Whom do Explanations Serve? A Systematic Literature Survey of User Characteristics in Explainable Recommender Systems Evaluation

KATHRIN WARDATZKY, Department of Informatics, University of Zurich, Zurich, Switzerland
OANA INEL, Department of Informatics, University of Zurich, Zurich, Switzerland
LUCA ROSSETTO, Dublin City University, Dublin, Ireland
ABRAHAM BERNSTEIN, Department of Informatics, University of Zurich, Zurich, Switzerland

Adding explanations to recommender systems is said to have multiple benefits, such as increasing user trust or system transparency. Previous work from other application areas suggests that specific user characteristics impact the users' perception of the explanation. However, we rarely find this type of evaluation for recommender systems explanations. This paper addresses this gap by surveying 124 papers in which recommender systems explanations were evaluated in user studies. We analyzed their participant descriptions and study results where the impact of user characteristics on the explanation effects was measured. Our findings suggest that the results from the surveyed studies predominantly cover specific users who do not necessarily represent the users of recommender systems in the evaluation domain. This may seriously hamper the generalizability of any insights we may gain from current studies on explanations in recommender systems. We further find inconsistencies in the data reporting, which impacts the reproducibility of the reported results. Hence, we recommend actions to move toward a more inclusive and reproducible evaluation.

CCS Concepts: • **Information systems** → **Recommender systems**; *Personalization*; • **Human-centered computing** → **User studies**; *Empirical studies in HCI*; • **General and reference** → **Surveys and overviews;**

Additional Key Words and Phrases: Explainable AI, recommender systems, user studies, literature survey

ACM Reference Format:
Kathrin Wardatzky, Oana Inel, Luca Rossetto, and Abraham Bernstein. 2025. Whom do Explanations Serve? A Systematic Literature Survey of User Characteristics in Explainable Recommender Systems Evaluation. *ACM Trans. Recomm. Syst.* 3, 4, Article 49 (April 2025), 35 pages. https://doi.org/10.1145/3716394

1 Introduction

Recommender systems have become part of many people's everyday online interactions. Whether it is suggestions for what movie to watch, what items to buy, which news to read, or which

This work was done while the Luca Rossetto was at the University of Zurich.

This work was partly supported by the Swiss National Science Foundation through project "MediaGraph" (contract no. 202125).

Authors' Contact Information: Kathrin Wardatzky, Department of Informatics, University of Zurich, Zurich, Switzerland; e-mail: wardatzky@ifi.uzh.ch; Oana Inel, Department of Informatics, University of Zurich, Zurich, Switzerland; e-mail: inel@ifi.uzh.ch; Luca Rossetto, Dublin City University, Dublin, Leinster, Ireland; e-mail: luca.rossetto@dcu.ie; Abraham Bernstein, Department of Informatics, University of Zurich, Zurich, Switzerland; e-mail: bernstein@ifi.uzh.ch.

ACM 2770-6699/2025/04-ART49
https://doi.org/10.1145/3716394

restaurant to visit, these systems generally reach and serve a broad audience in their decision-making process. Adding explanations to recommender systems is said to serve a multitude of benefits. First, by providing the reasoning behind specific recommendations, explanations can mitigate issues dealing with users' autonomy that can occur when recommendations nudge or persuade users in a particular direction that is not beneficial for them [90]. Carefully designed explanations can help the user to identify when the system is "wrong" and provide options to correct it (see the scrutability goal of Tintarev and Masthoff [133]). Second, they can help users understand why they receive a specific recommendation, increase their trust, or allow them to make better decisions [133, 145]. The quality of recommender systems explanations is often assessed based on whether it has such effects on a user. Measuring the explanation quality with offline metrics is still an open research challenge [156]. Text-based explanations have been evaluated with metrics from the natural language processing domain. Ariza-Casabona et al. [10], for example, use BLEU-n, ROUGE-n, and BERT-S along with content repetition, explanation length, and sentence uniqueness measures to measure and compare the text quality of the generated explanations. In addition to BLEU and ROUGE, Kokubo and Sugiyama [72] and Yu et al. [153] evaluate the diversity of the generated sentences in the explanation with the Distinct score. Yu et al. [153] additionally use METEOR, a score based on unigram matching between the generated text and a ground truth. These metrics are, however, limited to text-based explanations. Other metrics include ranking- and retrieval-based metrics, such as Recall and NDCG used in Yu et al. [154]. Except for the diversity-based metrics, these metric types have the downside that they require a ground truth to compare the generated explanations, which is often approximated through online reviews. The results from using online reviews as a ground truth need to be carefully discussed and interpreted as they can be biased [63] and might exclude certain sociodemographics and personality types [87]. Liao et al. [81] further point out that the usefulness of explanations for users is dependent on their motivation for using the system and their downstream actions, which cannot be captured by current offline evaluation metrics.

Previous work found that user characteristics, such as personality, expertise, or cognitive abilities, lead to differences in the effect of explanations on a user. Several works have shown that the Big Five personality traits can impact the persuasiveness of the recommender systems explanations [8, 49, 122]. The characteristics can also impact how visual interfaces are perceived and interacted with [36, 135], or what type of transparency mechanisms for algorithmic decision-making are preferred [123]. Users' cognitive abilities, such as perceptual speed, disembodiment, visual and verbal working memory, and personal characteristics, such as personality traits, impact, for example, the effectiveness of visualizations [37, 119]. Furthermore, interfaces tailored to users' cultural backgrounds improve their satisfaction and efficiency with navigating it [116].

The recommender systems community also has a long-standing history of investigating user characteristics to improve the quality of recommendations [62] and the overall **user experience (UX)** [137]. However, *only a few studies have been conducted in the explainable recommender systems community to investigate how different users perceive explanations* [99], which is why this issue is still widely unknown. This subject matter is, nevertheless, of particular importance, as recent research [77] suggests that the content, design, and goal of the explanations for typical decision-support systems should depend on all stakeholders' needs and interests.

One popular form of recruiting participants for human-subject evaluation is via crowdsourcing platforms such as Amazon **Mechanical Turk (MTurk)**[1] or Prolific[2], who offer access to a global workforce. Research, however, has shown that their participant pool is not as diverse as

[1] https://www.mturk.com/
[2] https://www.prolific.co/

advertised [40], crowd workers on MTurk being, on average, younger and with a higher income than the average population of their country. Further issues with crowdsourcing platforms were found regarding data quality and internal validity, including when measuring participants' personality [30, 45]. While several user characteristics have been identified as relevant in typical user studies, such as age, gender, geographic location, ethnicity, language proficiency, education level, expertise, political, religious, and sexual orientation, physical and mental health, and disability [14, 43, 67], only a few are actually considered when selecting study participants, namely geographic location and language proficiency [67]. Furthermore, research has shown that participant samples in human-subject experiments often come from **Western, Educated, Industrialized, Rich, and Democratic (WEIRD)** societies [57, 82, 127], which means that *the results of these experiments only represent a subset of the overall population.*

In this work, *we systematically analyze what we know about the participants of studies involving recommender system explanations regarding user characteristics such as demographics, prior knowledge, and personality traits.* We look at publications where the effects of explanations were measured with respect to different user characteristics. Often, these effects equal the goals or benefits of adding explanations to recommender systems as identified in Tintarev and Masthoff [133, 134]. One challenge that we face is that, to this date, there are no commonly accepted standards or guidelines for evaluating explanations. This led to differences in the evaluation approaches and made conducting a meta-analysis impossible. Therefore, *our research regarding the explanation effects focuses on exploring* ***which (user) characteristics have been analyzed for which explanation effect****.*

To achieve these goals, we systematically surveyed 124 peer-reviewed papers published between 2017 and 2022 in which explanations were evaluated in user studies. We analyzed them to answer the following research questions:

RQ 1: Who was recruited to evaluate recommender systems explanations?
RQ 2: Which user characteristics impact the effect of explanations in recommender systems?

In the first research question, we look at the description of the participants in the user studies. We analyze the participants' demographic information, personality traits, and prior experience to provide more context to the evaluated explanation effects. Subsequently, we discuss what user groups have not been included when evaluating recommender systems explanations based on the results of RQ 1. By creating awareness about these gaps, we hope to generate opportunities to specifically guide research endeavors in a direction that has improved coverage of the entire range of recommender system users. For the second research question, we look at the findings regarding the explanations' effects on users and investigate how these might differ for specific user characteristics. We look at and discuss the papers to show possible indications for the effects that have been evaluated and highlight gaps where more research is needed.[3]

We conclude with recommendations regarding participant recruitment, reporting of participant data, and evaluation of explanation effects. Hence, in addition to highlighting the interaction between user characteristics and explanations, this work also aims to raise awareness about reproducibility, result reuse, and inclusion issues related to evaluating recommender systems explanations and spark a discussion about addressing these issues in the research community.

