Text	Approach	Model Agnostic Intrinsic	Domain	Dynamic Static
Guesmi [11]	User-Centered	Both	NLP	Model-agnostic	Articles	static
Buzcu [5]	Contrastive	Text	Decision Trees	Model-agnostic	Food	static
Balog [2]	Knowledge-Based	Text	Knowledge-Based	Model-intrinsic	Movies	static
Shimizu [21]	Example-Based Knowledge-Based	Text	Knowledge-Graph	Model-intrinsic	Products	static
Tan [24]	Counterfactual	Text	Neural Network	Model-intrinsic	Movies	dynamic
Rago [17]	Content-based	Text	Knowledge-Graph	Model-intrinsic	Movies	static
Tran [26]	Counterfactual	Text	Neural Network	Model-intrinsic	Products	dynamic
Our Approach	Contrastive User-Centered Content-Based	Text	Clustering Random Forest	Both	Food	dynamic

3 Recommendation and Explanation Strategies

In this study, we adopt three basic recommendation strategies with aligned explanation approaches.

3.1 Baseline Recommendation and Explanations

Baseline Recommendation Generation is adopted from [3], which relies on filtering and scoring recommendations by considering underlying conditions and users' preferences. First, the system filters items (e.g., food recipe, movie) with respect to the user's constraints. In turn, the utilities of the remaining candidates are calculated through a scoring function. The items are sorted according to the computed utilities. The item with the highest utility, which was not recommended before, is selected as a recommendation, and the system retroactively generates an explanation in line with the recommendation's properties/features. For this baseline recommendation, Buzcu *et al.* introduce two types of explanation generation methods, which will be explained briefly below.

- **Item & User Explanations:** A decision tree is constructed from historical data in which recommendations are labeled with all users' decisions (i.e., accept or reject) in the user-based explanation approach. In contrast, items are labeled according to **the current user's constraints and feedback** in the item-based explanation generation approach. We can extract the importance of the features while building the decision tree. This approaches pick the most important three features to generate an explanation for the given recommendation.
- **Contrastive Explanations:** This type of explanations can be generated by referring a *contrastive item*, which is an item similar to the chosen one but fails to satisfy user constraints/preferences. For this purpose, the most similar item is selected from the aforementioned candidate set of items with the current recommendation. The features of the selected item with those of the recommendations are compared one by one. The features influencing

the user satisfaction positively or negatively are used to build explanations that highlight the positive side of the recommendation while sending away the contrastive item by emphasizing its opposing sides.

3.2 Enhanced Baseline Recommendations & Cluster-based Explanations

Recommendations is decided in a similar way to the aforementioned approach while explanations are generated by relying on the clustering approach proposed in [4]. The score function of the recommendation strategy is more comprehensive, thus we name it the Enhanced Baseline Recommendation. According to this approach, items are clustered with respect to the user's estimated preferences and desires. As usual, it is expected to have similar behavior or pattern in the same cluster, and someone can inquire which features distinguish those items in the same cluster from other clusters. In the proposed approach, each item is represented as a vector of evaluation criteria (e.g., the preference score for food recommendation). As illustrated in Figure 1, a clustering algorithm is applied to determine distinguishable items concerning users' preferences and needs. For each cluster a separate classifier (Random Forest Classifier) is trained to detect whether or not the item belongs to underlying cluster. The feature importances, particularly the most important feature, can be extracted from this classifier to generate the explanations.

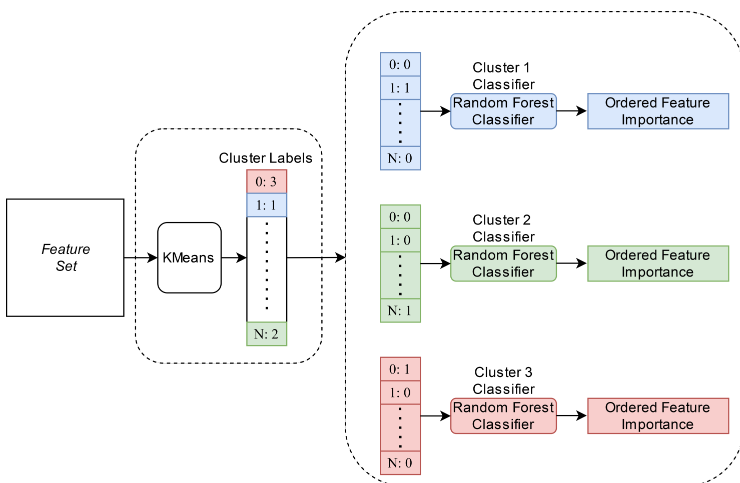


Fig. 1: Process of Clustered-Based Explanation [4]

While generating the contrastive explanations, the most similar item to the recommended item from another cluster is selected. By comparing the values of each feature of the contrastive item with those of the recommended item, positive and negative features are detected. While generating an explanation, the recommended item is promoted with the positively contrastive features, whereas negative features indicate why the system does not suggest the contrastive example.

3.3 Popularity-based Explanations

The explanation generation techniques employed thus far have been model-agnostic, which could result in explanations that do not accurately reflect the actual decision-making process. This discrepancy arises because the system's overall outcomes may need to align better with the individual recommendations made by the system. In this approach, we utilize a historical dataset capturing the acceptance of past recommendations and a machine learning approach for their classification. In particular, a Random Forest Classifier is utilized to generate recommendations and explanations. Therefore, the proposed structure is inherently connected to the recommendation process, as both are generated from the same model. The Random Forest classifier is a valuable choice for popularity-based recommendation systems because it handles large-scale datasets and provides robust predictions while being explainable [22]. While making a recommendation, we calculate the probability estimates of each item and sort them based on the probability of acceptance (labeled as "1"). The recommendation algorithm combines this knowledge with personalization derived from previous sections and recommends the recipe with the highest level of probability of acceptance. On the other hand, we utilize the feature importance generated by the same Random Forest model in generating an explanation.

Algorithm 1 details the recommendation selection and explanation generation process. First, we train a Random Forest Classifier using our popularity-labeled data (Line 1). Then, we calculate the posterior probability for each class (Line 2). The class labels are acceptable (1) or unacceptable (0). We select the item with the maximum probability of acceptance (Line 2). This item's features are then compared according to the Random Forest model's feature importance vector. Finally, the most important feature is selected to construct the explanation (Line 3-4).

Additionally, we utilize the Popularity-based recommendation approach to generate contrastive explanations befitting the context. Algorithm 2 explains how contrastive explanations are generated accordingly. First, we extract the subset of recipes labeled as not recommended by the algorithm (Line 1). Then, we select the item with the minimum distance in the feature space to the current recommendation (Line 2). Finally, we choose the most important feature according to the feature importance of the Random Forest Classifier (Line 2).

4 Case Study: Food Recommendation

For baseline recommendation strategy, we use the original feature set in [3] to represent each food recipes as follows: *calorie count*, *fat amount*, *carbohydrates amount*, *fibers*, *preparation time*, *protein amount*, *preference score*. For

Algorithm 1 Popularity-based Recommendation & Explanation

Require:
F: Features of items;
I: All items,
I_p: Subset of popular recipes that were recommended in previous experiments; labeled "1" if they were ever accepted or "0" if they were recommended but not accepted;

Ensure: ϵ: Selected feature; r: Selected recommendation;

$randomForest \leftarrow RandomForestClassifier(I_p)$
$R \leftarrow randomForest.predictproba(I)$
$r \leftarrow \text{argmax}\, R[1]$
$\epsilon \leftarrow max(randomForest.featureImportance(F)$
return ϵ, r

Algorithm 2 Popularity-based Contrastive Explanation Selection

Require:
F: Feature set of Popularity;
P_u: Users feature preference weights;
I: Set of scored items;
i: Recommended item;

Ensure: ϵ: Explanation feature; r': Contrastive item;

$C \leftarrow R[0]$
$r' \leftarrow \text{argmin}_{c \in C}\, distance(c, r)$
$\epsilon \leftarrow \text{argmax}_{f \in F}$ featureImportance(r)
return ϵ, r'

our clustered-based recommendation strategy, we consider a vector of *preference score*, *health score*, *price score*, *time score*, and *taste score* with respect to the user profile since those scores captures users' preferences more comprehensively. For the popularity-based strategies, we have conducted an additional experimental study to gather a labeled dataset where if any user ever accepts them, the recommended recipes are labeled as "1"; otherwise, they are labeled as "0". Note that the recipes that were never recommended are excluded from the dataset. We applied one-hot encoding to the features (e. g., *flavors of food*, *meal type*, *price*, and *cooking style*) to classify them accordingly. We utilized this data to train a Random forest classifier to predict whether a recommendation will be accepted or rejected by the user based on its popularity label in those experiments.

