Why or Why Not? The Effect of Justification Styles on Chatbot Recommendations

DARICIA WILKINSON, Clemson University, USA
ÖZNUR ALKAN, IBM Research, Ireland
Q. VERA LIAO, IBM Research, USA
MASSIMILIANO MATTETTI and INGE VEJSBJERG, IBM Research, Ireland
BART P. KNIJNENBURG, Clemson University, USA
ELIZABETH DALY, IBM Research, Ireland

Chatbots or conversational recommenders have gained increasing popularity as a new paradigm for Recommender Systems (RS). Prior work on RS showed that providing explanations can improve transparency and trust, which are critical for the adoption of RS. Their interactive and engaging nature makes conversational recommenders a natural platform to not only provide recommendations but also justify the recommendations through explanations. The recent surge of interest inexplainable AI enables diverse styles of justification, and also invites questions on how styles of justification impact user perception. In this article, we explore the effect of "why" justifications and "why not" justifications on users' perceptions of explainability and trust. We developed and tested a movie-recommendation chatbot that provides users with different types of justifications for the recommended items. Our online experiment (n = 310) demonstrates that the "why" justifications (but not the "why not" justifications) have a significant impact on users' perception of the conversational recommender. Particularly, "why" justifications increase users' perception of system transparency, which impacts perceived control, trusting beliefs and in turn influences users' willingness to depend on the system's advice. Finally, we discuss the design implications for decision-assisting chatbots.

CCS Concepts: • **Human-centered computing → User studies**; **Empirical studies in interaction design;**

Additional Key Words and Phrases: Conversational agent, chatbots, explanation, trust, human computer interaction, user study, user interface

ACM Reference format:
Daricia Wilkinson, Öznur Alkan, Q. Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P. Knijnenburg, and Elizabeth Daly. 2021. Why or Why Not? The Effect of Justification Styles on Chatbot Recommendations. *ACM Trans. Inf. Syst.* 39, 4, Article 42 (October 2021), 21 pages.
https://doi.org/10.1145/3441715

Authors' addresses: D. Wilkinson and B. P. Knijnenburg, Clemson University, 821 McMillan Rd, Clemson, SC, USA; emails: {dariciw, bartk}@clemson.edu; Ö. Alkan, M. Mattetti, I. Vejsbjerg, and E. Daly, IBM Research, IBM Dublin Technology Campus (Building 3), Dublin, Ireland; emails: OAlkan2@ie.ibm.com, massimiliano.mattetti@ibm.com, {ingevejs, elizabeth.daly}@ie.ibm.com; Q. V. Liao, IBM Research, 1101 Kitchawan Road, Yorktown Heights, NY, USA; email: Vera.Liao@ibm.com.

1046-8188/2021/10-ART42 $15.00
https://doi.org/10.1145/3441715

1 INTRODUCTION

Recently, chatbots have gained popularity. Among many other benefits, chatbot technologies are increasingly adopted by companies and organizations to reduce cost and time spent with customer service tasks. Globally, the chatbot market was valued at USD 1.17 billion in 2018 and by 2026 it is expected to reach USD 10.08 billion [47].

Chatbots offer a paradigm change from traditional interaction methods used in **Graphic User Interface (GUI)** of websites or other channels (such as calling a help desk) [29]. As users constantly face complex situations where finding information or making a decision is difficult, chatbots offer an efficient exchange of useful information to quickly resolve a complaint, find items to purchase, or receive recommendations for consumption [10].

Chatbots are powered by "conversational recommenders" that can provide personalized answers or recommendations. In this article, we address the issue that existing conversational recommenders often do not offer insights into why certain items were selected for the user. Providing explanations is key to users' understanding of actions made by a system, which has been underlined by the surging field of **explainable AI (xAI)**. One popular type of explanation, the *justification*, offers insight as to why a decision is a good one without necessarily describing the algorithmic details. Our work focuses on justifications as they may be particularly important for users of **Recommender Systems (RS)** to understand and accept the recommended items. Indeed, in traditional RS, researchers have found that justifications could support users' decision-making process, increases the perceived transparency of, and inspires users' trust in the system [5, 13, 55].

Although trust is multifaceted [31, 39, 56], most of the existing literature in xAI has focused on the influence of *overall trust* [7, 15, 41]. However, investigating the effects of the different facets of trusting beliefs independently would have theoretical and practical benefits. Moreover, conversational recommenders enable interactive and arguably social properties in explanations that may not present in traditional explanatory components in RS. Against this background, we conducted a user experiment in which we let participants interact with a conversational recommender that provided justifications for the recommended items. Controlling for algorithmic accuracy, we explored justifications explaining *why* an item was recommended or *why not*. As such, the objective of the study was to answer the following research questions:

RQ1. How do different justification styles affect users' perception and user experience with the system?

RQ2. Does algorithmic accuracy have an effect on the interaction with the justifications?

RQ3. What type of personal characteristics influence users' interaction with the respective justification styles?

By utilizing structural equation modeling, we were able to contribute empirical evidence of the relation between algorithmic accuracy, justification style and user trust. Specifically, the contributions of this article can be summarized as follows:

- We propose a method to generate *why not* justifications, which utilizes a system-controlled selection for the alternative item X to generate justification using "why not X?" questions
- We conducted a user study that explored and confirmed the influence of algorithmic accuracy and justification style on users' trust and user experience in a conversational recommender, particularly when users are given *why* justifications.
- We extend existing Human Computer Interaction theory [53] on the effect of justifications on trusting beliefs and system adoption intentions within conversational recommendation systems.

In the following section, we provide background into our work by reviewing existing literature and offer theoretical support for the framing of the concepts the work is centered around. We

then describe our methodology followed by a presentation of the results of our online study. We conclude with a discussion of the implications for design and proposals for future research.

2 RELATED WORK

Prior research relevant to our work is collated in two sub-sections. First, we present the related work around conversational recommender systems. Second, we review the research on explanations in recommender systems and other technologies. The latter part of this section highlights the differences to previous work and our hypotheses for the study.

2.1 Conversational Recommender Systems

Prior work has shown that conversational interactions have significantly contributed to establishing long-term rapport and trust with systems [8]. As such, conversational recommendation systems have been attracting attention with increasing levels of engagement due the potential benefits from dynamic conversational interactions with decision-making systems. Yan et al. [57] propose a conversational dialogue recommender framework in a mobile online shopping application. Besides being used in different modalities, researchers have also looked into implementing decision-assisting capabilities. Christakopoulou et al. [12] present a system that improves the recommendation results based on users' repeated input on whether they like an item or prefer one item to another. Similarly, Dodge et al. [14] discuss an end-to-end dialogue agent that can recommend movies based on context information using a memory network. Moreover, Zhang et al. [58] present a conversational search solution and a recommender system; their framework can select facets to ask the user about and recommend a list of items accordingly. There are also existing works around critiquing-based conversational recommendation, where users provide feedback on multiple cycles of recommendations [37, 38].

In critiquing-based conversational recommendation, interactions with users are performed mainly through graphical user interfaces that are presented on different platforms, such as mobile mockups or webpages, rather than through natural language dialogues. In contrast, we are particularly interested in exploring natural language dialogues through the conversational nature of chatbots. Rather than point and click interactions, we leverage the social benefits of conversational agents to offer justifications throughout the decision-making process.

2.2 Justifying Recommendations

Establishing trust is critical to the adoption of recommended items [6]. Explanations and justifications allow users to better understand and interpret the rationale of the recommender systems, which can lead to improved trust, transparency and user engagement [17, 24, 40, 45, 49, 53]. Kouki et al. [32] present a hybrid recommender system that is built on a probabilistic programming language, and they demonstrate that explanations improve the user experience of the recommender system. Friedrich et al. describe a taxonomy of explanation approaches, taking into account different dimensions like the style (e.g., collaborative, knowledge, utility or social explanation style), paradigm (e.g., content-based, knowledge or collaborative based) and the type of preference model [19]. In Reference [52], authors create explanations through capturing the interactions between users and their favorite features by constructing a feature profile for the users. Moreover, they use a feature-weighting scheme to reveal those features that better describe a user and those that better distinguish that user from the others. In addition, different visualization techniques are proposed for providing explanations for the generated recommendations, such as interfaces with concentric circles [26, 44], and pathways between columns [9]. Dominguez et al. [15] experiment with different interfaces with different levels of explainability and different algorithms for artistic image recommendation.