2 Related Work

In the literature, several surveys mention or review aspects of the participants in human-subject evaluation of recommender systems explanations or the effects these explanations have on the

[3] We provide an interactive version of this analysis here: https://kathriwa.github.io/interactive-survey-visualization/#/

user. The literature survey by Naiseh et al. [99] is probably closest to this work. They surveyed 48 publications on user needs and implementations of personalized explanations in recommender systems. The authors categorize the findings into users' motivation to interact with the system, their goals to examine the explanations, cognitive load, decision cost, and regulation compliance. They conclude that more user-based research is needed to learn more about how their perception and preferences relate to aspects such as personality, domain knowledge, and user goals.

In terms of general overviews about explainable recommender systems, Zhang et al. [156] point out the need for user behavior analysis and user perspectives as recommender systems are *"inherently human-computer interaction systems".* Nunes and Jannach [108] looked at the number of participants for different study designs in the evaluation of explainable decision support and recommender systems. They also investigated the dependent variables evaluated in their corpus but did not connect these findings with the participants. Several surveys focus on the design and evaluation of recommender systems explanations. For example, Tintarev and Masthoff [134] focus their investigations on the measured effects or goals that explanations are evaluated on but do not provide insights with regard to the participants conducting the evaluations. Mohseni et al. [94] follow a similar route and analyze design goals for AI explanations. They relate these goals to the targeted user type, which they classify based on their experience into AI novices, data experts, and AI experts.

Outside of the recommender systems community, Chromik and Schuessler [32] propose a general taxonomy for human-subjects evaluation for explainable AI. The participant dimension of their taxonomy includes aspects concerning the study type and design, such as the number of participants, their incentivization to participate, how they were recruited, and the participants' foresight (i.e., is the study assuming that all participants have the same knowledge about the context or can they draw from external facts, such as prior experiences). The taxonomy also includes the participants' AI and domain expertise but no other user characteristics. In contrast, Nauta et al. [102] do not focus their survey on the user-based evaluation. Still, they point out that in their corpus spanning the years 2014-2020, only one in five papers evaluate the explanations with users. Their analysis focuses on 12 properties for good explanations and how these can be evaluated quantitatively but do not include user-based evaluations.

Overall, the participants in human-subject evaluation of AI explanations and the effects that explanations can have on the users have been on the radar of some surveys, but these dimensions have not been connected yet.

Aside from survey papers, the impact of the Big Five personality traits on the persuasiveness of recommendations using explanations designed according to Cialdini's six persuasive principles (reciprocity, scarcity, authority, social proof, liking, and commitment) [33] has been evaluated by Alslaity and Tran [8], Sofia et al. [122], and Fatahi et al. [49]. Alslaity and Tran [8] found differences between the two evaluated domains—e-commerce and movie recommendations—within the same personality trait group for the majority of the persuasion profiles. They also found statistically significant interactions between personality traits and domain for three persuasive principles (reciprocity, liking, and scarcity). Sofia et al. [122] designed justifications following Cialdini's persuasive principles for a music recommender system to promote new artists' songs. They not only found differences in the reception of the justifications between participants with different personalities but also that the participants were not very good at assessing which justification type would be the most persuasive for them. Fatahi et al. [49] aimed to use the explanations to persuade users of movies that they were initially unmotivated to watch. They show that explanations containing influence strategies that are tailored to the respective user personality can successfully persuade them to interact with items that were previously not of interest and plan to extend the analysis to other personality traits, such as the need for cognition.

Table 1. Library Databases and the Number of Search Results

Database	**# Results**
ACM Digital Library[4]	905
IEEEXplore[5]	966
Taylor & Francis[6]	1,808
Web of Science[7]	1,380 8
Wiley Online[8]	2,467
SemanticScholar[9]	110,000
Total Retrieved Results	129,954
Final Corpus Size	**124**

These findings underline the motivation for this paper to systematically analyze which users were recruited and investigate what is known about the potential impacts of user characteristics on the measured explanation effects.

3 Methodology

This section describes the paper's methodology by explaining the data collection, selection, and annotation process.

3.1 Data Collection and Selection

We conducted a systematic literature survey of peer-reviewed papers published between 2017 and 2022 to answer the research questions outlined in Section 1. We opted to start the collection with the year 2017 as this year seems to have been the starting point of a steadily increasing number of publications in the explainability field [16, 145]. To collect the papers, we queried six library databases (see Table 1) with the following search term:

```
(explain* OR explanation* OR interpretab* OR intelligib* OR justification OR
transparen*) AND
(recommender OR recommendation OR personalization OR personalized)
```

We constructed the search term to contain words frequently used interchangeably to refer to both explainable AI and recommender systems or recommendations. The first part of the search term includes terms related to the explanation part. Given that there are currently no commonly agreed-upon definitions of AI explanations [54, 89], we included additional terms aiming to capture potentially relevant articles using a different terminology along with variants of *explainability* and *interpretability*. Some articles distinguish between *justifications* and *explanations* while others use *explanation* for *justifications*. Therefore, we opted to include *justifications* in the search. The variations of *transparency* aim to include those articles that opt for a different wording.

The six library databases were selected to cover major computer science venues and interdisciplinary outlets.

[4] https://dl.acm.org/
[5] https://ieeexplore.ieee.org
[6] https://www.tandfonline.com/
[7] https://www.webofscience.com
[8] https://onlinelibrary.wiley.com/
[9] https://www.semanticscholar.org/

Table 2. Exclusion and Inclusion Criteria for the Paper Selection Process

Exclusion Criteria (EC)	
EC-1	The paper is not written in English
EC-2	We have no access to the full paper
EC-3	The content of the paper was published in an extended paper that matches the inclusion criteria
EC-4	The paper was already retrieved from another database
Inclusion Criteria (IC)	
IC-1	The paper proposes an explainable or interpretable recommender system or an explanation generation method
IC-2	The paper presents explanations that are evaluated in a recommendation scenario
IC-3	The paper provides an example or detailed description of the explanation
IC-4	The explanations are evaluated on users
IC-5	The results were tested for statistical significance

This initial search resulted in a total of 129,954 returned articles. In the first round, we filtered out all papers that, based on title and abstract, were clearly not related to explainable AI or recommender systems. Second, we further filtered using the **exclusion** and **inclusion criteria** (**EC** and **IC**) summarized in Table 2. Hereby, we excluded all publications that were not written in English and where we could not access the full text through the licenses of our institution (EC-1 and EC-2). Furthermore, we excluded papers with results published in another paper matching our inclusion and exclusion criteria (EC-3). We kept the extended paper in our considered set of papers. Duplicated papers that appeared in multiple databases were only considered once (EC-4).

The remaining papers had to fulfill all following inclusion criteria to be selected. We included papers that propose an explainable or interpretable recommender system or an explanation generation method (IC-1), and the explanations are evaluated in a recommender systems context (IC-2). To ensure that the explanations are a focus of the papers, we only included publications providing examples or detailed descriptions of the explanations (IC-3). The final two inclusion criteria refer to the evaluation method. Our corpus includes papers in which the explanations were evaluated with a user study (IC-4), where the results were tested for statistical significance (IC-5).

The selection process narrowed the corpus to 124 papers containing a total of 158 user studies (some papers reported on multiple studies), from which we extracted the information reported on the user study participants and the results.[10] In Section 3.2, we elaborate on the annotation process of the papers in our corpus.

3.2 Data Annotation

We extracted general information on the application domain, the evaluation methodology, and the explanation that was evaluated.

To analyze the first research question, we focused on the participant descriptions from the 158 user studies in the papers of our corpus. We extracted all user characteristics mentioned and, if provided, the participants' distribution. The characteristics were then sorted into demographic, personality, and experience categories as suggested by Egan [47]. Table 3 provides an overview of the extracted characteristics.