In the following part, we explain how aforementioned scores are calculated to suit the clustering technique used in our user experiments as features.

Preference Score: To calculate the preference score of a user for the recipe dataset, we utilize a novel Active Learning framework [6] that is proven effective within our research project. The system first generates a diverse sample of recipes from the dataset. It asks the user to specify whether they like or dislike a given recipe. Afterward, the system shows a set of recipes to the user. The participants are asked to indicate the correct labels for the predictions made by the system by adjusting whether they like its respective features or not. This user feedback is utilized to generate synthetic data to enrich the user's preference data and increase the system's accuracy with a small dataset. Ultimately, the labeling generated from this process is used to train a supervised learning model, Logistic Regression. We use the positive class probability by the model as the indicator for the user preference score for a given recipe, which corresponds to the likelihood of user's acceptance of a recipe.

Health Score: In order to calculate the health score, we follow intuition, where we take the user's personal information to calculate their nutritional needs in a day. The final *healthScore* is the mean of *calorieScore* and respective nutrient scores for each nutrition value of the recipe. To calculate the *calorieScore* we use the daily active metabolic rate (AMR) described in the literature [5] and the final score is derived as described in Equation 1 where R corresponds to a recipe within the dataset. Essentially, we constrain the calorie score within the range of [0, 1], and we assume that a higher amount of calories is better as long as it is less than the active metabolic rate. For the nutrient scores, we use the nutrient density score [8] as described in Equation 2. To simplify the healthiness decision, each nutrition is scored higher if they have a higher density per calorie except for the amount of fat, which is reversed $(1 - fatScore)$. The $w_{nutrient}$ corresponds to the pre-defined weight for each nutrient. Currently, all the weights are equal given our use-case does not define a distinction for nutrient importance. Finally, all scores are clamped to the $[0, 1]$ range, then averaged to derive the final *healthScore*.

$$calorieScore_R = \left\{ \begin{matrix} \frac{R_{calories}}{max_{calories}}, \text{ if } calories_R \leq AMR \\ 0, \hfill \text{else} \end{matrix} \right\} \tag{1}$$

$$nutrientScore_R = \frac{amount_{nutrient}(gr)}{calories_R(kcal)} \tag{2}$$

$$healthScore_R = \sum_{nutrient} w_{nutrient} * score_{nutrient} \tag{3}$$

Price Score: We first label the recipes within three classes; cheap, standard and expensive (labelled as \$, \$\$, \$\$\$ in order). We assume that the cheaper is better for a given food recipe, therefore, we assign these classes the scores of 1, 0.67, 0.33 respectively.

Time Score: The time scores are a summation of the recipes preparation and cooking time. For the calculation of this score, we assume that the quicker is better. Therefore, we apply max-normalization on the time of preperation in terms of minutes and reverse the order of scores for all the recipes. Thus, the quickest recipe is scored as “1”.

Taste Score: The taste score corresponds to how well the flavor preferences of the user matches the flavor profile of a recipe. The flavor profile is comprised of the following tastes: *Savory*, *Bitter*, *Sour*, *Salty*, *Sweet* and *Spicy*, which are recognized as the main tastes the humans can distinguish [10]. Each food recipe holds boolean fields for the dimensions of a flavor profile. Table 2 shows an example recipe where each taste is labelled “1” if it is a part of the profile, or “0” if it is not. Finally, we ask user to elicitate their desired tastes in the same

Table 2: Flavors of a Tomato Soup

Recipe	Savory	Spicy	Sour	Salty	Sweet	Bitter
User Profile	1	1	1	1	0	0
Tomato Soup	1	0	1	1	0	0

form (e.g., they mark whether or not they want it in a boolean fashion). For each recipe, we apply the Jaccard Similarity [13] as described in the Equation 4 to calculate the final score within the range of [0, 1].

$$tasteScore_R = \frac{flavors_R \cap flavors_{user}}{flavors_R \cup flavors_{user}}, \tag{4}$$

5 Evaluation

In order to thoroughly evaluate the proposed explanation generation strategies we conducted user experiments via a Web-based interface for food recommendations [4]. The experimental setup is presented in Section 5.1, consecutively, Section 5.2 reports and discusses the experimental results elaborately.

5.1 Experimental Setup

Prior to commencing the experiments, each participant was required to fill out a pre-survey and registration form, wherein they provided details about their gender, age, height, weight, level of physical activity, dietary preferences, and any allergies they might have. Additionally, they were asked to rank food related factors, and their taste preferences. The system utilizes this information to score the recipes respective to each participant’s healthiness and preferences (Section 4). To evaluate the acceptability and effectiveness of the explanation-generation techniques proposed, we conducted a study involving participants experiencing three iterations of food recipes, each accompanied by three explanations. The system presents a recipe each time in the following order of recommendation strategies: (i) Baseline Recommendation (Section 3.1), (ii) Enhanced Baseline Recommendation and respective explanations compatible with the recommendation strategy (Section 3.2), and (iii) Popularity-Based Recommendation (Section 3.3):

[4] The user experiments in this study was reviewed and approved by the Ethics Committee of Özyeğin University, and informed consent was obtained from all the experiment participants.

- Baseline Recommendation: We use the following explanation methods: Item-based, User-based and Contrastive explanations (Section 3.1).
- Enhanced Baseline Recommendation (Section 3.2): We employ the following explanation methods: Cluster-based, Contrastive Cluster explanations, and Enhanced Item-based Section 3.1.
- Popularity Recommendation (Section 3.3): We apply the following explanation methods: Popularity-based, Contrastive Popularity, and Popularity User-based (Section 3.1) explanations.

Here, we adapted the User and Item-based explanation generation strategies described in Section 3.1 to generate explanations with the proposed recommendation strategies and their respective features. Afterward, the participants were asked to provide feedback on the perceived performance (effectiveness and convincability) of these explanations using a 5-point Likert scale. After completing the rating of the explanations, the system asks the user to choose their favorite recipe among the shown recipes with respective explanations. This design choice was based on research suggesting that users make better-informed decisions without experiencing excessive cognitive load when selecting from three items simultaneously [20]. As shown in the Figure 2, the user can easily view nutritional information, recipe ingredients, and various explanations. However, the food's picture and detailed recipe information are not immediately visible, but the user can still access them by clicking the respective buttons, similar to earlier studies. After concluding the experiment, the participants are requested to complete a questionnaire consisting mainly of 5-point Likert scale questions. The questionnaire aims to assess their experiences with the explanations provided by the system. Participants are shown an explanation generated by the system and asked to respond to seven questions designed to gauge the effectiveness and success of the explanations they received.

In total, there are 80 participants (25 female, 55 male) with diverse backgrounds and age groups took part in the test (mean 24.70, min: 18 and max:59). The participants were requested to order the importance of five criteria, relative to a given food recommendation: "Nutritional factors", "Past experience with taste", "How it looks", "Price of the ingredients", and "Cooking style". Participants were asked to rate various factors on a scale from 1 to 5, with 1 indicating the highest importance. The results show that a significant portion of the participants (specifically, 46%) considered their experience with the taste of such food to be the most critical factor influencing their cooking recipes. Additionally, 35% of the participants prioritized the healthiness of the food as their top concern. Conversely, 41% of the participants considered the time required to prepare the recipe the least important factor. In contrast, 28% of the participants rated the food price as the least significant factor in their decision-making process. The dataset used in the experiment is acquired from Diyetkolik and it is comprised of 1382 recipes, where 210 of them were recommended to the users. In total, 125 of those recipes were accepted by the users cumulatively from previous studies [3].