In this study, we consider the justification style in terms of the type of user *question* it answers. Beyond RS, explanations have been studied for broader AI and ML technologies, most notably in the field of xAI [22, 23, 42] as well as expert and systems [46, 48], adaptive agents [21], and context-aware technologies [36]. By conducting an extensive review of social science literature on how humans produce and consume explanations, Miller defines an explanation as "an answer to a why-question" and highlights the *contrastive* nature of most explanation demands. That is, a *why* question is often an implicit *why-not* question that refers to an alternative outcome *not* given by the system. This insight reveals a common gap in algorithmic explanations that often focuses on the *why* question only. For example, Lim et al. [34] studied what questions users may ask context-aware intelligent systems. They developed a toolkit that supports the generation of different styles of explanation for context-aware systems [35]. In an experimental study, Lim et al. compared the effect of explanation designs addressing four types of intelligibility (why, why not, how to, and what if), and found that for novice users, the Why explanation was preferred for its simplicity [36]. Moreover, Wang and Benbasat evaluated the use of three types of explanations in a recommender-*how*, *why*, and *tradeoff*- and found that the explanations impacted different facets of trust, namely, competence, benevolence, and integrity [55]. Similarly, by adapting McKnight's framework on trust [39], Kunket et al. found that different sources of recommendations influence users' trusting beliefs and trusting intention [33]. Likewise, Berkovsky et al. conducted a cross-cultural user study and found significant differences in explanation preference based on participants' perception of integrity, competence, and transparency [6]. To the best of our knowledge, our work is the first to explore how the different levels of trust are impacted when presenting justifications within conversational recommender systems.

2.3 Research Aims

The social aspects of conversational recommender systems may be beneficial in situations where the interaction needs to convey to users that they can depend on or trust the system to behave it their best interest. This is particularly important when offering insights into the recommendation process as end-users might be sometimes confused about why a system behaved a certain way or why it did not behave a certain way. Our work departs from prior work by using natural language to present justifications for recommended items rather than using web-based GUIs. In this way, we replicate the interactions of state-of-the-art chatbots that are commercially available. Our goal is to better understand how different approaches to justifying the behavior of the recommender can impact trust and potential adoption behaviors. As such, we conduct an online user experiment exploring the impact of *why* ("Why did the system do X?") and *why not* ("Why did the system not do X?") justifications, and algorithmic accuracy on users' trust and user experience.

3 HYPOTHESES

Several previous studies considered justifications for recommender systems, and they showed that justifications could increase the quality of user experience and interactions with a system [24, 53]. We assume that *justification efficiency* (i.e., to help users to quickly assess recommendations) and *justification effectiveness* (i.e., to help users in correctly determining the quality and suitability of recommendations) would be used to measure **User Perceived Quality (UPQ)**. As such, we hypothesize that UPQ will be higher for the manipulated conditions (why, why not, combination) compared to having no justification (*H1a*).

Further, compared to "Why Not" justifications, the "Why" justifications support users in making a clearer connection between the causes of the system's behavior. This may influence users' trusting beliefs and perceived transparency of the system [39]. As such, we expect that *why* justifications should improve users' perception of trusting beliefs such as integrity and transparency compared to *why not* styles (*H1b*). However, the combination of *Why and Why Not* justifications

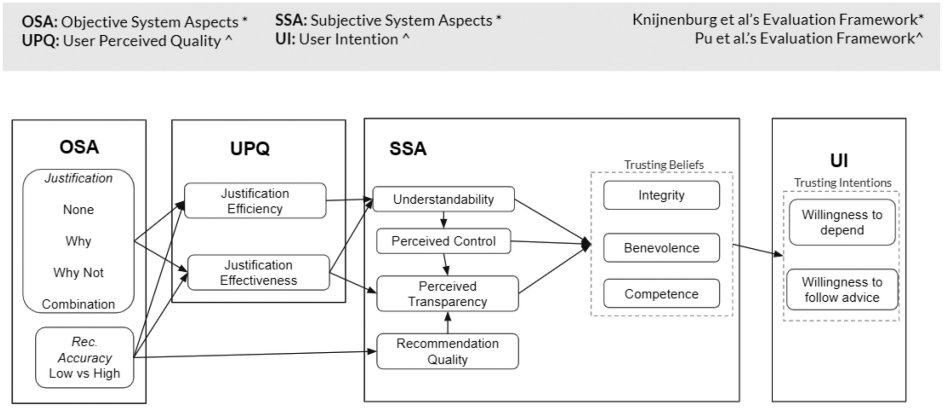


Fig. 1. Theorized model of the study.

can offer information about why an item was selected in addition to justifying why that was selected instead of another option. As such, we consider McKnight's framework on the different facets of trust and hypothesize that providing both types of justifications will improve users' perception of the system's competence (*H1c*).

Last, our study manipulates the accuracy of the algorithm (low versus high). High accuracy should result in recommendations that are better aligned with users' preferences and will therefore likely improve perceived recommendation quality (*H2*). As such, this setup allows us to test whether justifications are robust to fluctuations in recommendation quality.

To test these hypotheses, we created an interactive chatbot that provides different types of justifications with different levels of algorithmic accuracy. Our work is tested within existing recommender systems evaluation frameworks [30, 45] (see Figure 1). Next, we describe the testing infrastructure.

4 EXPERIMENTAL SETUP

For the study, we developed a movie recommending chatbot and employed a 4×2 between-subjects experimental design. First, we introduce the dataset chosen for the purpose of this study. Second, we describe the algorithms chosen. Third, we present an overview of the design choices for the interface and rationale behind the justification styles. Finally, the user study procedure is explained.

4.1 Data

We used the movie dataset available from Kaggle.[1] The dataset consists of movies released on or before July 2017. Data points include features such as cast, crew, plot keywords, budget, revenue, and so on. For our study, we used cast, director and genre features for both building the recommender algorithm and for generating the justifications. The content features for the movies are formed by processing the keywords, cast, and crew fields and they are treated equally as tags by the content-based recommender algorithm that is detailed in Section 4.2. The dataset contains *46,628* movies where each movie contains at most *8* features, and *5.6* features per movie on average.

4.2 Recommendation Techniques

The main motivation for the article is to evaluate different forms of explanations on the users' perception of explanability and trust. To achieve this, we used a content-based filtering algorithm

[1] https://www.kaggle.com/rounakbanik/the-movies-dataset.

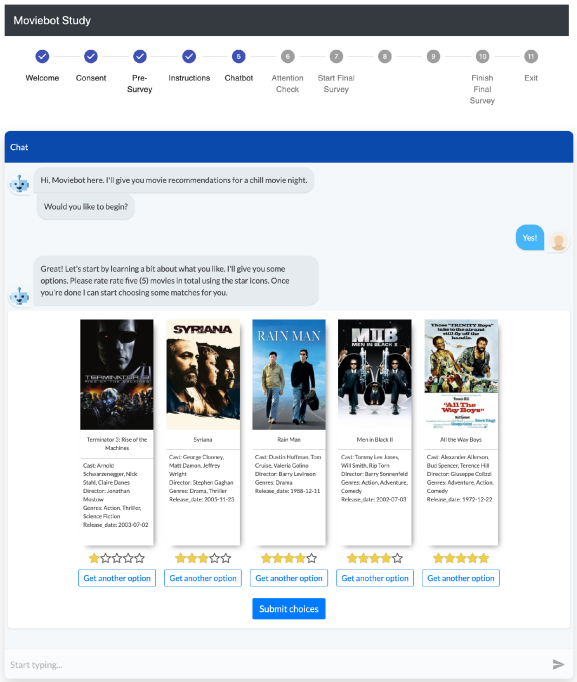


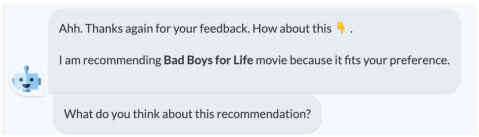
Fig. 2. Example for preference elicitation step.

for building the recommendation system [1]. The algorithm initially builds movie profiles in an offline manner, where each movie profile is represented by a vector of weights of all features of the domain. The features include all distinct *genres*, *actors*, and *directors* of the movies in the dataset. For each active user, a weighted mean of the rated item vectors constitutes the user's profile, where weights correspond to the ratings that the user has provided for the items during the preference elicitation step. The ratings are provided based on a *1–5* scale, where we consider *3* as neutral. During score calculation, we subtract *3* from users' ratings to be able to reflect the negativity and positivity of the user's preferences on the final recommendation score calculation.