To answer the second research question, we annotated the findings in the results sections of each paper. More precisely, we extracted the dependent and independent variables of each finding in which a user characteristic was part of a variable along with the outcome of the evaluation (i.e., if an effect was found or not). The dependent variables, or measured explanation effects, were then categorized in the next step. We noticed that the effects did not consistently follow the same definitions and were frequently named differently. Therefore, we referred to the evaluation task

[10] The full list of papers and categorization can be viewed here: https://doi.org/10.5281/zenodo.14771123

or question used for the evaluation, wherever possible (i.e., when they were made available by the authors of the papers). We categorized the explanation effects by the seven explanation goals defined by Tintarev and Masthoff [133]: effectiveness, efficiency, transparency, persuasiveness, trust, satisfaction, and scrutability. By looking at the user task design, we noticed that the papers stating to evaluate the scrutability of the system were doing so by assessing the perceived system control. We, therefore, combined the goals of scrutability and satisfaction, satisfaction being defined as *"increase the ease of usability or enjoyment"* [133], into a usability/UX category. Aside from these goals, we also observed frequent evaluations of the users' perceived explanation quality. Given that user characteristics might impact the user's perception, we added it as an additional factor to the effects we investigated.

As this analysis is a first step to gaining insights into how user characteristics interact with explanation effects, we omitted results for effects specific to the recommendation problem (e.g., group recommendations) or the application domain (e.g., change in driving behavior in autonomous driving).

4 Results

In this section, we present the results of the data analysis to answer our two research questions. We analyze the results and participants' information from the 158 user studies conducted in the 124 papers. First, we look at the information we could extract regarding study participants. Then, we analyze the user study findings in which the effects of explanations were measured by disaggregating different user characteristics.

4.1 Participants

The participants in the user studies presented in our corpus were primarily recruited on crowdsourcing platforms such as MTurk or Prolific and at universities.

Table 3 provides an overview of the extracted characteristics, the number of papers reporting this information about their participants, and the number of papers evaluating whether the given user characteristic impacts the explanation effect.

4.1.1 Demographics. As can be seen in Table 3, it is common practice to provide demographic information about the study participants. Only 18 of the 159 studies in our corpus did not disclose any demographic information about their participants. User demographics reported for most of the studies are the age (114 studies) and gender (105) of the participants. Other demographics that are less frequently mentioned are the country of residence (74), educational background (51), and ethnicity (3).

Age. Participants' age is reported in multiple ways in the reviewed studies. It was either stated as average (39 studies), distribution of age brackets (16), the entire age range of all participants (13), a combination of average age and the entire range (27), the majority age range (5), a combination of the entire age range and the majority age range (9), the lower age threshold (4), or as the percentage of participants younger than an age threshold (1). Figure 1 summarizes the reported age values and depicts the diversity in reporting. Depending on the information reported in a paper, the figure uses a different visualization for each presented age distribution. The least detailed information is only a mean-average age value, visualized using a diamond symbol ♦. This symbol can be combined with other information, such as a lower- and upper-bound, shown as a black interval. When two intervals are reported in a paper, referring to a combination of the entire range and a majority range, the former is shown as a grey interval. For papers that report a mean age and a standard deviation, Figure 1 shows a normal distribution centered around the mean, extending to ±2 standard deviations, using a blue color. When, in addition, explicit lower and upper bounds are

Table 3. Overview of the Recorded User Characteristics Reported in the Papers and the Papers in which the Characteristics Were Analyzed

Type of Characteristic	Characteristic	Papers Recording Characteristic	# Studies Recording Characteristic	Papers Analyzing Characteristic	# Studies Analyzing Characteristic
Demographic	Age	[2, 3, 5, 7, 9, 12, 15, 19–22, 24, 26–28, 31, 34, 35, 39, 41, 42, 44, 46, 50–53, 55, 56, 59–61, 64, 65, 68–71, 74–76, 78–80, 83–86, 88, 91–93, 97, 100, 101, 103–107, 109, 110, 112, 117, 120, 121, 124–126, 128, 131, 132, 136, 138–144, 146–149, 151, 155, 157, 159]	114	[106, 107, 132, 149]	4
	Gender	[2, 3, 7, 9, 12, 15, 19–22, 24, 26–28, 31, 34, 35, 39, 41, 42, 44, 46, 50–53, 55, 56, 59–61, 68–71, 74–76, 78–80, 83–86, 88, 91–93, 96, 97, 100, 101, 103–107, 109, 110, 112, 120, 121, 124, 126, 128, 129, 132, 136, 138–144, 146, 148, 149, 157, 159]	105	[34, 70, 84, 106, 107, 132, 139, 149]	7
	Location	[1, 2, 4, 6, 7, 13, 17, 18, 20, 21, 23, 24, 26–28, 34, 44, 46, 50, 52, 53, 55, 58–61, 64–66, 68, 69, 75, 86, 106, 107, 109, 112, 114, 115, 117, 118, 124, 126, 128, 130–132, 138, 139, 141–144, 147, 148, 151, 152, 155, 157, 159]	74	[20]	1
	Education	[1, 2, 4–7, 9, 15, 17, 24, 26, 27, 27, 28, 35, 39, 41, 42, 46, 52, 53, 58, 66, 71, 79, 80, 84–86, 93, 96, 103, 106, 107, 114, 115, 120, 128, 129, 136, 141–143, 146, 152, 157, 159]	51	[84, 107, 139, 158]	4
	Other	[12, 15, 70, 71, 76, 80, 97, 110, 111, 113, 124, 126, 129, 132, 158, 159]	22	-	0
Personality	Need for Cognition	[24, 51, 53, 79, 88, 91, 92, 129]	11	[26, 51, 53, 79, 88, 91, 92]	8
	Openness	[2, 19, 75, 88, 92, 100, 101, 106]	10	[2, 75]	3
	Conscientiousness	[2, 19, 75, 92, 100, 101, 106]	7	[2, 75, 106]	4
	Extraversion	[2, 19, 75, 92, 100, 101, 106]	8	[2, 101, 106]	3
	Agreeableness	[2, 19, 75, 92, 100, 101, 106]	9	[2, 100, 101]	3
	Neuroticism	[2, 19, 75, 92, 100, 101, 106, 129]	8	[2, 75, 100, 101, 106]	5
	Innovativeness	[26, 53]	2	[26, 53]	2
	Rationality	[59, 103]	2	[59, 103]	2
	Social Awareness	[59–61, 126]	5	[59–61, 126]	5
	Propensity to trust others	[9, 53, 61, 80, 84, 106, 149]	7	[53, 106]	2
	Trust in Technology	[83]	2	[83]	1
	Valence	[69]	1	[69]	1
	Decision-Making Type	[34, 59–61, 65]	7	[34, 60, 61, 65]	7
Experience	Domain Experience	[7, 9, 12, 21, 26–28, 31, 35, 38, 39, 41, 42, 50, 53, 64, 68, 75, 76, 78, 83, 85, 88, 91, 92, 97, 101, 109, 111, 120, 129, 132, 136, 141, 149, 151, 155, 157, 159]	39	[50, 53, 75, 83, 91, 92, 132]	7
	Technical Expertise	[19, 20, 35, 39, 53, 64, 75, 80, 84, 91, 93, 96, 109, 120]	14	[26, 53, 64, 75]	4
	Visualization Literacy	[9, 26, 53, 60, 61, 75, 91]	8	[26, 53, 60, 61, 75, 91]	6

reported, the distribution is capped at these values and shown in yellow. The line segments shown in green represent explicit age brackets. Whenever no explicit lower- or upper bound is given for a range or a collection of age brackets, the figure caps the lower-most at 0 and the upper-most at 100. Since these values are not based on reported information, they are labeled 'open end' in the figure. The visualizations are sorted by (actual or inferred) lower bounds of the stated ranges.

Due to these differences, it is impossible to determine the average or standard age of the participants, but we can see certain tendencies. Most user studies reported an average age between 20 and 40 years. Underage participants are rarely covered, while studies including participants above 60 are more frequent; exact participant numbers are seldom explicitly specified for the upper age range. Instead, we often find all participants above the age of a certain threshold grouped together, and it is unclear where the overall age range ends. Thill et al. [132] reported 13 of 123 participants above the age of 50, and Wilkinson et al. [149] had 13 of 310 participants above the age of 55. In both cases, we cannot say if all the participants in this age bracket were 51 or 56, respectively, or if they were significantly older. These open brackets of the oldest participants vary in their starting age as well. While most start in the early 50s, Coba et al. [34], for example, start at the age of 41, and the oldest recruited participants in Shmaryahu et al. [120] are *"above 30"*.