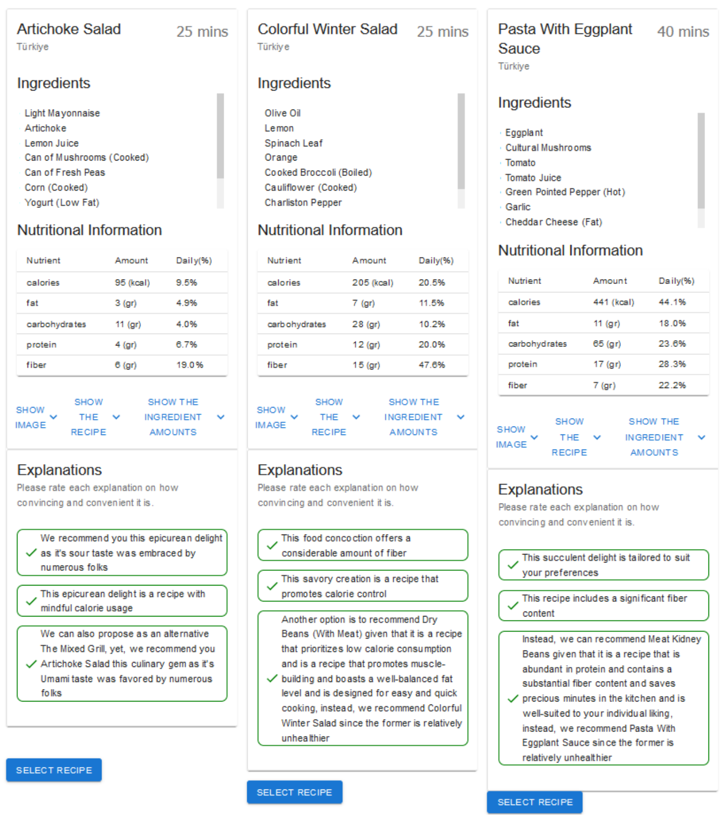


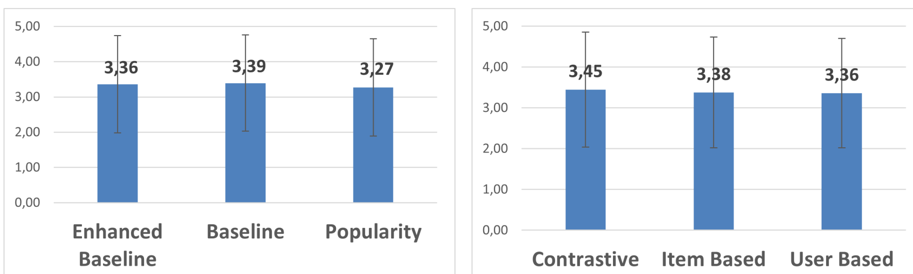
Fig. 2: User Interface for Recipe Selection Step

5.2 Experimental Results

The experimental setup is mainly comprised of experiment participants providing subjective input on explanations offered to recommendations. We applied the Repeated Measures ANOVA statistical test rejected the null hypotheses, which revealed a significant difference among the types of explanations (F=3.71, p=0.0003). For furter analysis, the data, as determined by the Kolmogorov normality test, does not conform to a normal distribution, a crucial assumption for conducting pairwise T-tests. Consequently, we opted for the appropriate non-parametric alternative, the Wilcoxon signed-rank test [14], for our statistical tests. In all our analyses, we set the Confidence Interval (CI) to 0.95, corresponding to a significance level of $\alpha = 0.05$. Our test results comprised of user responses to explanations are first analyzed by the recommendation strategy.

We conducted pairwise tests between these groups, yielding the following results: Enhanced Baseline vs. Popularity Recommendation ($p = 0.13$), Enhanced Baseline vs. Baseline Recommendation ($p = 0.44$), and Popularity vs. Baseline Recommendation ($\mathbf{p} = \mathbf{0.04}$). One could notice that the explanations generated along the Popularity-based recommendation have underperformed compared to the ones generated with the baseline recommendation strategy, as seen in Figure 3a. Before drawing such conclusions, we must look into further analysis. Additionally, we categorized explanation generation techniques based on their underlying differences and the form of labelling strategy they utilized.

- Item Based: Methods that utilize an item's attributes on-line (Basic Cluster, Enhanced Item-based, and Item-based).
- User Based: Techniques that use historical data (user acceptance) as labeling (Popularity, Popularity User-based, and User-based).
- Contrastive: Explanations that are in the contrastive form (Contrastive Cluster, Contrastive Popularity, and Contrastive).



(a) Recommendation (b) Explanation Generation

Fig. 3: Avg. Scores of Explanations Grouped per Technique

We conducted pairwise tests between these groups, leading to the following outcomes: Contrastive vs. Item-based ($\mathbf{p = 0.02}$), User-based vs. Item-based ($p = 0.84$), and Contrastive vs. User-based ($\mathbf{p = 0.02}$). Observing Figure 3b, we note that the contrastive explanations under-performed slightly compared to the other methods statistically. We notice a trend where the participants prefer explanations based on the attributes of the recommended item more. The participants did not appreciate both contrastive explanations and popularity-based metrics, potentially pointing to the fact that users care more about facts on their recommendation than comparative explanations.

Subsequently, we conducted pairwise tests to compare different types of explanation generation techniques within the same recommendation strategy individually as follows:

Popularity-Based Recommendation:

- Popularity vs. Popularity User-based ($p = 0.70$)
- Popularity vs. Contrastive Popularity ($p \leq \mathbf{0.0001}$)
- Contrastive Popularity vs. Popularity User-based ($p \leq \mathbf{0.0001}$)

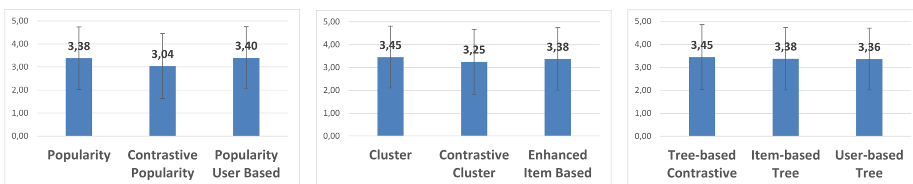
Enhanced Baseline Recommendation:

- Cluster vs. Enhanced Item-based (p = 0.78)
- Cluster vs. Contrastive Cluster (p = 0.13)
- Contrastive Cluster vs. Enhanced Item-based (p = 0.16)

Baseline Recommendation:

- Item-Based vs. User-Based (p = 0.92)
- Contrastive vs. Item-Based (p = 0.44)
- Contrastive vs. User-Based (p = 0.31)

These results provide insights into the comparative performance of explanation techniques within each recommendation strategy. The significant p-values (highlighted) indicate noteworthy differences deserving further investigation. We note that only in popularity group that the contrastive explanations significantly under-perform as shown in Figure 4. In other recommendation groups, there is no significant results observed.



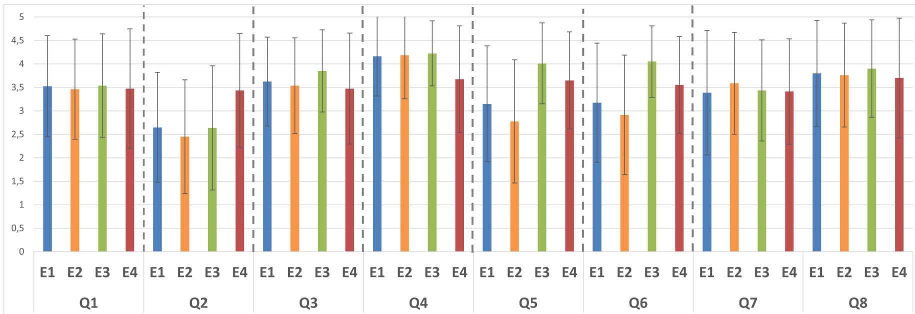
(a) Popularity Based (b) Cluster Based (c) Baseline

Fig. 4: Avg Scores of Explanations by Recommendation

Additionally, the distribution of accepted recipes are (i) Baseline recommendation 38%, (ii) Enhanced baseline 32% and (iii) Popularity-based 30%. Our take-aways may be further supported by this outcome given the simplest method of recommendation seems to be favored more than the others. However, this may be just a result of combination of explanations and the food recipe being more fitting to the users.

Moreover, we conducted an analysis of user responses to the post-experiment survey to assess their perceptions of the given explanations. Since each participant was given each form of explanation during the experiment, and the survey questions as well as the provided examples of the types were identical for all participants, we employed a within-subjects statistical comparison test. Figure 5 shows the average responses to the questionnaire questions, as well as the questions and respective explanations.

Table 3 shows the Wilcoxon paired test results for each type of explanation for each question. We observe significant differences between pairs of explanations on Q2, Q3, Q4, Q5 and Q6 whereas there is no significance within explanation



Id	Label
Q1	This type of explanation for recommendations has helped me choose the most convenient recipe.
Q2	This type of explanation for recommendations were too detailed.
Q3	This type of explanation displayed during the interaction were satisfactory.
Q4	This type of explanation for recommendations were clear and easy to understand.
Q5	This type of explanation were sufficient to make an informed decision for healthiness.
Q6	This type of explanation were realistic in terms of healthiness of given recipes.
Q7	This type of explanation let me know how convenient the recipe is.
Q8	Rate your appreciation of the idea of receiving this type of explanations in addition to recommendations.
E1	Popularity-based explanation sample
E2	Cluster-based explanation sample
E3	Baseline explanation sample
E4	Contrastive explanation sample

Fig. 5: Questionnaire Responses About Explanations

pairs for Q1, Q7 and Q8. The Q3 tells us that baseline explanations were the most satisfactory explanations, it is also the simplest explanation generation method. One could draw the conclusion that the participants favor simplistic methods over complicated ones. This finding is further supported by Q4, where contrastive explanations were found too complicated and the fact that they were rated lowest among the explanation types. The simpler explanations were found to be more effective in coming to healthy decisions, as it is seen from Q5. Finally, Q8 shows us that the users would still use this system despite it's short-comings with no significant difference among pairs of explanations.