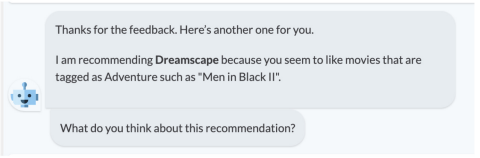
We implemented two levels of accuracy that the recommender algorithm employs while calculating the alignment between the user profiles and item vectors. For both versions, whenever a request is received to generate recommendations for a target user, the *cosine similarities* between user's profile vector and all candidate movie profile vectors are calculated. Candidate movies are all the movies the user has not rated yet.

The *High* accuracy recommendations are the typical *top-n* movies with the highest similarity scores. This condition represents the standard behaviour of recommender algorithms. The *Low* accuracy recommendations, however, are the *bottom-n* positively scored movies. Note that by selecting positively scored movies, the recommended items are prevented from including *bad* items—the recommendations still have a predicted score higher than 3/5 stars. For both techniques, the recommender algorithm selects movies within the *top-n* or *bottom-n* such that the list of recommendations covers all different regions of the user profile space.

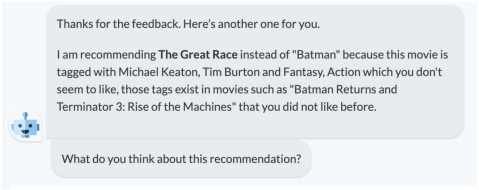
It is important to emphasize that the proposed explanation technique reveals information around the alignment between the user's interests and recommended items' features. Therefore the conclusions we reach can be extended to other recommender algorithms that present similar explanations to their generated suggestions.



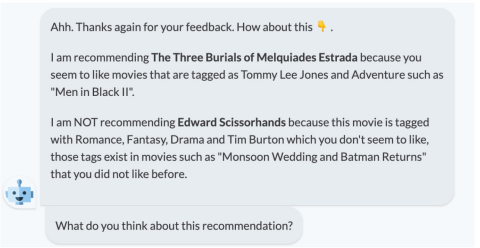
(a) None (no justification)



(b) Why justification



(c) Why not justification



(d) Combination justification

Fig. 3. Examples of our four different variations of justifications.

4.3 Preference Elicitation

We implemented a system-controlled preference elicitation mechanism to draw out users' movie preference information for generating recommendations. We presented movies in decreasing order of popularity (i.e., total number of ratings) and asked users to either provide a 1- to 5-star rating or ask for different movie. This process continues until users have rated *five* movies, after which the system proceeds to the recommendation part. A screenshot of the preference elicitation process is shown in Figure 2.

4.4 Justification Generation

We generated four types of justifications: *None*, *Why*, *Why not*, and *Combination*. In the *None* condition, the system recommends movies without any justification, and we present the same text for all users and for all recommended items, as shown in Figure 3(a).

ALGORITHM 1: Generating a Why justification

```
1:  procedure GenerateWhyExp(user, item, model)
2:      itemFL ← model.getFeatList(item)
3:      userFPL ← model.getUserFeatPos(user)
4:      sort userFPL in the order of desc. feature scores
5:      commonFeat ← {}
6:      for each feature f in userFPL do
7:          if f is in itemFL then
8:              add f to commonFeat
9:              f.movieList ← findContItems(f, user)
10:         end if
11:     end for
12:     return formExp(commonFeatures)
13: end procedure
```

The *Why* justification takes the following form: *"I am recommending x because…"* In this condition, we support users in tracing the causes of why a movie is recommended. To do so, we follow the algorithm given in Algorithm 1). We first find the features of the item being recommended (Line 2) and positive features that exist in the user profile (Line 3)—the ones that have positive accumulated scores, as described in Section 4.2. Positive features are sorted in descending order of their scores within the user profile (Line 4) and the algorithm subsequently tries to find the common features between the user profile and the item profile (Lines 6—11). The highest-scored features of the user profile are prioritized in the generation of the final justification string. The algorithm also finds movies in the user's profile that contain the corresponding common features, which are used to form the justification text. Figure 3(b) shows an example of a *Why* justification.

ALGORITHM 2: Generating a Why not justification

```
1:  procedure GenerateWhyNotExp(user, item, model)
2:      itemFL ← model.getFeatList(item)
3:      userFPL ← model.getUserFeaturesPos(user)
4:      userFNL ← model.getUserFeaturesNeg(user)
5:      recList ← model.getRec(user)
6:      sort recList in the order of asc. rec. scores
7:      for each itemToCompare in recList do
8:          itemToCFL ← model.getFeatList(itemToCompare)
9:          if userFNL is ∅ then
10:             featExc ← itemToCFL − userFPL
11:             featExc ← featExc − itemFL
12:             if featuresExc ≠ ∅ then
13:                 return formExp(itemToCompare, item, featExc)
14:             end if
15:         else
16:             commonNegFeat ← {}
17:             for each feature f in itemToCFL do
18:                 if f ∈ userFeaturesNL and f ∈ itemFL then
19:                     add f to commonNegFeat
20:                     f.movieList ← findContItems(f,user)
21:                     return formExp(item, itemToCompare,
22:                 commonNegFeat)
23:                 end if
24:             end for
25:         end if
26:     end for
27: end procedure
```

The *Why not* justification takes the following form: *"I am recommending x but not y because…"* In this condition, the aim is to provide a comparison between the actual recommended item x and another potential item y. y is considered as a possible recommendation for the user, but it is not recommended, because x has a better alignment with the user's preferences.

While some existing justification mechanisms allow the user to select the y item [36], in our system it is implemented as a system-controlled item selection mechanism. This reduces

uncontrolled variations that may result from users' selection of item y and avoids burdening the user with the additional cognitive effort of selecting item y during the user study.

The procedure for generating *Why not* justifications is outlined in Algorithm 2 (see Appendix B). The algorithm first finds the features in the recommended item profile (Line 2), positive features of the user profile (Line 3) and the negative features of the user profile (Line 4). Next, it processes potential y's that can be compared with the actual recommended item x. To achieve this, it calls the model to generate recommendations (Line 5) and sorts the returned list such that the lowest-scored items appear first in the list (Line 6).

Ideally, the algorithm points out negative aspects of item y, but if the user has only rated movies with a positive score, the algorithm has no information about potential negative features. In this case, the algorithm instead tries to find features that exist in the potential item y but not in the user's positive features and the feature set of x (Line 10). The algorithm tries to find an item y that has at least one of such features (Line 11). Once a potential y with such features is found, those features can be used to form a justification string for the *Why not* justification (Lines 12 and 13).

If, however, the user has provided negative ratings for some movies (and hence some features) during the preference elicitation step, then for each item in the list of potential y's, the algorithm checks whether y has any features that the user does not like and the actual recommended item x does not contain (Lines 16–20). Once such an item is found, the algorithm forms a justification text using three objects: the recommended item (x), item to be compared (y), and the common negative features between the profile of y and user's negative features in their profile (Lines 21 and 22). In addition, the algorithm also finds movies rated by the user that contains those negative features to add into the final justification text. Figure 3(c) shows an example of the Why not justification.