Gender. We observed that most studies (93) reported the participants' gender in a binary way. We either see a statement of the number or percentage of participants identifying as one gender (45) or the number or percentage of both male and female participants (48). Out of these, seven studies reported participants who did not disclose their gender identity. Eleven studies reported options for non-binary, diverse, self-described, non-listed, or other gender identities. These participants are

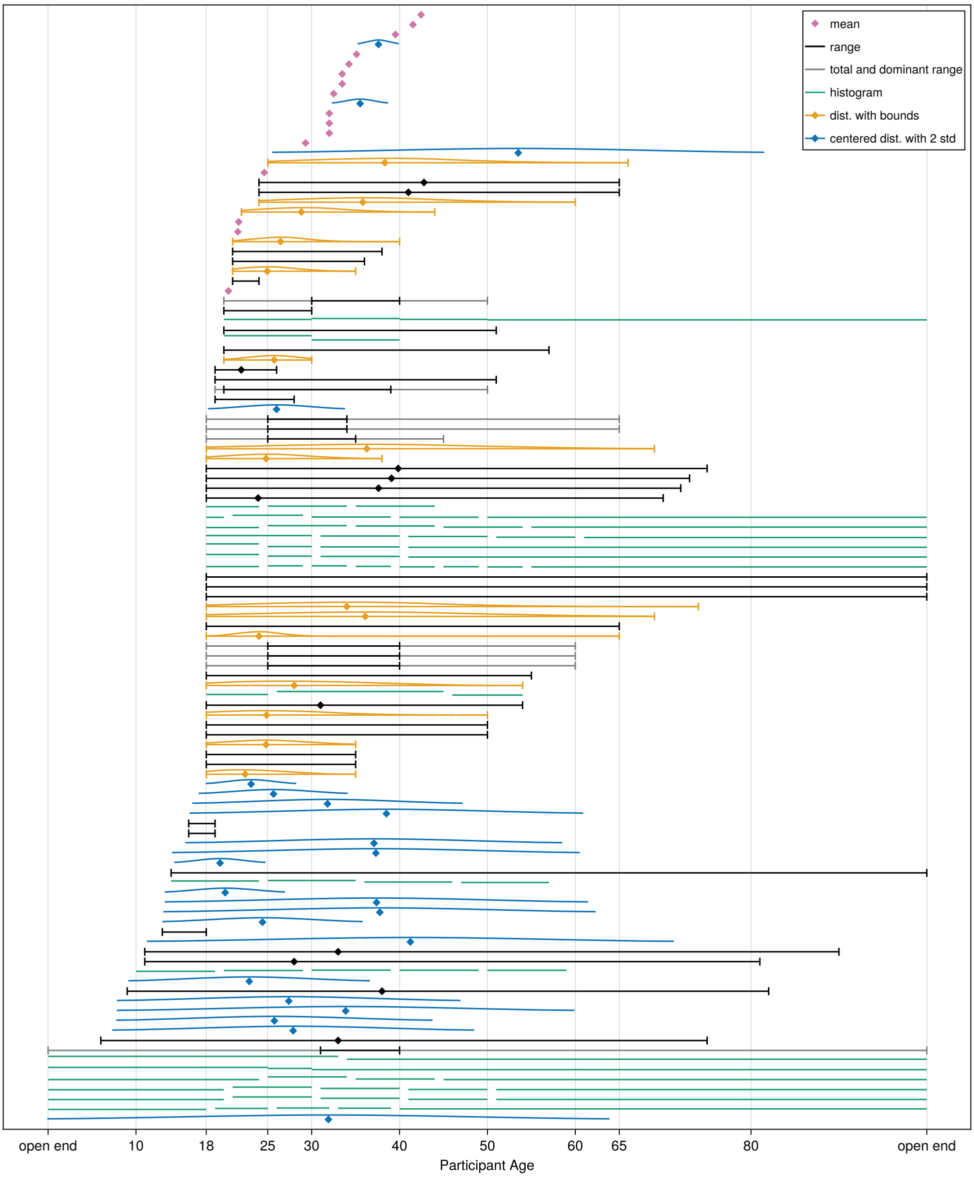

Fig. 1. Different ways of reporting the participants' age. Each visualization corresponds to how the information was reported in the papers, where one row represents one study sorted by lower-bound age. Black bars denote age ranges, and combined black and gray bars denote total and 'dominant' ranges. Blue curves denote centered distributions with two standard deviations around a mean. Yellow curves denote distributions with explicit lower- and upper bounds. Green lines represent explicit age histogram buckets, and diamonds ♦ represent mean values.

in the minority, with most studies reporting under 3% of the overall participants not identifying as male or female. Regarding the balance between male and female participants, we found that 55 of the 104 studies that reported the gender had a ratio within 10% difference in participant numbers. We found 49 studies that reported a difference between male and female participants higher than

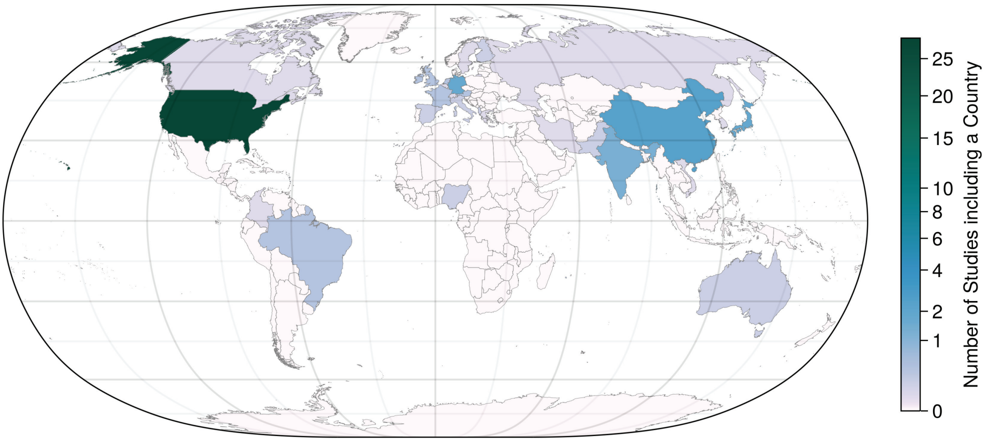

Fig. 2. Number of studies involving participants from different countries.

10% of the number of participants identifying with a binary gender; 30 studies had a male majority, while 19 had a female majority. One study [20] reported their gender distribution as "comparable distribution across countries in terms of gender" from which we were not able to determine the options they provided in their questionnaire.

Location. Seventy-four studies reported the location of the participants. These 8,446 participants came from 28 different countries, covering 40% of the total number of participants in our corpus. Figure 2 depicts the number of studies in which a participant's country of residence was mentioned. We notice that Africa, Central and South America, Eastern Europe, and South-East Asia are poorly represented. The USA is the country where, by far, most studies have recruited participants. Twenty-four studies recruited participants in the US, followed by China with six studies.

The participants for 24 studies were recruited from multiple countries. The highest number of countries stated in a paper was 25 countries [138]. This is the exception, though. In studies with multi-national participants, the majority recruited participants from two to five countries. However, not all papers reported the location of the participants on a country level but as aggregated. For example, Abdulrahman et al. [2] described the location of their participants on a broader regional level (Oceania, South-East Asia, Northern-Western Europe), Wibowo et al. [148] reported the number of participating countries. In contrast, Wang et al. [147] used a country descriptor (e.g., English-speaking countries).

We further analyzed the distribution of participants across the countries. Overall, we could extract the location for 7,646 participants, meaning that 800 participants were listed among "other" or the exact distribution was not reported in the paper. In terms of numbers, the USA dominates with 3,381 recruited participants. In the country with the second most participants, Japan, only 967 participants were recruited. In fact, the sum of participants from the countries with second to ninth most participants is 3,391, only 10 participants more than the US.

Education. The educational background of the participants is mainly associated with their highest degree. Out of the 53 studies in which the participants' educational background was addressed, 28 did not disclose any distribution. In most cases (22), the educational background is homogeneous as the researchers exclusively recruited students. The remaining six studies provide a general overview of the educational background, a lower threshold which the majority passed (*"almost all attained their high school diploma"* [131]), or a general description of the educational degrees *"they also have different education majors (IT, education, finance, sociology, literature, mathematics, biology, etc.), degrees (high school, associate degree, Bachelor, Master, PhD)" [27]*). All studies that reported information about the distribution predominantly recruited participants pursuing or having at least a bachelor's degree.

Other Demographics. We found the following demographic information that was only reported by a few studies: (1) the ethnic or cultural background of the participants, (2) their occupation or employment status, (3) the income level, (4) the participants' sexuality, (5) their marital status, (6) the household income, (7) the participants' body mass index, (8) their housing situation, and (9) whether they have a driver's license.