Table 3: Pairwise Wilcoxon Test Results

P-Value	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**
E1 vs E2	0.565	0.045	0.549	0.679	**0.007**	0.063	0.294	0.909
E1 vs E3	0.986	0.683	**0.010**	0.573	$\leq$ **0.000**	$\leq$ **0.001**	0.902	0.579
E1 vs E4	0.878	$\leq$ **0.001**	0.426	**0.001**	**0.003**	**0.012**	0.859	0.635
E2 vs E3	0.524	0.053	**0.034**	0.621	$\leq$ **0.001**	$\leq$ **0.001**	0.498	0.475
E2 vs E4	0.855	$\leq$ **0.001**	0.638	$\leq$ **0.001**	$\leq$ **0.001**	$\leq$ **0.001**	0.284	0.751
E3 vs E4	0.656	$\leq$ **0.001**	**0.017**	$\leq$ **0.001**	**0.023**	**0.001**	0.641	0.228

6 Conclusion

In conclusion, this research contributes to the ongoing dialogue about incorporating explanation generation strategies into recommendation systems, especially those focused on health-aware recommendations. As we search to enhance the transparency and effectiveness of recommendation systems, we find that user-centrism, simplicity, and clarity are crucial for effective explanations. Despite these findings, it is important to acknowledge that the effectiveness of explanation strategies may vary depending on the specific user, rather than the collective user opinion on recommendation items guiding explanations. This study does not particularly focus on a group of individuals and it involves participants from diverse backgrounds and dietary preferences. Such diversity could affect their perspectives on the styles of explanations. Future studies could delve deeper into fine-tuning the explanation strategies toward user profiles and preferences, where we offer different styles of explanations at varying degrees to diverse profiles of users.

References

1. Ancona, M., Ceolini, E., Öztireli, A.C., Gross, M.H.: A unified view of gradient-based attribution methods for deep neural networks (2017)
2. Balog, K., Radlinski, F., Arakelyan, S.: Transparent, scrutable and explainable user models for personalized recommendation. In: Proceedings of the 42nd international acm sigir conference on research and development in information retrieval. pp. 265–274 (2019)
3. Buzcu, B., Tessa, M., Tchappi, I., Najjar, A., Hulstijn, J., Calvaresi, D., Aydoğan, R.: Towards interactive explanation-based nutrition virtual coaching systems. Autonomous Agents and Multi-Agent Systems **38**(1), 5 (Jan 2024). https://doi.org/10.1007/s10458-023-09634-5
4. Buzcu, B., Tessa, M., Tchappi, I., Najjar, A., Hulstijn, J., Calvaresi, D., Aydoğan, R.: User-centric explanation strategies for interactive recommenders. In: The 23rd International Conference on Autonomous Agents and Multi-Agent Systems (2024)
5. Buzcu, B., Varadhajaran, V., Tchappi, I., Najjar, A., Calvaresi, D., Aydoğan, R.: Explanation-based negotiation protocol for nutrition virtual coaching. In: International Conference on Principles and Practice of Multi-Agent Systems. pp. 20–36. Springer (2022)
6. Cantürk, F., Aydoğan, R.: Explainable active learning for preference elicitation p. 25 (08 2023). https://doi.org/10.21203/rs.3.rs-3295326/v1
7. Cemiloglu, D., Catania, M., Ali, R.: Explainable persuasion in interactive design. In: 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW). pp. 377–382 (2021)
8. Drewnowski, A., Fulgoni, V.L.: Nutrient density: principles and evaluation tools123. The American Journal of Clinical Nutrition **99**(5), 1223S–1228S (2014). https://doi.org/https://doi.org/10.3945/ajcn.113.073395, https://www.sciencedirect.com/science/article/pii/S0002916523050748
9. Gedikli, F., Ge, M., Jannach, D.: Understanding recommendations by reading the clouds. In: E-Commerce and Web Technologies. pp. 196–208. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)

10. Gravina, S.A., Yep, G.L., Khan, M.: Human biology of taste. Annals of Saudi Medicine **33**(3), 217–222 (2013). https://doi.org/10.5144/0256-4947.2013.217, https://www.annsaudimed.net/doi/abs/10.5144/0256-4947.2013.217
11. Guesmi, M., Chatti, M.A., Vorgerd, L., Ngo, T., Joarder, S., Ain, Q.U., Muslim, A.: Explaining user models with different levels of detail for transparent recommendation: A user study. In: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization. pp. 175–183 (2022)
12. Herlocker, J., Konstan, J., Riedl, J.: Explaining collaborative filtering recommendations. Proceedings of the ACM Conference on Computer Supported Cooperative Work pp. 241–250 (01 2001). https://doi.org/10.1145/358916.358995
13. Jaccard, P.: The distribution of the flora in the alpine zone.1. New Phytologist **11**(2), 37–50 (1912). https://doi.org/https://doi.org/10.1111/j.1469-8137.1912.tb05611.x, https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x
14. Lazar, J., Feng, J., Hochheiser, H.: Research Methods in Human-Computer Interaction (04 2017)
15. Millecamp, M., Htun, N.N., Conati, C., Verbert, K.: To explain or not to explain: The effects of personal characteristics when explaining music recommendations. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. p. 397–407 (2019)
16. Pu, P., Chen, L.: Trust building with explanation interfaces. In: International Conference on Intelligent User Interfaces, Proceedings IUI. vol. 2006, pp. 93–100 (01 2006). https://doi.org/10.1145/1111449.1111475
17. Rago, A., Cocarascu, O., Bechlivanidis, C., Lagnado, D., Toni, F.: Argumentative explanations for interactive recommendations. Artificial Intelligence **296**, 103506 (2021)
18. Saarela, M., Jauhiainen, S.: Comparison of feature importance measures as explanations for classification models. SN Applied Sciences **3** (02 2021)
19. Sharma, A., Cosley, D.: Do social explanations work? studying and modeling the effects of social explanations in recommender systems. WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web pp. 1133—1144 (04 2013)
20. Shimazu, H.: Expertclerk: Navigating shoppers' buying process with the combination of asking and proposing. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2. p. 1443–1448. IJCAI'01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
21. Shimizu, R., Matsutani, M., Goto, M.: An explainable recommendation framework based on an improved knowledge graph attention network with massive volumes of side information. Knowledge-Based Systems **239**, 107970 (2022)
22. Speiser, J.L., Miller, M.E., Tooze, J., Ip, E.: A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications **134**, 93–101 (2019). https://doi.org/https://doi.org/10.1016/j.eswa.2019.05.028, https://www.sciencedirect.com/science/article/pii/S0957417419303574
23. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Moviexplain: A recommender system with explanations. In: Proceedings of the Third ACM Conference on Recommender Systems. p. 317–320. RecSys '09, Association for Computing Machinery, New York, NY, USA (2009)
24. Tan, J., Xu, S., Ge, Y., Li, Y., Chen, X., Zhang, Y.: Counterfactual explainable recommendation. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 1784–1793 (2021)

25. Tintarev, N., Masthoff, J.: Explaining recommendations: Design and evaluation. In: Recommender systems handbook, pp. 353–382. Springer (2015)
26. Tran, K.H., Ghazimatin, A., Saha Roy, R.: Counterfactual explanations for neural recommenders. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1627–1631 (2021)
27. Xu, Y., Collenette, J., Dennis, L., Dixon, C.: Dialogue-based explanations of reasoning in rule-based systems. In: 3rd Workshop on Explainable Logic-Based Knowledge Representation (Jun 2022)
28. Zhu, X., Wang, D., Pedrycz, W., Li, Z.: Fuzzy rule-based local surrogate models for black-box model explanation. IEEE Transactions on Fuzzy Systems **31**(6), 2056–2064 (2023)

Effects of Communication Channels on Explainable Food Recommendation Systems

Berkecan Koçyiğit[1[0009−0005−3201−5834]], Berk Buzcu[2[0000−0003−1320−8006]], Davide Calvaresi[2[0000−0001−9816−7439]], Zehra Kesemen[1[0009−0000−3103−9096]], Joris Hulstijn[4[0000−0002−5476−1062]], Igor Tchappi[3[0000−0001−5437−1817]], and Reyhan Aydoğan[1,5[0000−0002−5260−9999]]

[1] Özyeğin University, Türkiye
[2] HES-SO Valais-Wallis, Switzerland
[3] University of Luxembourg, Luxembourg
[4] Utrecht University, Netherlands
[5] Delft University, Netherlands

berkecan.kocyigit@ozu.edu.tr,berk.buzcu@ozu.edu.tr,davide.calvaresi@hevs.ch, zehra.kesemen@ozu.edu.tr,j.hulstijn@uu.nl,igor.tchappi@uni.lu, reyhan.aydogan@ozyegin.edu.tr

Abstract. This study examines the effects of communication channels on user engagement, acceptance, and satisfaction within a personalized food recommendation system, that provides explanations. We tested two conditions: a tablet Web interface and a humanoid robot along with a tablet. Our findings demonstrate that, despite occasional user frustrations with the robot's slower response time and overall interaction style, the robot's presence enhances the social perception of the interaction. Overall, the results suggest potential for socially interactive robots.