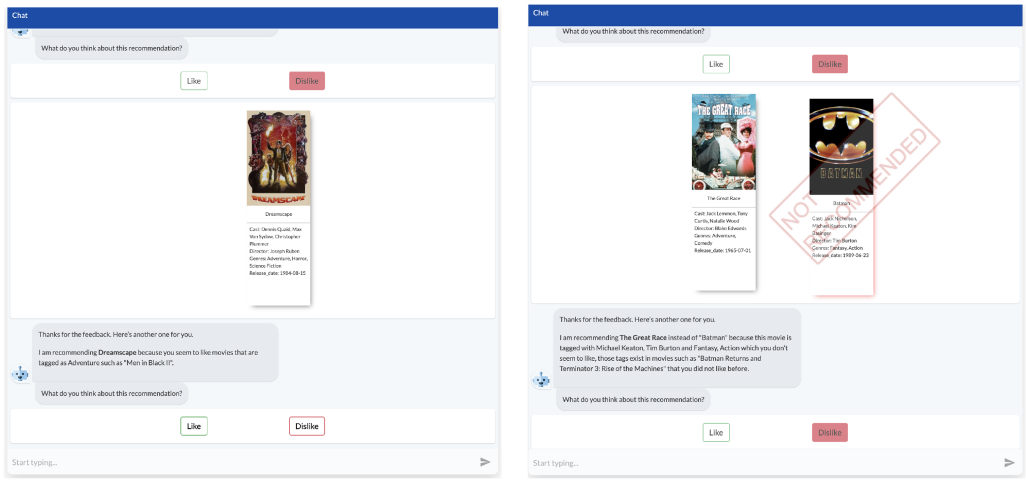
The *combination* justification refers to the experimental condition that presents both the Why justification and the Why not justification at the same time. To achieve this, both procedures given in Algorithms 1 and 2 are called, and their justifications are presented together. Figure 3(d) shows an example.

4.5 Design Rationale

To address our research questions, we developed a movie recommending chatbot that was button-based to improve the ease of use compared to a text-based interface.

To avoid a cluttered interface, we chose to provide meta-data about each recommendation using a movie card that included the movie poster, title, year, genre, release date, and the cast. After each movie card was presented, the justification was provided in a message bubble. After rating like or dislike the user was provided with another movie option until they were shown five movies. While all the other design components remained the same, the presentation of the movie card differed slightly depending on the justification style (see interfaces in Figure 4). In the *why not* and *combination* conditions, the recommended item was shown in addition to the alternative that was not recommended (Figure 4(b)). The item that was not recommended was prominently labeled and highlighted in red. Whereas, users in the *why* and *none* conditions were only shown the movie card for the recommended item (Figure 4(a)).

Initially, we ran a small pilot study with ten participants to ensure there were no usability issues. These participants were asked to interact with the system but verbalize their thought process as they proceeded in the study. This think-aloud approach has been shown as a useful usability tool to evaluate systems [3]. As a result, we learned that offering more than five movies in the preference elicitation steps could lead to participant fatigue. Therefore, the design was updated to feature five movies.



(a) Chat Interface: Why justification (b) Chat Interface: Why not justification

Fig. 4. Examples of the interface for the different justification styles. The movie card displays relevant information about the movie such as the title, year, genre, and year.

4.6 Participants and Procedure

For the experiment, we recruited 317 participants from Amazon's Mechanical Turk.[2] The study finished with 310 participants after checking validation metrics. In total, we used three validation metrics to check for participants' attention and potential bad data points.

We considered the time taken to complete the three main stages of the study, an attention-checking question based on interaction with the stimulus, and we reviewed the open-ended questions to rule out bots. Participants were paid US$2.00 for the task, which took around 15 minutes to complete. Of the participants who successfully passed the validation metrics, 62% identified as men, 37% as women while 0.3% reported as non-binary and 0.3% preferred not to disclose. Most of the participants tended to be younger: 18–24 (11%), 25–29 (26%), 30–34 (26%), 35–39 (15%), 40–44 (5%), 45–49 (8%), 50–54 (4%), and 55+ (4%). We summarize the distribution of participants in Table 1.

Figure 5 provides an overview of the study procedures. After accepting the consent form, participants were directed to complete the pre-survey that collected demographic (age and gender) and personal characteristic information (disposition to trust [43] and technology experience [18]). We adapted items for gender based on recent recommendations on gender inclusion in surveys [50]. We then provided an overview of the instructions for the stimulus stage. After reviewing the instructions, participants were directed to a chatbot interface where they were greeted by Moviebot.

To learn about participants' preferences, they were asked to rate five movies on a scale of 1 to 5 stars. If participants were not familiar with a movie option, then they were given the opportunity to replace it with a new option. Once the ratings for five options were submitted, participants could proceed.

Moviebot subsequently presented a total of five movie recommendations with either Low or High accuracy (randomly assigned) and with a randomly assigned justification style (None, Why, Why not, or combination). We chose to present five recommendations so as to not overwhelm participants; this was based on feedback from a pilot study. Aside from the "None" condition, a

[2] https://www.mturk.com.

Table 1. The Distribution of Participants across the Different Conditions

Justification Style	Low Accuracy	High Accuracy	Total
None	38	39	77
Why	39	36	75
Why Not	38	42	80
Combination	39	39	78

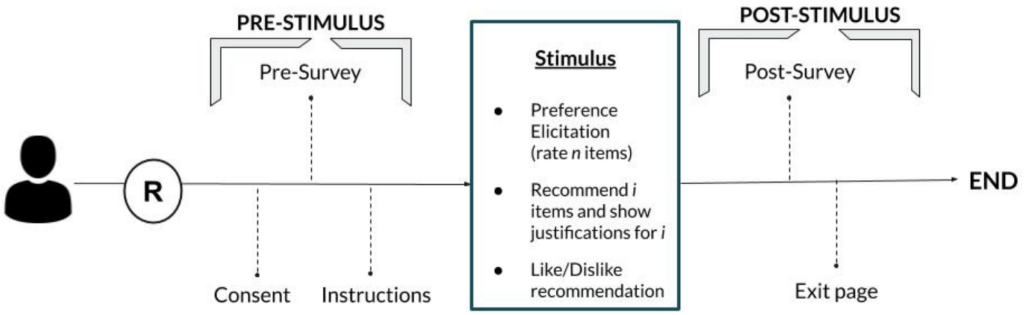


Fig. 5. Study procedure. Participants were randomly assigned to one of eight conditions. They were exposed to one of the four levels of justification styles and we also controlled the performance of the algorithm (low or high).

justification preceded each recommendation. For each recommendation, participants were asked "What do you think about the recommendation?," which they could answer by pressing a like or dislike button. After completing the interaction with Moviebot, participants were directed to complete the post-stimulus survey. Next, we describe the measures that were adopted in the study.

4.7 Measures

A key component of evaluating any online system involves exploring the subjective factors that influence the user experience, also known as constructs. We consider leading frameworks within the field that consider how objective aspects of our system (e.g., different justifications or the accuracy of the recommendations) are perceived by users, and how their personal characteristics lead to specific user experiences and interactions with the system [30, 45].

As such, we consider pre-validated measurement scales to operationalize each factor/construct that would help to evaluate users' perceptions of the system. Each construct was measured using multiple items that were phrased as statements, with a 5-point response format from "strongly disagree" to "strongly agree" (see Table 2). We tested the construct for reliability by examining the Cronbach's α value for each construct.

Perceived Justification Efficiency: We adapted four items from Vandenbosch et al. [54]. This construct focused on whether respondents found the justification helpful in *quickly* assessing a recommendation (**average variance extracted (AVE)** AVE = 0.908, Alpha = 0.926). Exemplar items included "The chatbot saves me time" and "The chatbot's justification speeds up my decision making."