We found two papers that reported the distribution of the ethnic background of their participants [80, 129]. Both studies had predominantly Caucasian participants. A third study [143] stated that their participants had a diverse cultural background without going into more detail. Twelve studies recorded participants' occupations, but no information on the participants' distribution over the occupations was provided. The remaining demographics were recorded when they were relevant to the application domain. In online dating scenarios, studies reported the income level [124], sexuality [70, 71], and marital status of the participants [70, 71]. The household income was also recorded in an e-commerce setting [159]. The body mass index [97] and employment status [111] were recorded in studies in food recommendation settings. A study on electricity saving recommendations [126] recorded the housing situation of the participants, and a study collected information on the participants' driver's licenses on recommendations during autonomous driving [132].

4.1.2 Personality. We found 36 studies in which information about the participants' personalities was reported across 27 publications. Table 3 provides an overview of the recorded personality traits reported in our corpus. These traits, outlined below, are often evaluated using standardized questionnaires. The most frequent personality traits evaluated were the need for cognition, the so-called Big Five or OCEAN characteristics (openness, conscientiousness, extraversion, agreeableness, and neuroticism), and the participant's propensity to trust others.

Need for Cognition. **Need for Cognition (NFC)** is the *"Tendency for an individual to engage in and enjoy effortful cognitive activities"* [25]. Eleven studies (8 papers) stated that they measured the NFC of their participants. We found that NFC is frequently evaluated by splitting the participants' scores by the median to form two groups with high and low NFC, respectively. Hence, the reported distribution of participants is approximately equal. Millecamp et al. [92] was the only paper in which the average NFC score was reported instead of the distribution.

Big Five. Nine of the eleven studies that reported Big Five or OCEAN personality traits evaluated all five traits. The exceptions were Martijn et al. [88], who focused on openness, and Sun and Sundar [129], who recorded the anxiety level of their participants, which can be classified as part of the neuroticism trait.

We found different approaches to report the participant distribution. Two studies stated the mean and standard deviation of the respective scores [92, 106], both studies in Abdulrahman et al. [2] provided the distribution of participants in low, medium, and high score-regions of each trait, and Martijn et al. [88] performed a median split and grouped all participants in low and high openness groups. The remaining six studies did not provide details on the distribution.

Trust Propensity. Trust propensity is defined as *"Level of intensity of an individual's natural inclination to trust other parties in general"* [73]. Eight studies (8 papers) looked into this personality trait. However, we could only extract the distribution across participants for one paper.

Andjelkovic et al. [9] reported that the participants were approximately evenly distributed across high, medium, and low propensity scores.

Other Personality Traits. We identified six other personality traits for which no participant distributions were reported. Seven studies looked at the decision-making style of the participants. We found two different dimensions concerning participants' decision-making strategy in our corpus. Four studies from three papers classify their participants as rational or intuitive decision-makers. The remaining three studies split their participants into people who make decisions to maximize the outcome (maximizers) and satisficers who settle for the first choice that is good enough. We did not find further information on how the participants were distributed across the Social Awareness (four studies), Personal Innovativeness (two studies), Trust in Technology (one study), and Valence (one study) traits.

4.1.3 Experience. We identified 53 studies in which the participants' prior experience was recorded. The notions of experience that we observed can be grouped into three categories: domain experience, technical expertise, and visualization literacy.

Domain Experience. Domain experience is most commonly reported in our corpus. Forty-five studies included domain experience as an aspect of describing their participants. We found that the researchers had different focus areas and notions when measuring and reporting the domain experience of their participants. The knowledge about the domain can be assessed in certain domains with standardized questionnaires. Liang and Willemsen [78] and Martijn et al. [88], for example, both assess the musical sophistication of their participants measured by the Musical Sophistication Index [98]. Others assess the domain knowledge of their participants by having them report their years of working in the field [38, 64] or self-assess their experience on a scale [41, 85]. The level of the participants' experience is also measured by asking the participants how frequently they interact with the domain. This is often the case in everyday-life recommendation scenarios such as movie or music streaming [39, 76, 120, 136] or e-commerce [28, 155]. Overall, the papers reported that their participants are familiar with the evaluated domains and interact with them regularly. This does not necessarily apply to the familiarity with a specific application. Tsai and Brusilovsky [141] reported no participant had prior experience with the application.

Technical Expertise. The participants' technical expertise was assessed in 13 studies. The majority of the studies reported self-assessed technical expertise on varying scales. We additionally found that different aspects of the technical expertise were evaluated. Mishra et al. [93] and Jacobs et al. [64] asked their participants about their experience with AI or Machine Learning and reported a general familiarity with AI. Damak et al. [39] and Coba et al. [35] specifically asked about familiarity with recommender systems, while Shmaryahu et al. [120] combined both aspects and recorded if their participants have taken courses in deep learning, information retrieval, or recommender systems. These papers also reported a general familiarity with the aspects in question. Berkovsky et al. [19, 20] asked about the computer literacy of their participants. They reported that 90% of their participants have a high to very high computer literacy. Five studies did not provide information on the distribution of their participants regarding technical expertise.

Visualization Literacy. The visualization literacy of the participants was recorded in eight studies. However, only Millecamp et al. [91] reported the distribution of their participants between low (21 participants) and high (51 participants) visualization literacy.

4.2 The Impact of User Characteristics

We analyzed the findings reported in our corpus to investigate which user characteristics introduced in Section 4.1 might impact the effect of explanations. Overall, we extracted 165 findings from 31 papers in which the effect of explanations was evaluated on different user characteristics. As seen in Table 3, only a fraction of the papers that record a characteristic of their participants also report measurements on how the effect of explanations differs when disaggregating the participant data.

We present these results in Tables 4, 5, and 6 with different context variables[11]. More precisely, we present the evaluated explanation effects on different user characteristics using as context variable(s) the evaluated explanation type and explanation method in Table 4, the application domain in Table 5, and the type of recommender systems in Table 6. At first glance, the data is sparse for all combinations of effects and characteristics, making it impossible to determine a statistically significant impact of user characteristics on explanation effects. We can identify some indications that might lead to the generation of hypotheses that can be tested in further experiments.

4.2.1 *Demographics.*

Notably, the demographic information, the category recorded most frequently in Section 4.1.1, is only evaluated by a minority of the papers. The possible impacts of demographic characteristics on the effectiveness and efficiency of explanations were rarely evaluated.

We extracted nine findings from six papers that analyzed whether the participants' gender influenced the explanation effects. Differences between genders were found in user experience, transparency, and trust. No impact was found on persuasiveness, and perceived explanation quality has a slight majority of findings in which no effect was found. Due to the low number of studies, these can be merely seen as indications.

The number of findings and studies is even lower for the other user characteristics. Location, education, and age were only evaluated by one-to-two studies. Berkovsky et al. [20] assessed the impact of their participants' location on transparency and trust. We extracted four findings that indicate that the location impacts trust for participants from France, Japan, and the USA. At least for Japanese participants, this also seems to be the case for transparency. Two papers looked at the impact of the education level on transparency, trust, and usability/UX perception. Ma et al. [84] compared graduate with undergraduate students and found higher levels of transparency, trust, and usability/UX perception in undergraduate students. Zheng and Toribio [158] compared students with instructors and found mixed results on transparency. Regarding age, Nelekar et al. [106] found a correlation with persuasiveness, and Wilkinson et al. [149] found that younger participants perceive the explanation quality as higher than older participants.

4.2.2 *Personality.*

The impact of personality traits on explanation effects was evaluated by 18 papers from which we could extract 107 findings. In contrast to the demographics, the difference between the number of papers in which the characteristic was recorded and the number of papers in which it was evaluated is much smaller.

Most papers and results in the personality category were extracted for NFC. Six papers reported 17 results in which the impact of NFC was evaluated on all seven effects stated in Table 4. There seems to be a general tendency that no impact can be found, except for transparency and perceived explanation quality, where the results were mixed.