Keywords: XAI · Food Recommender Systems· Human Robot Interaction · Effect of Communication Medium

1 Introduction

With the increasing integration of artificial intelligence (AI) technologies into everyday life, there is a growing demand for systems that not only provide precise recommendations but can also deliver transparent, understandable explanations [4]. Explainable AI (XAI) has been employed to address this need by making complex algorithmic processes more accessible to the end-user, thus promoting user trust and understanding [2]. However, explainability alone does not compose the effectiveness of the content, but also how the explanations are communicated plays a crucial role in how users perceive, engage with, and ultimately accept these recommendations [7]. Previous research has shown that physical presence of a robot can increase user engagement by providing a richer, more interactive experience that fosters a sense of connectedness [13]. Moreover, there is a notable effect of different communication channels (e.g., physical embodiment versus purely digital) for users' within personalized meal recommendation systems [15]. However, there is still a gap in understanding how embodiment translates to practical applications, particularly in scenarios where complex, personalized information needs to be communicated efficiently and effectively. To investigate this, we have developed two different interface configurations: (i) one based on a standard Web interface accessible with a tablet, and (ii) another that complements this tablet interface with a humanoid robot capable of providing food recommendations along with explanations through speech and gestures. The robot introduces physical embodiment and aims to create a more human-like interaction by utilizing multi-modal communication. By examining user reactions and interactions with both systems, we aim to determine whether physical embodiment influences user engagement and affects perceptions of the recommendations, including acceptance and satisfaction with the system.

The remainder of the paper is organized as follows. Section 2 presents related work. Section 3 describes the set-up of the study, detailing the components and configurations used. Section 4 provides an evaluation and discussion of the experimental results. Section 5 concludes the paper and suggests future research.

2 Related Work

The increasing adoption of artificial intelligence in day-to-day applications has developed an interest in understanding the most effective ways to deliver personalized and explainable recommendations. In particular, research has focused on the communication mediums: the means of presenting the recommendations and corresponding explanations to the user. Various communication mediums, such as Web-based systems [9, 4, 12] and mobile interfaces [3, 16, 14], have been studied for their effects on recommender systems. Moving beyond these conventional mediums, recent research has focused on robots as a promising alternative for recommendation delivery, particularly in scenarios where physical presence may improve effectiveness with enhanced user engagement.

Kamei et al. [9] experiment with robots recommending items based on customers' purchasing behaviors tracked by networked sensors in a shop. Using multiple robots, they show that participants lingered longer near shelves when robots interacted with them, often mirroring previous purchasing behaviors. Herse et al. [6] investigate embodiment's role in social robots' persuasiveness within a service setting. They conduct an experiment comparing a human, robot, and kiosk for restaurant recommendations. The results reveal that human-like embodiment enhances persuasiveness, though only with specific recommendation phrasing. They suggest human-like traits can boost recommendation impact, with a dependence on the choice of language. Sakai et al. [15] conduct an experiment with robots in a both virtual and embodied setting with visitors in conversations about food preferences before recommending dishes. While behavioral differences were minimal, the study found that physical robots notably improved satisfaction and agreement with recommendations, suggesting that embodiment has a positive impact on user engagement. Interestingly, in our setting, user engagement is higher when participants interact solely with a conventional tablet interface compared to when they interact with the robot and tablet together.

Embodiment research also takes place in education and negotiation. Çakan et al. [5] study the difference in negotiation styles, when interacting with virtual robots and physically embodied robots. Human participants took the negotiation more seriously against physically embodied robots and made more collaborative moves in the virtual robot setting. Survey responses indicate that participants perceived the robot as more human-like when it is physically embodied.

Köse et al. [11] investigate embodiment and gesture effects in child-robot interactions with a child-sized humanoid robot in an interactive drumming game. They studied three forms of communication: (i) direct interaction with an embodied system, (ii) a sound only, and (iii) a real-time virtual avatar. Through mixed-design measures, data from the experiments reveal that physical embodiment, especially with gestures, significantly enhances interaction quality, performance, and user enjoyment. Ultimately, a physically embodied robot with gesture capabilities, enhances the perceived intelligence of the agent and impoves engagement in human-computer interactions. In a similar set-up Keskin et al. [10] compared various forms of embodied negotiation opponents, to test the hypothesis that the appearance of the robots would change participants' impressions and attitudes. Ultimately, it did not alter the final session results.

In contrast, our study focuses on recommender systems. We examine how explainable recommender systems are experienced through different interfaces, namely a conventional tablet and a robot combined with a tablet, and how these interfaces affect user engagement, satisfaction, and overall system performance.

3 Research set-up and Hypotheses

In this study, we utilize the personalized explainable recommendation framework developed by Buzcu et al. [4], designed to support users in making food choices that meet both health requirements and personal preferences. The interaction between users and system is governed by the protocol illustrated in Figure 1. The interaction is initiated by a user request consisting of several constraints and the system provides an explanation along with the given recommendation. The user can provide feedback about recommended recipes and their corresponding explanations or end the interaction by either accepting or terminating the session without any agreement. With the given feedback, the recommendation strategy revises its recommendation, thus engaging in a subtle negotiation. This process lasts until they reach a termination condition.

The recommendation generation process begins with filtering. The system processes each user's dietary restrictions, allergies, and ingredient preferences and filters out items that don't align directly. For instance, vegan users will only receive plant-based recipes. This filtering is enabled by an RDF ontology-based database, which defines complex relationships between food entities and user dietary restrictions to refine the available choices to meet each group of users' requirements. The system evaluates the remaining items using a multi-criteria

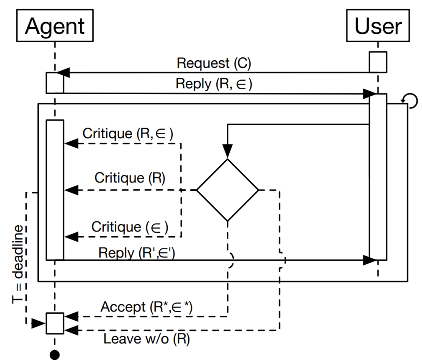


Fig. 1: FIPA description of the Negotiation Protocol where C corresponds to the user constraints, R is a recipe recommended by the agent, and ϵ is an explanation that comes with the recipe.

utility function, which rates each option based on (i) **Nutritional Value**, (ii) **Active Metabolic Rate (AMR)** and (iii) **User Satisfaction Scores**. Ultimately, these scores are combined via their respective weights to calculate an overall additive utility score for each item, which determines a heuristic ranking among the filtered options. The highest-ranked item that hasn't been recently recommended is then selected to be recommended. The system generates post-hoc explanations to clarify the reasoning behind the choice from the selected recommendation. This step includes two types of explanations:

– **Item & User-Based Explanations**: Using historical data, a decision tree is constructed to identify the key features that influenced the recommendation in a post-hoc manner.
– **Contrastive Explanations**: The system compares the chosen item with a similar alternative that didn't meet the decision criteria or was filtered otherwise. This contrast is used to show why the recommended item is the a more preferable option by highlighting features.

So far, Buzcu et al. [4] has implemented the Recommendation Framework with a Web Interface. Figure 2 shows the Web Interface as in the recommendation state. The Web Interface provides a medium for users to interact with the recommender system with the means of allowed actions in the protocol (see Figure 1). The Interface allows users to see recommended recipes in detail (e.g. name of the recipe, recipe ingredients, nutritional information) along with the explanation of the recommendation and give feedback about both recommendation and explanation. The recipe feedback section consists of options and the recipe ingredients. Users can pick the appropriate option (e.g. "I don't like...", "I ate the following recently...") with preferred ingredients. The explanation feedback can also be given by selecting the preferred option in the explanation feedback section. Users have three options to respond to the given recipe: (i) **Accepting the**

recommendation and ending the session, (ii) **Submitting the feedback and requesting new recommendation**,(iii) **Terminating the session**.