Perceived Justification Effectiveness: We adapted four items from Vandenbosch et al. [54]. This construct focused on whether respondents found the justification helpful in *correctly* determining the suitability and quality of the recommended items (AVE = 0.866, Alpha = 907). For example, "The chatbot's justification improves the quality of my movie decisions."

Table 2. The Survey Items Per Construct with Item Factor Loadings

Construct	Items	Factor Loading
Perceived Justification Efficiency	The chatbot's explanation saves me time.	0.93
	The chatbot's explanation improves the chatbot's efficiency.	0.88
	The chatbot's explanation speeds up my decision making.	0.91
	The chatbot's explanation enables me to react more quickly to movie recommendations.	0.90
Perceived Justification Effectiveness	The chatbot's explanation helps me to respond more appropriately to movie recommendations.	0.87
	The chatbot's explanation improves the quality of my movie decisions.	0.90
	The chatbot's explanation helps me to determine which movies to choose.	0.85
	The chatbot's explanation improves the chatbot's effectiveness.	0.84
Transparency	I understood why the movies were recommended to me.	0.79
	The information provided for the recommended movie is sufficient for me to make a decision.	
	The movies recommended to me had similar attributes to my preference.	0.95
Trusting Beliefs (Integrity)	This chatbot provides unbiased movie recommendations.	0.70
	This chatbot is honest.	0.91
	I consider this chatbot to be of integrity.	0.91
(Benevolence)	This chatbot puts my interest first.	0.90
	This chatbot keeps my interests in its mind.	0.90
	This chatbot wants to understand my needs and preferences.	0.80
(Competence)	This chatbot is like a real expert in assessing movies.	0.85
	This chatbot has the expertise to understand my needs and preferences about movies.	0.92
	This chatbot has the ability to understand my needs and preferences about movies.	0.87
	This chatbot has good knowledge about movies.	0.73
	This chatbot considers my needs and all important attributes of movies.	0.85
Perceived Control	I had limited control over the way the chatbot made explanations.	
	The chatbot does what I want.	0.90
	I would like to have more control over the chatbot.	
	I had full control over the chatbot.	0.82
Recommendation Quality	I liked the movies recommended by the movie recommender.	0.93
	I found the recommended movies appealing.	0.94
	The recommended movies fit my preference.	0.91
	The recommended movies were relevant.	0.93
	I didn't like any of the recommended movies.	

(Continued)

Table 2. Contnued

Construct	Items	Factor Loading
Social Presence	There is a sense of human contact when using the chatbot.	0.86
	There is a sense of personalness when using the chatbot.	0.95
	There is a sense of sociability when using the chatbot.	
	There is a sense of human warmth when using the chatbot.	0.90
	There is a sense of human sensitivity when using the chatbot.	
Trusting Intention (Willingness to Depend)	When I need a movie suggestion, I would feel comfortable depending on the recommendations provided by this chatbot.	
	I can always rely on this chatbot to help me make a decision about tough movie choices.	0.90
	I feel that I could count on this chatbot to help when I do not know what movie to watch.	0.92
	Faced with a situation that required me to use a movie recommendation service (for a fee), I would use the company that integrates this chatbot.	0.87
(Follow Advice)	If I had a challenging movie choice, I would want to use this chatbot again.	0.91
	I would feel comfortable acting on the movie recommendations given to me by this chatbot.	0.87
	I would not hesitate to use the movie recommendations this chatbot supplied me.	0.90
	I would confidently act on the movie recommendations I was given by this chatbot.	

Removed items are colored in grey.

Transparency: We adapted three items from Millecamp et al. [41]. This construct was focused on measuring users' understanding of the recommendation rationale (AVE = 0.869, Alpha = 0.819). For example, "I understood why the movies were recommended to me."

Perceived Control: Four items from Knijnenburg et al. [27]. This construct measures respondents' perception that the system allows them to control the recommendation process (AVE = 0.856, Alpha = 787). For example, "I had limited control over the way the chatbot made justifications" (reversed).

Recommendation Quality: Five items from Reference [27]. This construct measures how useful the recommendations are perceived to be by users (AVE = 0.926, Alpha = 0.948). For example, "I liked the recommended movies."

Social Presence: We adapted five items from Gefen et al. [20]. This construct measured whether respondents felt like the chatbot cared for them (AVE = 0.905, Alpha 0.822). For example, "There was a sense of personalness when using the chatbot."

Chatbot Experience: Four items adapted from Reference [2]. For example, "I use chatbots frequently."

Trusting Beliefs: Using McKnight et al.'s framework on trust [39, 55], trusting beliefs consists of *integrity* (three items, AVE = 0.842, Alpha 0.901), *benevolence* (three items, AVE = 0.868, Alpha = 0.867), and *competence* (five items, AVE = 0.843, Alpha = 0.903) . Exemplar items from the integrity construct include "The chatbot provides unbiased movie recommendations."

Trusting Intention: This aspect consists of two constructs: *willingness to depend on the advice of the system* (AVE = 0.894, Alpha = 0.9) and *willingness to follow advice* (AVE = 0.896, Alpha = 0.908)

[39]. For example, "When I need a movie suggestion, I would feel comfortable depending on the recommendations provided by the chatbot."

After completing these questions, participants were asked to answer four open ended questions about their experience interacting with the system. Exemplar questions included: "What aspects of the system do you think were most helpful in understanding your recommendations?" These items were adapted from Millecamp et al. [41].

5 RESULTS

First, we conducted a **Confirmatory Factor Analysis (CFA)** and examined the validity and reliability scores of the constructs measured in our study. Upon inspection of the CFA model, we merged the "justification efficiency" and "justification effectiveness" factors due to a lack of discriminant validity (i.e., a correlation between them that was larger than the square root of the AVE of each factor). This resulted in the *justification quality* factor. Similarly, "willingness to depend" and "willingness to follow advice" were merged into one stable factor. In inspecting the remaining factors, we found the values of Cronbach's α were high.[3] Furthermore, the AVE for all factors exceeded 0.50, indicating convergent validity.

We then subjected the nine factors, the experimental conditions, and selected interaction behaviors to Structural Equation Modeling. The model has a good model fit[4]: $\chi^2(772) = 1158.937$, $p < .01$; RMSEA = 0.032, 90% CI: [0.028, 0.037], CFI = 0.990, TLI 0.991. The corresponding model is shown in Figure 7.[5] For clarity, we report significant direct effects from left to right.

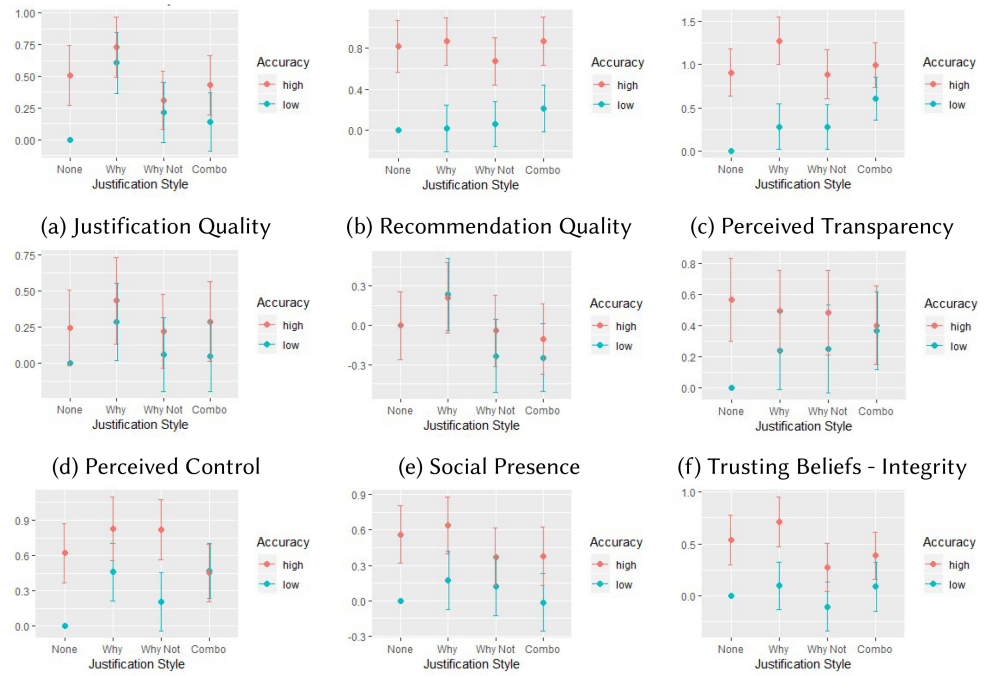
Effects of Justification Style: The results indicate that there were no significant interaction effect between justification style and algorithmic accuracy on the justification quality factor (all things held constant) ($\chi^2(3) = 0.567$, $p > 0.05$). The model shows that the manipulation of the justification style had a positive independent main effect on the perceived quality of the justification compared to having no justification (H1a supported). Participants with justifications around why an item was recommended perceived about 0.39 standard deviation higher levels of justification quality compared to having no justifications—a small to medium sized effect. However, for the "why not" and the combination conditions, the effect was not significant. For marginal effects of justification style on justification quality see Figure 6(a).