The Big Five traits were analyzed solely for a possible impact on perceived explanation quality and persuasiveness. Neuroticism was evaluated most frequently. Regarding persuasiveness, two papers found an impact of neuroticism; one did not. Both papers that looked into the impact of neuroticism on the perceived explanation quality did not find statistically significant results. The results for extraversion were similar; no impact was found on perceived explanation quality, but

[11] An interactive version of these tables is available at https://kathriwa.github.io/interactive-survey-visualization/#/characteristics

Table 4. Number of Results and Publications in which the Effect of User Characteristics on Explanation Effects was Analyzed with the **Explanation Type** and **Explanation Generation Method** as Context Variables

	User Characteristic	Effect Found	Effectiveness	Efficiency	Perceived Explanation Quality	Persuasiveness	Transparency	Trust	Usability/UX	Total
Demographic	Age	✓			[149]$^{t}_{p}$	[106]$^{h}_{-}$				2
		×								
	Location	✓					[20]$^{ht}_{p}$	[20]$^{ht}_{p}$		6
		×								
	Gender	✓			[70]$^{t}_{p}$		[84]$^{h}_{i}$	[84]$^{h}_{i}$	[84]$^{h}_{i}$, [139]$^{t}_{-}$	9
		×			[149]$^{t}_{p}$, [34]$^{h}_{-}$	[106]$^{h}_{-}$				
	Education	✓					[158]$^{t}_{-}$		[84]$^{h}_{i}$	6
		×					[84]$^{h}_{i}$, [158]$^{t}_{-}$	[84]$^{h}_{i}$		
Personality	Agreeableness	✓			[100]$^{t}_{-}$					4
		×			[101]$^{t}_{-}$	[2]$^{h}_{-}$				
	Conscientious-ness	✓			[75]$^{tv}_{i}$	[2, 106]$^{h}_{-}$				4
		×				[2]$^{h}_{-}$				
	Decision-Making Type	✓	[61]$^{stv}_{p}$, [60]$^{hv}_{p}$	[60]$^{hv}_{p}$	[61]$^{stv}_{p}$, [60]$^{hv}_{p}$, [34]$^{h}_{-}$		[61]$^{stv}_{p}$	[61]$^{stv}_{p}$, [60]$^{hv}_{p}$	[61]$^{v}_{p}$	17
		×			[34]$^{h}_{-}$		[60]$^{hv}_{p}$			
	Extraversion	✓				[106]$^{h}_{-}$				4
		×			[101]$^{t}_{-}$	[2]$^{h}_{-}$				
	Need for Cognition	✓			[88]$^{h}_{p}$		[51]$^{t}_{-}$			17
		×	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[79]$^{t}_{-}$, [92]$^{v}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[51]$^{t}_{-}$, [26]$^{hv}_{-}$, [53]$^{h}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	
	Neuroticism	✓				[106]$^{h}_{-}$,[75]$^{t}_{i}$				7
		×			[100, 101]$^{t}_{-}$	[2]$^{h}_{-}$				
	Openness	✓			[75]$^{tv}_{i}$					3
		×				[2]$^{h}_{-}$				
	Personal Innovativeness	✓	[26]$^{hv}_{-}$	[26]$^{hv}_{-}$			[53]$^{h}_{-}$		[26]$^{hv}_{-}$	13
		×	[53]$^{h}_{-}$	[53]$^{h}_{-}$		[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[26]$^{hv}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	
	Trust Propensity	✓						[106]$^{h}_{-}$	[53]$^{h}_{-}$	8
		×	[53]$^{h}_{-}$	[53]$^{h}_{-}$		[53]$^{h}_{-}$	[53]$^{h}_{-}$	[53]$^{h}_{-}$	[53]$^{h}_{-}$	
	Rationality	✓								3
		×			[59]$^{t}_{-}$			[103]$^{t}_{-}$		
	Social Awareness	✓	[61]$^{stv}_{p}$, [60]$^{hv}_{p}$		[61]$^{stv}_{p}$, [60]$^{hv}_{p}$	[61]$^{stv}_{p}$	[61, 126]$^{stv}_{p}$, [60]$^{hn}_{p}$	[59]$^{t}_{-}$, [61]$^{stv}_{p}$, [60]$^{hv}_{p}$		15
		×		[60]$^{hv}_{p}$	[59]$^{t}_{-}$					
	Trust in Technology	✓					[83]$^{v}_{i}$			1
		×								
	Valence	✓			[69]$^{v}_{p}$					1
		×								
Experience	Domain Experience	✓					[83]$^{v}_{i}$			11
		×	[53]$^{h}_{-}$	[53]$^{h}_{-}$	[75]$^{tv}_{i}$	[53]$^{h}_{-}$, [75]$^{tv}_{i}$	[53]$^{h}_{-}$	[53]$^{h}_{-}$	[53]$^{h}_{-}$	
	Technical Expertise	✓	[64]$^{tv}_{-}$	[26]$^{hv}_{-}$		[75]$^{t}_{i}$				16
		×	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[53]$^{h}_{-}$	[75]$^{tv}_{i}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	
	Visualization Literacy	✓	[53]$^{h}_{-}$	[53]$^{h}_{-}$	[60]$^{v}_{p}$,[75]$^{v}_{i}$		[53]$^{ht}_{-}$	[53]$^{h}_{-}$	[53]$^{h}_{-}$, [61]$^{stv}_{p}$	18
		×	[26]$^{hv}_{-}$	[26]$^{hv}_{-}$	[75]$^{v}_{i}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	[26]$^{hv}_{-}$	[26]$^{hv}_{-}$	[26]$^{hv}_{-}$, [53]$^{h}_{-}$	
	Total (number of results)	✓	7	4	21	9	13	12	10	
		×	8	8	18	18	13	12	14	

Rows with ✓indicate that an impact of the user characteristic was found; rows with × indicate that no impact was found. The last column sums the number of findings for each user characteristic. The last row sums up the findings where an effect was found or not found. The superscript letter indicates the explanation type: t: textual, v: visual, h: hybrid, s: schematic (i.e., a table), and n: numerical. Findings can contain multiple explanations of different types that were grouped together. Subscript letters show the explanation method: p: post-hoc, i: intrinsic. Articles where information on any of the categories was not provided or the recommendations were simulated are marked with –.

for persuasiveness, one paper found an impact while another did not. Kouki et al. [75] indicated that conscientiousness and openness influence the perceived explanation quality. However, it was the only paper in our corpus analyzing this combination of variables. Regarding persuasiveness, no impact was found for openness, while the analyses of conscientiousness returned mixed results. Only one of three studies found an impact of agreeableness on perceived explanation quality. Only one study evaluated the impact of agreeableness on persuasiveness without finding evidence.

Five studies examined the effect of social awareness on transparency, trust, effectiveness, efficiency, persuasiveness, and perceived explanation quality. All studies evaluating transparency, trust, effectiveness, and persuasiveness found that the results differ depending on the participants' social awareness scores. The one study that evaluated a possible interaction of social awareness and efficiency did not find a significant effect [60]. The impact of social awareness on perceived explanation quality was found in four studies [60, 61] but not in one study [59]. Personal innovativeness was analyzed by two papers [26, 53]. Both results contradicted for effectiveness, efficiency, transparency, and usability/UX. Both papers did not find an impact on trust and persuasiveness. The two papers that evaluated the propensity to trust others [53, 106] did not find an impact on effectiveness, efficiency, persuasiveness, and transparency. The results were contradictory for trust and usability/UX. No effect was found for rationality on perceived explanation quality and trust. The only evaluation of trust in technology found that it can impact the transparency perceived by the participants [83]. One study [69] found an impact of valence on perceived explanation quality. The potential impact of the decision-making strategy was evaluated in three papers [34, 60, 61] on all effects but persuasiveness. Hernandez-Bocanegra and Ziegler [60, 61] found impacts on usability/UX perception, effectiveness, efficiency, and trust. Perceived explanation quality and transparency yielded mixed results.

4.2.3 Experience. The prior experience of the participants was recorded by 61 studies out of which 17 analyzed whether it impacts the explanation effect being evaluated. The difference in the number of papers recording the characteristic and analyzing it is lower for the experience category than in the demographic but higher than in the personality category.

Seven studies evaluated if the domain experience of their participants had an impact on all of the explanation effects listed in Table 3. Contrasting results were found for transparency; evaluations for all other effects did not find significant results. Four studies evaluated the impact of technical expertise on the explanation effect. There was no impact on perceived explanation quality, transparency, trust, and usability/UX. Some studies found an impact of technical expertise on effectiveness, efficiency, and persuasiveness; others did not. Six studies looked into the impact of visualization literacy on the explanation effects. No significant impact was found for persuasiveness; all other six effects had studies in which an impact was found and those in which it was not.