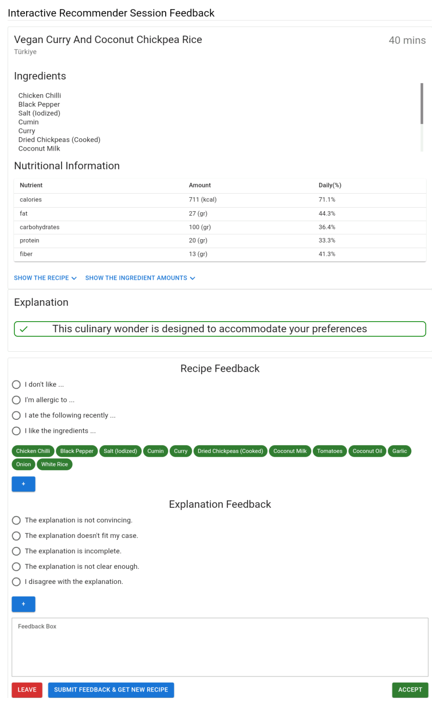


Fig. 2: Tablet Web Interface

Apart from the content of the explanations, how they are delivered also plays a crucial role in their effectiveness. Therefore, it is essential to investigate which communication medium/modality would establish more effective interaction with the user when building personalized explainable systems. This study mainly focuses on personalized explainable food recommendation systems and examines the effect of communication medium/modality (i.e., the effect of physical embodiment and textual/speech-based communication). Consequently, this study aims to investigate the following research hypothesis through user studies. For sake of readability, hereafter, tablet (touch-screen-based devices) only is referred to as a conventional interface.

- **Hypothesis-1 (H1)**: Incorporating a physically embodied robot into an explainable food recommendation system affects user engagement. *(Metric: Amount of User Feedback)*
- **Hypothesis-2 (H2)**: Incorporating a physically embodied robot into a conventional interface shifts the original recommendation perception and acceptance. *(Metric: Recommendation Acceptance Rate)*

- **Hypothesis-3 (H3)**: Interacting with a physically embodied robot with a conventional interface could result in different satisfaction levels with the personalization of recommendations compared to conventional interfaces. *(Metric: Qualitative User Feedback)*

Accordingly, we compare two settings in which the user interacts with the system via only a tablet (only visual and textual data is delivered) in one setting, whereas they interact with a QT robot via speech in addition to the use of a tablet. Note that we utilize the same strategies for recommendation and explanation generation in both settings. Ultimately, we aim to study user engagement, acceptability of the explanations, and communication effectiveness by comparing the interactions with the user when the explanations are delivered by a tablet-based medium or a physical humanoid robot with a tablet interface.

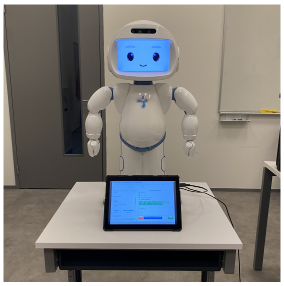
4 Experimental Evaluation

In this section, we first introduce our Methodology (Section 4.1) and discuss our results with the experimental settings (Section 4.2, and Section 4.3).

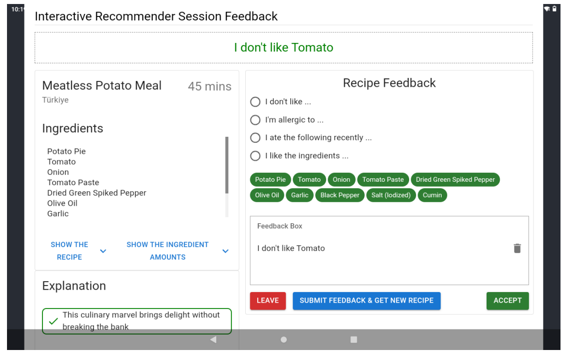
4.1 Methodology

In order to test our hypotheses, we utilize a QT robot[6] supporting various interaction services such as text-to-speech, speech-to-text, emotions display and gestures. Human-robot interaction requires speech communication, emotions and gestures to capture the users' attention and create a more human-like interaction. On the other hand, grasping the provided material within the dialogue and reasoning them may require some visual deliverables. That is, if we provide the same context in the form of visual and textual, it might be easier for a human user to evaluate. While determining what content should be delivered in textual format via tablet and what content to be delivered by the robot via speech, we conducted a pilot user study where the users give feedback about the effectiveness of communication. We observed that users had difficulty in understanding the details of the given recipe or keeping in mind the related details affecting their decisions. We found out that utilizing speech-based dialogue for short explanations and structured user feedback creates a rapport between system and users. Therefore, we employ the speech-base interaction for the following tasks:

- Greeting the user (e.g. "Hello! My name is QT. I am a nutritionist! I am here to help you about your food selection! Can you hear me?")
- Providing the name of the recipe and its corresponding explanations verbally (e.g. "Great! I can hear you too! Now I recommend you Meatless Potato Meal. Because I think this culinary marvel brings delight as an affordable option. Did you like this recipe?").
- Receiving structured feedback (e.g. "I am curious to hear your thoughts! Would you like to give me a feedback before we continue?").



(a)



(b)

Fig. 3: Robot with Tablet Session Setup and Tablet Interface

On the other hand, the content of the full recipe is displayed on the tablet screen as shown in Figure 3b so that the users can asses the recommendation. Moreover, the robot also uses gestures and expresses feelings, emotions via facial expressions alongside the recommendations to add human-likeliness to the system for a more fluent experience. Furthermore, the user can track the interaction states from the tablet interface by displaying the speech recognition outputs simultaneously and options for structured explanations. It is worth noting that both settings (Only Tablet versus Robot and Tablet) provide the same content to the users. The only difference is that some of the visible content changes with the interaction state in the robot setting, while all information is displayed on the tablet in the only tablet setting.

4.2 Experiment Setup

The primary objective of our experiment is to analyze the effects of different communication channels on the efficiency of the recommender system. As described above, we have two communication settings of our food recommender: *QT Robot with tablet* and *Only Tablet*. Therefore, our study comprises two subject groups; one only experienced the tablet, whereas the other experienced the robot with the tablet in order to get a food recipe recommendation along with its explanations. We employ between-subject design for our experimental setup to reduce (i) learning effect, (ii) cognitive overload, and (iii) comparison bias so that each participant interacted with only one setting. During our experiment, each participant goes through the steps depicted by Figure 4.

1. **Pre-Experiment Survey**: Participants fill out a survey to provide demographic information and their initial perceptions of robots and technology. This survey includes questions about age, gender, education level, and familiarity with technology.

[6] https://luxai.com/humanoid-social-robot-for-research-and-teaching/

2. **Experiment Sessions**: Participants (i) are informed about the underlying setting through a brief explanatory video, and (ii) practice with a short demo version to familiarize how to interact with the system effectively. Then they start the experiment choose a food preference profile from a pool of predefined profiles (e.g. vegetarian, fast-food lover, sports). According to the chosen profile, the system makes recommendations along with their explanations. Note that the content of the recommendations are generated in the same way for the same profile. Only the communication channels are different for different subject groups.
3. **Post-Experiment Survey**: After their interaction with the system, they fill out post-experiment survey regarding the quality of the recommendations and explanations as well as the likability of the interaction.
4. **Informal Interviews**: At the end of the experiment, participants are interviewed informally to gather qualitative data on their experiences.

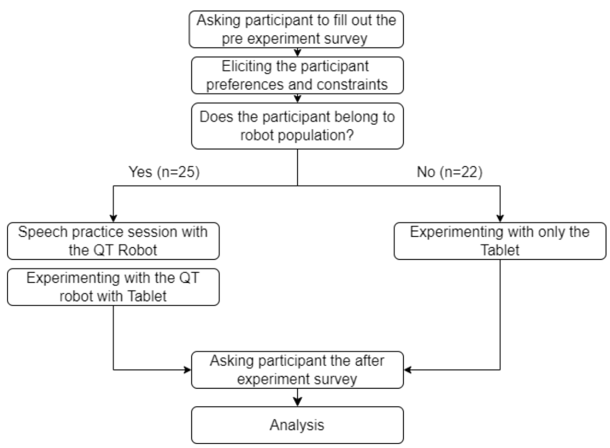
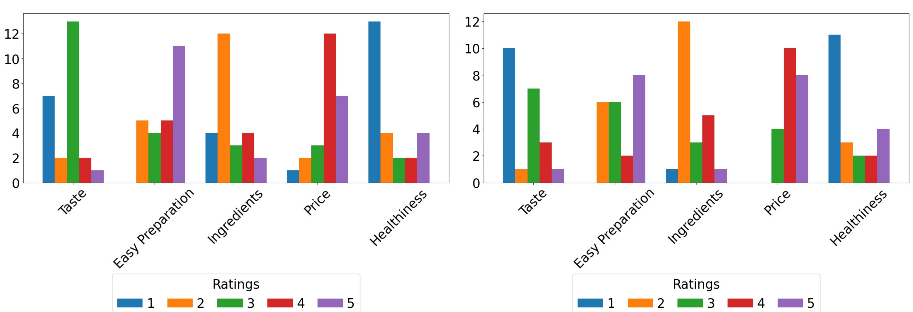


Fig. 4: Experiment Procedure

In the experiments, we used a 5-point Likert scale to measure participants' responses. After collecting the data, we assessed the distribution to determine the appropriate statistical tests for analysis. Using the Kolmogorov-Smirnov Test of Normality, we found that our data did not follow a normal distribution. This lack of normality precludes the use of parametric tests, which generally assume a normal distribution of data. The previous lead to select Mann-Whitney U Test, a non-parametric alternative suitable for comparing differences between two independent groups without assuming normal distribution.