Meanwhile, there was positive direct effect of the perceived justification quality on recommendation quality, perceived transparency and perceived control: The more users perceive quality from the justification and the recommendations, the more transparent the system was perceived to be and that increased participants feelings that they had control over what was being recommended. There was no significant direct effect of any of the justification styles on transparency. However, there was a significant indirect effect of the *why* justification style on transparency via justification quality ($\beta = 0.639$ $p < 0.05$) (H1b supported). Regarding interaction behavior, higher perceived justification quality had a *negative* effect on liking recommended items. A potential explanation for this effect may be that higher quality justifications may help users to identify items that may not be relevant. Note that the recommendation quality yields a competing impact ($\beta = 0.813$ $p < 0.001$), hence the overall effect of justification quality on liking behavior is positive.

[3]For alpha, >.70 is acceptable, >.80 is good, >.90 is excellent.

[4]A model should not have a non-significant χ^2, but this statistic is regarded as too sensitive [4]. Hu and Bentler [25] propose cutoff values for other fit indices to be: CFI > .96, TLI > .95, and RMSEA < .05, with the upper bound of its 90% CI below 0.10.

[5]Significance levels: $^{***}p < .001$, $^{**}p < 0.1$, $^{*}p < 0.05$. R^2 is the proportion of variance explained by the model. Numbers on the arrows represent the β coefficients (and the standard error) of the effect. Aspects represented: Objective System Aspects, UPQ, Subjective System Aspects, Personal Characteristics, Situational Characteristics, Interaction, User Experience.



(a) Justification Quality (b) Recommendation Quality (c) Perceived Transparency

(d) Perceived Control (e) Social Presence (f) Trusting Beliefs - Integrity

(g) Trusting Beliefs - Benevolence (h) Trusting Beliefs - Competence (i) Willingness to Depend

Fig. 6. Marginal effects of justification styles and algorithmic accuracy on the subjective factors. The effects of the "no justification, low accuracy" condition is set to zero, and the y-axis is scaled by the sample standard deviation.

Effect of Algorithmic Accuracy: The performance of the algorithm had a positive effect on recommendation quality (H2 supported). Unsurprisingly, the algorithm that better aligned with participants' interests was perceived to provide higher quality recommendations—a large, 1.0 standard deviation difference. Subsequently, the perceived recommendation quality had an effect on perceived transparency and perceived control. Hence, indirectly, the high accuracy algorithm improved participants' view on the system's transparency and their feeling of control. For marginal effects of algorithmic accuracy on the subjective constructs see Figure 6.

Effect on Trusting Beliefs: Trusting beliefs were represented with three trust facets - integrity, benevolence, and competence. There were no significant direct main effects of justification style on trusting beliefs, but there is a indirect effect on all facets of trust via the route [justification style → justification quality → perceived transparency → (trusting beliefs)] (H1a supported; H1c not supported). Integrity was positively influenced by perceived control and trust disposition. Meanwhile, integrity and transparency have a positive direct effect on benevolence. This suggests that participants view that the system as acting in their best interests if they perceive integrity and transparency. In turn, benevolence, social presence, perceived control, transparency, and justification quality positively affected users' perception of competence. The negative direct effect of integrity on competence suggests that users can view a system as acting in line with a set of accepted principles but that this may negatively impact their perceptions of the system's abilities. Figure 6(f), (g), and (h) provide an overview of the marginal effects of the justification styles and algorithmic accuracy on trusting beliefs.

Effect on Trusting Intention: We observed four positive direct effects on participants trusting intentions. Participants trusting beliefs, particularly integrity and competence, significantly

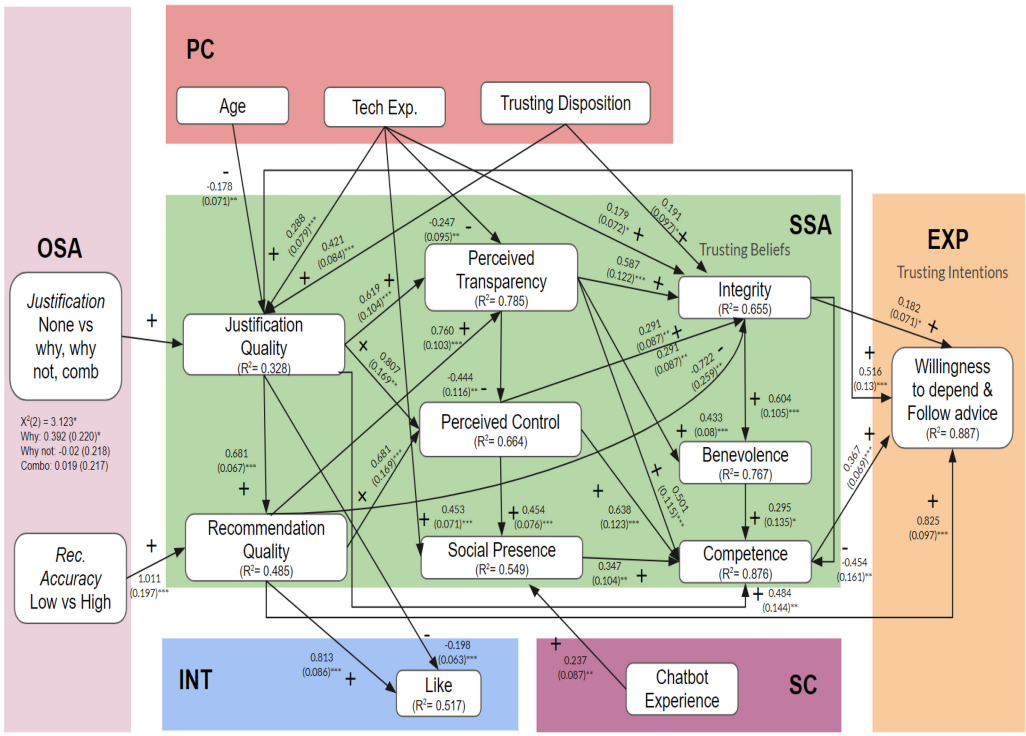


Fig. 7. The structural equation model for the experiment using Knijnenburg et al.'s framework [30].

influenced participants' willingness to depend on and follow advice from the system. Benevolence was the only facet of trust that did not yield a significant effect. Similarly, the perceived quality of the recommendation and the justification also had a positive effect. According to the parameter weight, recommendation quality has the strongest influence ($\beta = 0.825$ $p < 0.001$), but the quality of the justification also had a strong effect ($\beta = 0.516$ $p < 0.001$).