4.2.4 Explanation Context. We looked at four additional variables that provide context to the results: explanation type, explanation generation method, recommender type, and the application domain in which the explanation was evaluated. In terms of explanation type, we found 68 results in which multiple explanations of different types were combined in the result. We cannot determine any patterns in instances where the result addresses only one explanation. We often see that an impact of a user characteristic was found and not found for the same explanation type (see the trust or usability/UX effects for the trust propensity user characteristic in Table 4 for an example), but no explanation type seems to be more likely to evoke an effect. The explanation generation method is predominantly either unknown or the explanations were simulated for the user evaluation. Out of 165 findings, 103 are based on these explanations. Intrinsically explainable recommender systems are rare in this evaluation. Only 21 of the findings stem from a model intrinsic explanation such that a comparison between intrinsic and post-hoc explanations is not possible. Along with the explanation type and generation method, we investigated the results presented in Table 4 in the context of the application domain in which the explanations were evaluated (Table 5) and the type of recommender system that was explained (Table 6). We identified a total of nine different application domains that were evaluated, with point-of-interest recommendations being the most common (44 findings), followed by document (39 findings), and social recommendations (27 findings). We notice the dominance of some domains in certain user characteristics.

Table 5. Number of Results and Publications in which the Effect of User Characteristics on Explanation Effects was Analyzed with the **Application Domain** as a Context Variable

	User Characteristic	Effect Found	Effectiveness	Efficiency	Perceived Explanation Quality	Persuasiveness	Transparency	Trust	Usability/UX	Total
Demographic	Age	✓			[149] 🎬	[106] ⚕				2
		×								
	Location	✓					[20] 🎬	[20] 🎬		6
		×								
	Gender	✓			[70] ♥		[84] 🎓	[84] 🎓	[84] 🎓, [139] ⦿	9
		×			[149] 🎬, [34] ⦿	[106] ⚕				
	Education	✓					[158] 🎓		[84] 🎓	6
		×					[84, 158] 🎓	[84] 🎓		
Personality	Agreeableness	✓			[100] ⦿					4
		×			[101] ⦿	[2] ⚕				
	Conscientious-ness	✓			[75] 🎼	[2, 106] ⚕				4
		×				[2] ⚕				
	Decision-Making Type	✓	[60, 61] ⦿	[60] ⦿	[34, 60, 61] ⦿		[61] ⦿	[60, 61] ⦿	[61] ⦿	17
		×			[34] ⦿		[60] ⦿			
	Extraversion	✓				[106] ⚕				4
		×			[101] ⦿	[2] ⚕				
	Need for Cognition	✓			[88] 🎼		[51] ⚕			17
		×	[26] 👥, [53] 🗎	[26] 👥, [53] 🗎	[79] 🛒, [92] 🎼	[26] 👥, [53] 🗎	[51] ⚕, [26] 👥, [53] 🗎	[26] 👥, [53] 🗎	[26] 👥, [53] 🗎	
	Neuroticism	✓				[106] ⚕,[75] 🎼				7
		×			[100, 101] ⦿	[2] ⚕				
	Openness	✓			[75] 🎼					3
		×				[2] ⚕				
	Personal Innovativeness	✓	[26] 👥	[26] 👥			[53] 🗎		[26] 👥	13
		×	[53] 🗎	[53] 🗎		[26] 👥, [53] 🗎	[26] 👥	[26] 👥, [53] 🗎	[26] 👥, [53] 🗎	
	Trust Propensity	✓						[106] ⚕	[53] 🗎	8
		×	[53] 🗎	[53] 🗎		[53] 🗎	[53] 🗎	[53] 🗎	[53] 🗎	
	Rationality	✓								3
		×			[59] ⦿			[103] 🛒		
	Social Awareness	✓	[60, 61] ⦿		[60, 61] ⦿	[61] ⦿	[126] 💡, [60, 61] ⦿	[59–61] ⦿		15
		×		[60] ⦿	[59] ⦿					
	Trust in Technology	✓					[83] 🎬			1
		×								
	Valence	✓			[69] 🎬					1
		×								
Experience	Domain Experience	✓					[83] 🎬			11
		×	[53] 🗎	[53] 🗎	[75] 🎼	[53] 🗎, [75] 🎼	[53] 🗎	[53] 🗎	[53] 🗎	
	Technical Expertise	✓	[64] ⚕	[26] 👥		[75] 🎼				16
		×	[26] 👥, [53] 🗎	[53] 🗎	[75] 🎼	[26] 👥, [53] 🗎	[26] 👥, [53] 🗎	[26] 👥, [53] 🗎	[26] 👥, [53] 🗎	
	Visualization Literacy	✓	[53] 🗎	[53] 🗎	[60] ⦿,[75] 🎼		[53] 🗎	[53] 🗎	[53] 🗎, [61] ⦿	18
		×	[26] 👥	[26] 👥	[75] 🎼	[26] 👥, [53] 🗎	[26] 👥	[26] 👥	[26] 👥, [53] 🗎	
	Total (number of results)	✓	7	4	21	9	13	12	10	
		×	8	8	18	18	13	12	14	

Rows with ✓indicate that an impact of the user characteristic was found; rows with × indicate that no impact was found. The last column sums the number of findings for each user characteristic. The last row sums up the findings where an effect was found or not found. Application domains are depicted as follows: 🗎: document, 🛒: e-commerce, 🎓: education, 💡: energy saving, ⚕: health, 🎬: movie, 🎼: music, ♥: online dating, ⦿: point-of-interest, 👥: social.

The decision-making type and social awareness, for example, were almost exclusively evaluated in the point-of-interest domain and trust propensity and the education level of participants in the document and education domain, respectively. The dominant type of recommender system is content-based, with 107 out of 165 findings. Pure collaborative filtering systems are only used to generate the recommendations for 10 findings.

5 Discussion

Overall, we observe that the data about the participants varies between the publications in terms of completeness. Table 3 illustrates that none of our analyzed characteristics were recorded by all

Table 6. Number of Results and Publications in which the Effect of User Characteristics on Explanation Effects was Analyzed and the **Type of Recommender System** as Context Variable

	User Characteristic	Effect Found	Effectiveness	Efficiency	Perceived Explanation Quality	Persuasiveness	Transparency	Trust	Usability/UX	Total
Demographic	Age	✓			[149]*cb*	[106]–				2
		×								
	Location	✓					[20]–	[20]–		6
		×								
	Gender	✓			[70]*cb*		[84]*cf*	[84]*cf*	[84]*cf*, [139]–	9
		×			[149]*cb*, [34]–	[106]–				
	Education	✓					[158]*hy*		[84]*cf*	6
		×					[84]*cf*, [158]*hy*	[84]*cf*		
Personality	Agreeableness	✓			[100]–					4
		×			[101]–	[2]*cb*				
	Conscientious-ness	✓			[75]*hy*	[106]–, [2]*cb*				4
		×				[2]*cb*				
	Decision-Making Type	✓	[60, 61]*cb*	[60]*cb*	[60, 61]*cb*, [34]–		[61]*cb*	[60, 61]*cb*	[61]*cb*	17
		×			[34]–		[60]*cb*			
	Extraversion	✓				[106]–				4
		×			[101]–	[2]*cb*				
	Need for Cognition	✓			[88]*hy*		[51]–			17
		×	[26, 53]*cb*	[26, 53]*cb*	[79]*cbcf*, [92]–	[26, 53]*cb*	[51]–, [26, 53]*cb*	[26, 53]*cb*	[26, 53]*cb*	
	Neuroticism	✓				[106]–,[75]*hy*				7
		×			[100, 101]–	[2]*cb*				
	Openness	✓			[75]*hy*					3
		×				[2]*cb*				
	Personal Innovativeness	✓	[26]*cb*	[26]*cb*			[53]*cb*		[26]*cb*	13
		×	[53]*cb*	[53]*cb*		[26, 53]*cb*	[26]*cb*	[26, 53]*cb*	[26, 53]*cb*	
	Trust Propensity	✓						[106]–	[53]*cb*	8
		×	[53]*cb*	[53]*cb*		[53]*cb*	[53]*cb*	[53]*cb*	[53]*cb*	
	Rationality	✓								3
		×			[59]–			[103]–		
	Social Awareness	✓	[60, 61]*cb*		[60, 61]*cb*	[61]*cb*	[126]–, [60, 61]*cb*	[59]–, [60, 61]*cb*		15
		×		[60]*cb*	[59]–					
	Trust in Technology	✓					[83]*cf*			1
		×								
	Valence	✓			[69]*cb*					1
		×								
Experience	Domain Experience	✓					[83]*cf*			11
		×	[53]*cb*	[53]*cb*	[75]*hy*	[53]*cb*, [75]*hy*	[53]*cb*	[53]*cb*	[53]*cb*	
	Technical Expertise	✓	[64]–	[26]*cb*		[75]*hy*				16
		×	[26, 53]*cb*	[53]*cb*	[75]*hy*	[26, 53]*cb*	[26, 53]*cb*	[26, 53]*cb*	[26, 53]*cb*	
	Visualization Literacy	✓	[53]*cb*	[53]*cb*	[60]*cb* ,[75]*hy*		[53]*cb*	[53]*cb*	[53, 61]*cb*	18
		×	[26]*cb*	[26]*cb*	[75]*hy*	[26, 53]*cb*	[26]*cb*	[26]*cb*	[26, 53]*cb*	
	Total (number of results)	✓	7	4	21	9	13	12	10	
		×	8	8	18	18	13	12	14	

Rows with ✓indicate that an impact of the user characteristic was found; rows with × indicate that no impact was found. The last column sums the number of findings for each user characteristic. The last row sums up the findings where an effect was found or not found. The recommender type is shown after the domain and indicated by *cb* for content-based, *cf* for collaborative filtering, and *hy* for hybrid systems. Articles where information on any of the categories was not provided or the recommendations were simulated are marked with –.

papers in the corpus. Reporting participants' age, gender, location, education, and prior domain experience seems standard practice. However, disaggregating the data by these features to see if the evaluated effect differs between sub-groups is rare.