The study was conducted at Özyeğin University and involved 59 participants, including students and employees. Participation was voluntary, with optional credits for social sciences students. In order to validate participants' attention, we included a question inside the pre-experiment survey that asks, "If you are paying attention, please select 2.”. 9 participants failed this test. Moreover, 3 participants had to end the experiment due to personal circumstances. Thus, 47 participants were evaluated for the experiment. Participants were divided into two groups. The Robot with Tablet group consisted of 25 participants: 15 male, 10 female; 20 bachelor’s, 4 master’s, and 1 PhD student; 9 aged 18-21, 12 aged 22-25, 3 aged 26-30, and 1 aged 36-40; 13 from engineering, 2 from mathematics, and 10 from social sciences. The Only Tablet group included 22 participants: 13 male, 9 female; 16 bachelor’s, 5 master’s, and 1 PhD student; 7 aged 18-21, 13 aged 22-25, and 2 aged 26-30; 13 from engineering, 1 from mathematics, and 8 from social sciences. The university ethics committee approved the study[7].

Finally, Figure 5 presents a histogram analysis of the questionnaire results, where participants rated factors on a scale from 1 to 5, with 1 indicating the highest importance. The analysis reveals that, in both groups, the healthiness of a recipe is considered to be the most vital factor in choosing it (52% for robot with tablet group and 50% for tablet only group). In contrast, 44% of participants within the Robot with Tablet group rated ease of preparation as the least important factor and 36% of the participants voted equally Price and Easy Preparation to be the least important factor. The users’ priority on the decision criteria for choosing different recipes are similar for both groups.



(a) Robot with Tablet (b) Tablet Only

Fig. 5: Histogram analysis of the pre-experiment survey for each group

4.3 Results and Discussions

Within the analysis of self-explanatory systems, success is commonly measured through two categories of metrics: objective and subjective [8]. On one hand, objective metrics are derived from participants’ actions during their interactions

[7] Participant data is anonymized and securely stored. They can withdraw anytime without consequences. The study follows ethical guidelines to ensure no physical or psychological harm to participants.

with the system, including measures like success rate (e.g., percentage of accepted recipes), number of interaction rounds per session, or feedback during sessions. Subjective metrics, on the other hand, are gathered from qualitative data such as the post-experiment surveys (see Figure 8 below), where participants rate aspects such as perceived effectiveness, satisfaction, and ease of use. These subjective ratings assess whether the incorporating of the robot affects the user satisfaction with the recommendations, explanations and the general working of the system.

First and foremost, we consider the users' feedback to the system's recommendation. Recall that the feedback is comprised of any verbal (Robot with tablet session) or non-verbal (Only tablet session) user response to a given recommendation, which we take as an indicator of engagement levels across groups. Meanwhile it is difficult to compare the quality of feedback, we can still consider the amount of feedback where the higher number means the user is engaged more with the system since they actively partake. This gives us a quantifiable measure of user engagement while testing **H1**. Figure 6 shows the boxplot of the feedback counts per group. We note that there is a significant statistical difference between the groups ($p = 0.04 < 0.05$) where the tablet group provided more feedback to the system. We observe that the users engaged more in the Tablet Only group in contrast the Robot with Tablet group, thus satisfying the **H1** (on average, 14.72 vs 6.52 respectively).

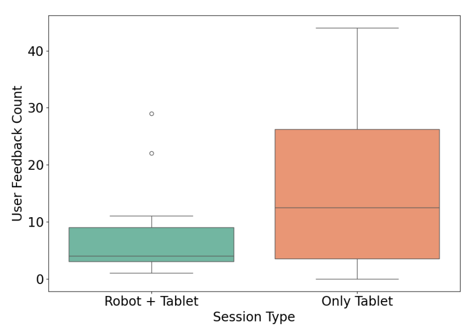


Fig. 6: Amount of feedback per type of session

On the other hand, higher acceptance counts could indicate that users find the recommendations more convincing, so they might be more willing to adopt them. Acceptance counts are defined as the number of times users explicitly agree with a food recipe. Ultimately, the comparative study between the groups would allow us to assess whether physical embodiment affects users' acceptance of the system's recommendations. Our results indicate that the users accepted 68% of the recommendations where as tablet acceptance was 95%. Consequently, we note that this supports the **H2** as we found a significant difference among the two groups. The acceptance rate and feedback counts could be related to the fact that robots are scrutinized by humans since embodiment brings additional expectations from the system [1]. This could also signal a cognitive load (i.e., may get tired from interacting with the robot)and more tendency to terminate the interaction.

Finally, we consider the number of turns between the groups, as shown in Figure 7. Here, we define turns as each time a user responds to a recommendation. As a side observation, we noticed a minor difference between the groups (p=0.72, 6.76 vs 5.94 interactions on average). This may suggest different engagement levels, potentially influenced by user expectations and the mentioned cognitive load during interactions.

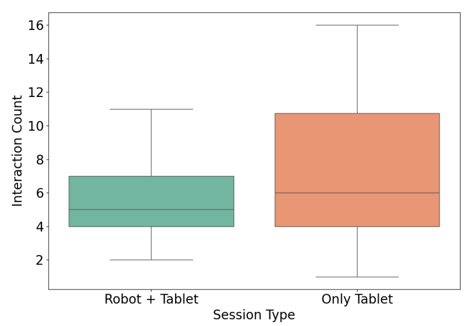
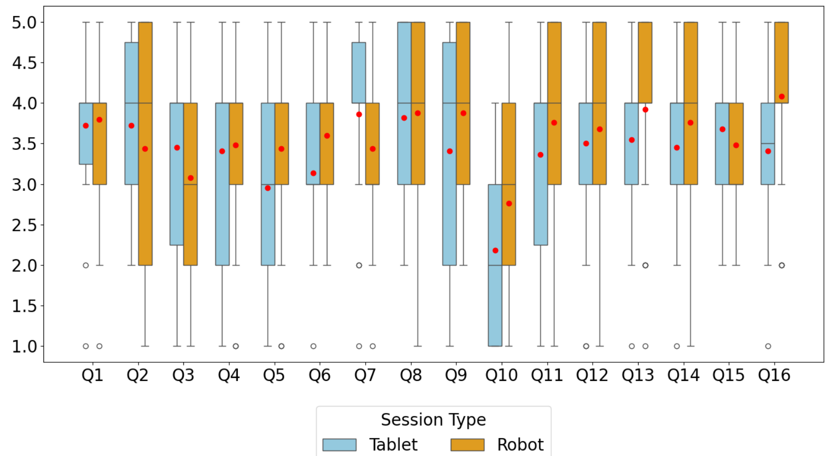


Fig. 7: Amount of interaction per type of session

Figure 8 illustrates the results of our post-experiment survey questions with regards to our subjective metrics. We link some of the questions to **H3** (Q1, Q11, Q12, Q13, Q14, and Q16) that assess various aspects of recommendation satisfaction and personalization between the groups. These questions provide insights into participants' perceived relevance and accuracy of the recommendations, which reflects the user satisfaction with the system's communication method when compared between the groups. Q1 (p=0.69, on average, 3.73 vs 3.80) gauges initial liking for the recommended items, setting a foundation for perceived quality, Q11 (p=0.26, 3.36 for the Tablet Only vs 3.76 for the Robot with Tablet) assesses alignment with the scenario, highlighting satisfaction with situational appropriateness. Q12 (p=0.54, 3.5 vs 3.68) and Q13 (p=0.20, 3.54 vs 3.92) measure participants' perceptions of how accurately and personally the system recognized and incorporated their individual preferences. Finally, Q14 (p=0.24, 3.46 vs 3.76) and Q16 (**p=0.03**, 3.41 vs 4.08) focus on the perceived effectiveness of the personalization process, specifically regarding dietary preferences and constraints. Ultimately, **H3** is only partially supported since we found a significant difference for Q16.