There were no direct effects of the justification style on participants' willingness to depend on and follow advice from the system ($\chi^2(3) = 5.097$, $p > 0.05$). However, the "why not" justification style had a negative effect ($\beta = -0.331$, $p < 0.01$) while "why" and "combination" had negative but non-significant effects. The higher justification quality of the "why" justification ultimately gives this explanation type the upper hand when it comes to users' willingness to depend on and follow the advice of the RS. See Figure 6(i) for the marginal effects of justification style and algorithmic accuracy on willingness to depend on and follow advice.

Personal and Situational characteristics: Participants with more experience with technology perceived more social presence in their interactions with the system, and judged the system to have a higher level of integrity. They also perceived less system transparency, though. Participants' trusting disposition positively influences their perception of the justification quality as well as the integrity of the system, which suggests those with a greater tendency to trust technology would assume the system acts with a set of principles that are acceptable to participants (e.g., transparency and honesty). Last, age influenced perceived justification quality (with younger participants perceiving a higher justification quality), while gender did not significantly influence any factor ($p > 0.05$) (RQ3).

6 DISCUSSION

Based on the results of our experiment, we describe the benefits of justifications in conversational recommenders and the relationship with algorithmic accuracy. We offer insights into the impact on users' perceptions and present preliminary suggestions on the use of "why" versus "why not" justification styles. We also describe the significance of algorithmic accuracy. Last, we discuss how our results are influenced by users' personal characteristics.

6.1 Impact of Justifications

Our results show that only the "why" justifications are considered to have a higher quality than no justification at all (R1). In effect, the study provides supporting evidence that the use of "why" style justifications enhances users' trusting beliefs, perception of system transparency, and user experience with conversational recommenders, with other justifications having no significant effect. Moreover, "why not" justifications negatively influenced users' willingness to depend on and follow advice from the system. For explanatory tools this distinction is important when considering effect designs. Justifications may become particularly important to people when there are problematic predictions from the system. Therefore, efforts to alleviate concerns would be critical to maintain a positive experience and to retain that consumer. Being clear with "why" a recommendation was delivered could offer insight into the underlying workings of the system. As such, "why" justifications significantly influenced users' perception of system transparency and control. The results align with existing work that showed that what is perceived to be a "better explanation" would positively impact trusting beliefs and adoption intentions [33, 49, 55]. For future intelligent systems,

Furthermore, describing the rationale behind "why" a recommendation was made versus "why not" may be cognitively easier to process, thus, quicker for a user to identify if a system correctly or incorrect inferred preferences. This could be important for incorporating direct feedback to improve predictions and building trust with the system. A system built with scrutability (allowing users to tell the system if it is wrong) as a feature may be appropriate for the conversational nature of a chatbot agent. Thus, this may improve perceived accountability in future intelligent systems.

Prior work has found that users may adopt normative or pragmatic views regarding the explanations in intelligent systems: Users with normative views are motivated by detailed and comprehensive explanations while those with a pragmatic view are motivated by benefits for usability and efficient use [16]. As such, we suspect that "why" justifications may be viewed as more pragmatic and therefore more efficient and succinct. Indeed, Lim et al. also found that "why" explanations were preferred over "why not" explanations [36] for their simplicity. Moving forward, we anticipate that this would have implications for how justifications would be designed, since simplicity and familiarity appear to be sensitive to the effectiveness of the justification. We note that for users in the "why not X" conditions, the option that was not recommended (item X) was chosen by the system. Future work could investigate the possibility of having the user ask the "why not" question about a particular item of their own choosing.

6.2 Impact of Algorithmic Accuracy

One of the primary functions of a recommendation agent is improving the agent's ability to predict to what extent an item will match a person's preferences, interests, and goals. Accuracy is a common metric used to test the effect of variations in prediction performance. By distinguishing between high and low algorithmic accuracy, we were able to demonstrate that there was a significant effect of the performance of the algorithm on the perceived recommendation quality (see marginal effects in Figure 6).

Our results indicate that being able to accurately predict users' preferences significantly impacts how people perceive the system's ability to suggest useful items that are of quality. This, in turn, helps to establish long-term trust toward the agent. At the same time, even with less relevant items, providing justifications had an independent effect on users trusting beliefs and their willingness to follow the system's recommendations. For stakeholders, this suggests that offering insights into why a recommendation is made may influence consumers' trusting beliefs and their willingness to follow through with recommendations, even if the recommendation is not perfectly aligned with the user's preferences. This may prove to be vital for conversational recommenders that are used commercially as justifications could assist in building trust and influencing system adoption. System designers could consider prior work that have called for systems for self-actualization that not only focus on providing the best recommendation but offering support to users in exploring, developing, and understanding their preferences [28, 51]. Care must be taken, though, not to cajole users into accepting items that they will eventually not like, as this would erode trust.

6.3 Impact of Personal Characteristics

Another contribution of our work is an empirical investigation into how users' prior positions on trust in technology impacts their reaction to different justifications (RQ3). Users with more technology experience perceived less transparency from the system. Having more experience with different systems may increase the expectations of interactions online. In line with the Expectancy Violation Theory, users form expectations based on their interaction with other actors and they evaluate their current experience based on their expectations [11]. As such, their evaluation of met-expectations or violated-expectations affect interaction outcomes. Therefore, as more tools around transparency and explainability become available, developers should consider that users expectations may also increase.

Similarly, users who were more trusting of technology perceived more control, integrity, and justification quality from the chatbot. Although this is positive, designers should consider possible consequences for users who may be more skeptical about technology.

7 LIMITATIONS

We performed our study with MTurk workers who may be more open to trying and thus trusting technology. Also, our sample is slightly skewed to younger males and as such that may affect the generalizability of our results. The interactive capabilities of the chatbot were also limited to allow the researchers to control for the user experience. Additionally, there are many styles and elements of justifications not evaluated in the study.

Another potential limitation is that we use a content-based recommendation algorithm rather than a more common collaborative filtering algorithm that may impact accuracy. The reason for this choice is that content-based algorithms are easier to explain; although there is some existing work on the explanation of collaborative filtering algorithms [24].

8 CONCLUSION

In this article, we have studied the effect of providing justifications in conversational recommenders by employing three different justification styles, as well as interactions with two different recommendation algorithms: one with high accuracy, and another with lower. Our results show that *why* justifications (rather than *why not*) significantly influence users' perception of system transparency, influences perceived control, and in turn affects users' trusting beliefs and intentions. The contributions of this work are multi-disciplinary as we contribute to the fields of Human-Computer Interaction and Recommender Systems by offering a better understanding of user perceptions towards justifications in conversational recommenders that could benefit the

design of explainability-related tools for recommender systems. Future research should investigate the possibilities of other methods for explaining recommendations considering both the explanation style and other interactive methods.

REFERENCES

[1] Titipat Achakulvisut, Daniel Acuna, Tulakan Ruangrong, and Konrad Kording. 2016. Science concierge: A fast content-based recommendation system for scientific publications. *PLoS One* 11, 04 (2016). https://doi.org/10.1371/journal.pone.0158423

[2] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*.

[3] Kathy Baxter, Catherine Courage, and Kelly Caine. 2015. *Understanding Your Users: A Practical Guide to User Research Methods*. Morgan Kaufmann.

[4] Peter M. Bentler and Douglas G. Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* 88, 3 (1980), 588.

[5] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to recommend?: User trust factors in movie recommender systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces (IUI'17)*. ACM, New York, NY, 287–300. https://doi.org/10.1145/3025171.3025209 event-place: Limassol, Cyprus.

[6] Shlomo Berkovsky, Ronnie Taib, Yoshinori Hijikata, Pavel Braslavsku, and Bart Knijnenburg. 2018. A cross-cultural analysis of trust in recommender systems. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP'18)*. Association for Computing Machinery, New York, NY, 285–289. https://doi.org/10.1145/3209219.3209251

[7] Anol Bhattacherjee and Chieh-Peng Lin. 2015. A unified model of IT continuance: Three complementary perspectives and crossover effects. *Eur. J. Inf. Syst.* 24, 4 (2015), 364–373. DOI: https://doi.org/10.1057/ejis.2013.36

[8] Timothy Bickmore and Justine Cassell. 2001. Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 396–403.

[9] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: A visual interactive hybrid recommender system. In *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys'12)*. ACM, New York, NY, 35–42. https://doi.org/10.1145/2365952.2365964

[10] Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Why people use chatbots. In *Proceedings of the International Conference on Internet Science*. Springer, 377–392.

[11] Judee K. Burgoon, Joseph A. Bonito, Paul Benjamin Lowry, Sean L. Humpherys, Gregory D. Moody, James E. Gaskin, and Justin Scott Giboney. 2016. Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *Int. J. Hum.-Comput. Stud.* 91 (Jul. 2016), 24–36. https://doi.org/10.1016/j.ijhcs.2016.02.002

[12] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. ACM, New York, NY, 815–824. https://doi.org/10.1145/2939672.2939746

[13] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model. User-Adapt. Interact.* 18, 5 (Nov. 2008), 455–496. https://doi.org/10.1007/s11257-008-9051-3

[14] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. arXiv preprint arXiv:1511.06931.

[15] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. 2019. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI'19)*. ACM, New York, NY, 408–416. https://doi.org/10.1145/3301275.3302274

[16] Malin Eiband, Hanna Schneider, and Daniel Buschek. 2018. Normative vs. pragmatic: Two perspectives on the design of explanations in intelligent systems. In *Proceedings of the IUI Workshops*.

[17] Alexander Felfernig and Bartosz Gula. 2006. An empirical study on consumer behavior in the interaction with knowledge-based recommender applications. In *Proceedings of the 8th IEEE International Conference on E-Commerce Technology and the 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)*. IEEE, 37–37.

[18] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *Int. J. Hum.–Comput. Interact.* 35, 6 (2019), 456–467.

[19] Gerhard Friedrich and Markus Zanker. 2011. A taxonomy for generating explanations in recommender systems. *AI Mag.* 32 (2011), 90–98.
[20] David Gefen and Detmar W. Straub. 2004. Consumer trust in B2C e-commerce and the importance of social presence: Experiments in e-products and e-services. *Omega* 32, 6 (2004), 407–424.
[21] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*. ACM, 227–236.
[22] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 5 (2019), 93.
[23] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (2019).
[24] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW'00)*. ACM, New York, NY, 241–250. https://doi.org/10.1145/358916.358995
[25] Li-tze Hu and Peter M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Eq. Model.* 6, 1 (1999), 1–55.
[26] Antti Kangasrääsiö, Dorota Glowacka, and Samuel Kaski. 2015. Improving controllability and predictability of interactive recommendation interfaces for exploratory search. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI'15)*. ACM, New York, NY, 247–251. https://doi.org/10.1145/2678025.2701371
[27] Bart P. Knijnenburg, Niels J. M. Reijmer, and Martijn C. Willemsen. 2011. Each to his own: How different users call for different interaction methods in recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems*. ACM, 141–148.
[28] Bart P. Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. 2016. Recommender systems for self-actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 11–14.
[29] Bart P. Knijnenburg and Martijn C. Willemsen. 2016. Inferring capabilities of intelligent agents from their external traits. *ACM Trans. Interact. Intell. Syst.* 6, 4 (Nov. 2016). https://doi.org/10.1145/2963106
[30] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Model. User-Adapt. Interact.* 22, 4–5 (2012), 441–504.
[31] Sherrie Y. X. Komiak and Izak Benbasat. 2006. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly* 30, 4 (2006), 941–960. http://www.jstor.org/stable/25148760.
[32] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI'19)*. ACM, New York, NY, 379–390. https://doi.org/10.1145/3301275.3302306
[33] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, New York, NY, 487:1–487:12. https://doi.org/10.1145/3290605.3300717 event-place: Glasgow, Scotland Uk.
[34] Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*. ACM, 195–204.
[35] Brian Y. Lim and Anind K. Dey. 2010. Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. ACM, 13–22.
[36] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*. ACM, New York, NY, 2119–2128. https://doi.org/10.1145/1518701.1519023
[37] Tariq Mahmood and Francesco Ricci. 2009. Improving recommender systems with adaptive conversational strategies. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (HT'09)*. ACM, New York, NY, 73–82. https://doi.org/10.1145/1557914.1557930
[38] Kevin McCarthy, James Reilly, Lorraine McGinty, and Barry Smyth. 2004. On the dynamic generation of compound critiques in conversational recommender systems. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, Paul M. E. De Bra and Wolfgang Nejdl (Eds.). Springer, Berlin, 176–184.
[39] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Info. Syst. Res.* 13, 3 (Sep. 2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81
[40] Sean M. McNee, Shyong K. Lam, Joseph A. Konstan, and John Riedl. 2003. Interfaces for Eliciting new user preferences in recommender systems. In *Proceedings of the 9th International Conference on User Modeling*. Springer, Berlin, Heidelberg, 178–187.
[41] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. 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 (IUI'19)*. ACM, New York, NY, 397–407. https://doi.org/10.1145/3301275.3302313

[42] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* (2018).
[43] Cecilie Bertinussen Nordheim. 2018. *Trust in Chatbots for Customer Service—Findings from a Questionnaire Study.* Master's thesis.
[44] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. 2008. PeerChooser: Visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08).* ACM, New York, NY, 1085–1088. https://doi.org/10.1145/1357054.1357222
[45] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI'06).* ACM, New York, NY, 93–100. https://doi.org/10.1145/1111449.1111475 event-place: Sydney, Australia.
[46] Ashwin Ram. 1989. Question-driven understanding: An integrated theory of story understanding, memory and learning. Ph.D. Dissertation. Yale University.
[47] Reports and Data. 2019. Global Chatbot Market Size & Analysis | Industry Report, 2019-2026. https://www.reportsanddata.com/report-detail/chatbot-market.
[48] Milene Selbach Silveira, Clarisse Sieckenius de Souza, and Simone D. J. Barbosa. 2001. Semiotic engineering contributions for designing online help systems. In *Proceedings of the 19th Annual International Conference on Computer Documentation.* ACM, 31–38.
[49] Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems.* ACM, 830–831.
[50] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. 2019. How to do better with gender on surveys: A guide for HCI researchers. *Interactions* 26, 4 (2019), 62–65.
[51] Emily Sullivan, Dimitrios Bountouridis, Jaron Harambam, Shabnam Najafian, Felicia Loecherbach, Mykola Makhortykh, Domokos Kelen, Daricia Wilkinson, David Graus, and Nava Tintarev. 2019. Reading news with a purpose: Explaining user profiles for self-actualization. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization.* 241–245.
[52] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos. 2008. Providing justifications in recommender systems. *IEEE Trans. Syst. Man, Cybernet. A: Syst. Hum.* 38, 6 (2008), 1262–1272.
[53] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Model. User-Adapt. Interact.* 22, 4 (01 Oct. 2012), 399–439", issn=.
[54] Betty Vandenbosch and Michael J. Ginzberg. 1996. Lotus notes and collaboration: Plus ça change... *J. Manage. Inf. Syst.* 13, 3 (1996), 65–81.
[55] Weiquan Wang and Izak Benbasat. 2007. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *J. Manage. Inf. Syst.* 23, 4 (May 2007), 217–246. https://doi.org/10.2753/MIS0742-1222230410
[56] Jingjun David Xu, Ronald T. Cenfetelli, and Karl Aquino. 2016. Do different kinds of trust matter? An examination of the three trusting beliefs on satisfaction and purchase behavior in the buyer–seller context. *J. Strategic Inf. Syst.* 25, 1 (2016), 15–31.
[57] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the AAAI Annual Conference on Artificial Intelligence.*
[58] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM'18).* ACM, New York, NY, 177–186. https://doi.org/10.1145/3269206.3271776

Received May 2020; revised October 2020; accepted December 2020