Besides the mentioned questions, we observe no significant difference among the questions. However, we would like to note that there was a higher average for the Robot with Tablet group among Q4 (p=0.92, 3.41 vs 3.48), Q5 (p=0.16,



Q1 I liked the recommended recipes.
Q2 I enjoyed using the recommender system.
Q3 Given interface was comfortable to navigate.
Q4 I liked the explanations that were given.
Q5 I find the given explanations convincing.
Q6 I find the given explanations intuitive.
Q7 The interaction with the recommender was fluent.
Q8 I enjoyed the interaction with the recommender.
Q9 I would like to engage in such an interaction in the future.
Q10 The interaction with the recommender was frustrating for me.
Q11 The recommendations were right for the given scenario.
Q12 Rate the food recommendation system's ability to recognize your food preferences accurately.
Q13 Recommendations were personalized according to my preferences.
Q14 Rate the effectiveness of the food recommendation system in personalizing food recommendations based on your profile.
Q15 What is your opinion about the duration of the experiment?
Q16 How well did the food recommendation system address your specific dietary preferences and constraints?

Fig. 8: Post-Experiment Survey Results for Only Tablet and Robot with Tablet

2.95 vs 3.44), Q6 (p=0.09, 3.13 vs 3.60). These results could signal that the explanations had a slightly higher impact among the participants when they are delivered in an embodied manner. On the other hand, users reported higher scores in Q2 (p=0.66, 3.73 vs 3.44) and Q3 (p=0.27, 3.45 vs 3.08). The frustration with the interaction measured by Q10 (p=0.12, 2.18 vs 2.76) and higher average for the robot supports our finding with the objective metrics. This indicates that either system was not perceived particularly more interesting than the other.

During our unstructured interviews, it became clear that the design of the state machine communication in the robot with tablet configuration may cause impatience for some participants. This aligned with the survey results for Q10 (Figure 8) where 7 out of 25 participants gave a score of more than 3. The slower pace of interaction combined with occasional speech recognition issues made the session seem less fluent compared to the tablet-only configuration. It was observed that the simple nature of the tablet-only setup made it easy for users to provide feedback, which explains why engagement remained high. In addition, some participants struggled to engage with the system due to their varying levels of English, which may have impacted their ability to fully engage. These challenges highlight the need for more user-friendly communication strategies and better support for non-native speakers.

5 Conclusion and Future Work

This study investigated the effects of different communication mediums, a tablet-only interface, and a robot-enhanced setup on the user experience within a personalized food recommendation system. Differences were observed in the frequency of user engagement between the two setups, indicating that the choice of communication medium can significantly impact how actively users participate in interactions. Acceptance levels also varied across the setups, suggesting that communication medium influences users' willingness to adopt the system's recommendations. On the other hand, we looked at how users perceived the system's alignment with their preferences and needs, focusing on their satisfaction with the interaction. The robot-enhanced setup affected user satisfaction minimally. Satisfaction remained largely consistent across both setups, suggesting that physical embodiment alone did not significantly enhance users' contentment or sense of alignment with the system's functionality. These findings emphasize the importance of selecting suitable communication mediums to optimize user experience and perceived effectiveness in recommendation systems. Future designs may benefit from simplifying communication flows in embodied setups to minimize user impatience, enhancing response fluency, and implementing strategies that address language barriers. These improvements could further increase accessibility and engagement across diverse user backgrounds, supporting the development of more intuitive and responsive interactive systems.

6 Acknowledgments

This work has been supported by the grant CHIST-ERA-19-XAI-005, and by the Swiss National Science Foundation (G.A. 20CH21_195530), the Luxembourg National Research Fund (G.A. INTER/CHIST/19/14589586), and the Scientific and Research Council of Turkey (TÜBİTAK, G.A. 120N680).

Bibliography

[1] Muneeb Imtiaz Ahmad, Jasmin Bernotat, Katrin Solveig Lohan, and Friederike Eyssel. Trust and cognitive load during human-robot interaction. *ArXiv*, abs/1909.05160, 2019. URL https://api.semanticscholar.org/CorpusID:202558882.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012.

[3] Berk Buzcu, Yvan Pannatier, Reyhan Aydoğan, Michael Ignaz Schumacher, Jean Paul Calbimonte, and Davide Calvaresi. A framework for explainable multi-purpose virtual assistants: A nutrition-focused case study. In Davide Calvaresi, Amro Najjar, Andrea Omicini, Rachele Carli, Giovanni Ciatto, Reyhan Aydogan, Joris Hulstijn, and Kary Främling, editors, *Explainable and Transparent AI and Multi-Agent Systems - 6th International Workshop, EXTRAAMAS 2024, Revised Selected Papers*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 58–78, United States, 2024. Springer.

[4] Berk Buzcu, Melissa Tessa, Igor Tchappi, Amro Najjar, Joris Hulstijn, Davide Calvaresi, and Reyhan Aydogan. Towards interactive explanation-based nutrition virtual coaching systems. *Autonomous Agents and Multi-Agent Systems*, 38(5), 2024. https://doi.org/10.1007/s10458-023-09634-5.

[5] Umut Çakan, M. Onur Keskin, and Reyhan Aydoğan. Effects of agent's embodiment in human-agent negotiations. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, IVA '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399944. https://doi.org/10.1145/3570945.3607362. URL https://doi.org/10.1145/3570945.3607362.

[6] Sarita Herse, Jonathan Vitale, Daniel Ebrahimian, Meg Tonkin, Suman Ojha, Sidra Sidra, Benjamin Johnston, Sophie Phillips, Siva Leela Krishna Chand Gudi, Jesse Clark, William Judge, and Mary-Anne Williams. Bon appetit! robot persuasion for food recommendation. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 125–126, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356152. https://doi.org/10.1145/3173386.3177028.

[7] Sarita Herse, Jonathan Vitale, Meg Tonkin, Daniel Ebrahimian, Suman Ojha, Benjamin Johnston, William Judge, and Mary–Anne Williams. Do you trust me, blindly? factors influencing trust towards a robot recommender system. In *2018 27th IEEE International Symposium on Robot*

and Human Interactive Communication (RO-MAN), pages 7–14, 2018. https://doi.org/10.1109/ROMAN.2018.8525581.

[8] Joris Hulstijn, Igor Tchappi, Amro Najjar, and Reyhan Aydoğan. Metrics for evaluating explainable recommender systems. In *AAMAS, EXTRAAMAS 2023, London, England, May 29, 2023*. Springer, 2023.

[9] Koji Kamei, Kazuhiko Shinozawa, Tetsushi Ikeda, Akira Utsumi, Takahiro Miyashita, and Norihiro Hagita. Recommendation from robots in a real-world retail shop. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450304146. https://doi.org/10.1145/1891903.1891929. URL https://doi.org/10.1145/1891903.1891929.

[10] M. Onur Keskin, Selen Akay, Ayse Dogan, Berkecan Koçyigit, Junko Kanero, and Reyhan Aydogan. You look nice, but i am here to negotiate: The influence of robot appearance on negotiation dynamics. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 598–602, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703232. https://doi.org/10.1145/3610978.3640759.

[11] Hatice Köse, Ester Ferrari, Kerstin Dautenhahn, Dag Sverre Syrdal, and Chrystopher Nehaniv. Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*, 23: 1951–1996, 10 2009. https://doi.org/10.1163/016918609X12518783330360.

[12] Dennis Lawo, Thomas Neifer, Margarita Esau, and Gunnar Stevens. Buying the 'right' thing: Designing food recommender systems with critical consumers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[13] Alessandra Rossi and Silvia Rossi. Engaged by a bartender robot: Recommendation and personalisation in human-robot interaction. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '21, page 115–119, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383677. https://doi.org/10.1145/3450614.3463423. URL https://doi.org/10.1145/3450614.3463423.

[14] Christine Rzepka and Benedikt Berger. User interaction with ai-enabled systems: A systematic review of is research. 12 2018.

[15] Kazuki Sakai, Yutaka Nakamura, Yuichiro Yoshikawa, and Hiroshi Ishiguro. Effect of robot embodiment on satisfaction with recommendations in shopping malls. *IEEE Robotics and Automation Letters*, 7(1):366–372, 2022. https://doi.org/10.1109/LRA.2021.3128233.

[16] Wan-Shiou Yang, Hung-Chi Cheng, and Jia-Ben Dia. A location-aware recommender system for mobile shopping environments. *Expert Systems with Applications*, 34:437–445, 01 2008. https://doi.org/10.1016/j.eswa.2006.09.033.