5.1 Who Was Recruited to Evaluate Recommender Systems Explanations?

In this subsection, we discuss our results to answer **RQ 1**. Our analyses demonstrate that it is challenging to aggregate the data in our corpus due to non-uniform ways of measuring and reporting it. Therefore, we cannot conclusively determine an average participant for each application domain. To the best of our knowledge, there is no related work that investigated how representative study participants are in recommender systems evaluation, which leaves us without a reference point to determine whether the reported participant data in our corpus is sampled accordingly. Instead, we focus our discussion on the characteristics reported by at least one-quarter of the papers in our corpus. We specifically discuss the participants' age, gender, location, education, and prior domain experience.

Age. Regarding the participants' age, we made two main observations. First, almost all participants are above the age of 18. One possible explanation for this might be age restrictions for crowd workers. Our second observation is that the average age of the participants does not seem to reflect the average age of the overall population. In our corpus, the age averages are predominantly between the ages of 20 and 40, while the average age of the overall population, especially in the countries where the participants were recruited, is shifting towards an older average. A possible reason for this could be the recruiting method. A large proportion of the papers in our corpus recruited their participants on crowdsourcing platforms. The findings of Difallah et al. [40] support our observation, as they found that MTurk workers tend to be younger than the overall population. In their study, conducted in 2018, 60% of the workers were born after 1980, which also aligns with our findings. Studies that did not rely on crowdsourcing platforms frequently recruited participants at universities. In such cases, students often comprised most of the sample. This can be another explanation for the comparatively low age average.

Gender. The difference between male and female participants was balanced in about half of the papers that reported the participant distribution for gender. For the other half of the papers, a male majority was more common than a female majority. We found very few papers report non-binary options in their gender distribution. This does not necessarily imply that these options were not available in the study, but it could also mean that all participants identified as male or female, and the possibilities with no answer were omitted from the paper. Especially with the demographic data, we rarely had access to the questions and response options given in the study. As the exact questions are not provided, no conclusion can be drawn about the options provided.

Location. The results for the participants' location show a clear dominance of participants living in the USA. Again, the recruiting mechanism might impact the location as Amazon Mechanical Turk is a frequently used crowdsourcing platform in our corpus with a predominantly US population [40].

Education. The results regarding the education level of the participants show that the majority of the participants evaluating recommendation explanations are highly educated. Participants with at least a bachelor's degree were the dominant group in our sample. The average participant has a higher education level than the average population.

Domain Experience. Regarding prior experience with the domain, the results show that most participants frequently interact with the application domain. One reason might be that the evaluated domains are typically accessible to various people and predominant in the users' leisure time. The most common are point-of-interest (e.g., restaurants or tourist attractions), e-commerce, movie, or music recommendations. Another reason for the dominance of these domains might be their popularity in the recommender systems community. Chin et al. [29] analyzed popular datasets for recommender systems evaluation, which almost exclusively came from these domains.

The domain experience might also be related to the participant's age group, assuming that younger people who grew up with online shopping and media streaming services interact with these applications more frequently than the average above 60-year-old person.

Summary. In conclusion, the results of the frequently recorded characteristics show that the most common participant is an educated male or female in their twenties or thirties, living in the USA, and regularly interacting with everyday recommender systems. Due to the low number of papers that reported the participant distribution for the remaining characteristics we analyzed (10% or less of the total corpus size), we cannot draw other observations about other demographic characteristics or their personality traits. However, the characteristics we have provided enough information to match the WEIRD societies outlined in Henrich et al. [57].

5.2 What User Groups Were Not Represented?

After discussing what we know about the participants on which explanations are evaluated, we now focus on possible blind spots where user characteristics are underrepresented or missing.

Overall, for each of our analyzed characteristics, we found many papers that did not report recording it. Therefore, we only refer to the papers in which the characteristic was reported when pointing out a gap. We cannot completely rule out that these user groups were part of the participant sample of the papers that did not disclose the information.

Regarding participant diversity, the papers in our corpus did not include people from the Global South and Eastern Europe, a significant share of the population. This issue has been pointed out in multiple publications such as Linxen et al. [82], Sturm et al. [127] and is generally known in the HCI community. Given the culturally-dependent differences in user interaction [116], this may have a profound implication on the generalizability of the results to global settings.

Children, teenagers, and older people were rarely part of the participant sample. People in these age groups might have different needs than younger adults regarding explanation content and display, so it would be interesting to see an evaluation of the same explanation that includes these user groups. Ekstrand et al. [48], for example, looked at the recommendation effectiveness and found that the recommendation accuracy differs when disaggregating the data by age and gender. While one might argue that children only start independently interacting with most recommender systems applications from a certain age, this might change. Furthermore, investigating explainable recommender systems in use cases where parents use them together with their children opens up a new research direction.

The comparisons regarding the participants' education level mainly investigate differences between participants with an above-average education level and participants with an even higher education level (e.g., undergraduate compared to graduate students as seen in Ma et al. [84]). Participants without an academic degree are in the minority in most studies. Instead of comparing participant groups where the overall education level is high already, an evaluation of the effect of explanations on participants with an average or below-average education level would be interesting to see.

Regarding gender, we found only a few studies reporting non-binary participants. As mentioned in Section 5.1, most papers did not report any gender other than male/female.

Only a small fraction of our corpus reported the personality traits of their participants. However, we cannot identify user groups that might have been left out due to how these traits are reported. In most cases, we do not know anything about the participant distribution. We noticed, however, that the participants are frequently split into high and low-scoring, often based on a median split.

In terms of prior experience, we mainly found tech-savvy participants with domain experience in the studies of our corpus. The technical expertise might be related to recruiting participants online, who are often experienced crowd workers, which requires a certain level of expertise. However, in real life, many recommender systems applications also target less experienced users or might not know that an algorithm selects the items they see. Although recruiting these participants might be more challenging, depending on the application domain, it could be relevant to find out how findings generalize to the general populations with all kinds of domain experience.

We do not have enough information about the participant distributions of the remaining characteristics to infer user groups that were not represented. As of today, we do not know if these characteristics might impact the explanation effects, so we encourage reporting participant distributions and exploring possible impacts on explanation effects in future work.

5.3 Which User Characteristics Impact the Effect of Explanations in Recommender Systems?

In this subsection, we discuss the results of Section 4.2 and answer **RQ 2**. Generally, the number of findings we could extract in which the impact of user characteristics on the explanation effect was evaluated is very low. In our corpus, we have at most five studies that evaluated the same characteristic and the same effect (impact of decision-making strategy on perceived explanation quality). It is far more frequent that only one or two studies examined the same characteristic-effect combination. Therefore, it is not possible to derive significant results from our data. Results, where multiple articles either all found or did not find an impact of a user characteristic on an effect, might indicate a tendency that can be formulated into hypotheses for further experimentation. Table 4 shows the gaps the community should close and allows us to spot these indications easily. Looking at *social awareness*, for example, all three articles that investigated the impact on transparency and trust and both articles that looked into effectiveness found an effect. This can be taken as motivation for further research that systematically evaluates the impact of social awareness on these explanation effects.

In the instances where an impact was observed in one study and not found in another, other factors than the user characteristics might have impacted the results. If we look, for example, at the impact of extraversion on the persuasiveness of the explanations, Nelekar et al. [106] found an effect while Abdulrahman et al. [2] did not. The work of Nelekar et al. [106] replicates the study performed in Abdulrahman et al. [2], focusing on participants from India. While their participants had a similar average age and education level, they were from different locations. Additionally, the participant sample in Nelekar et al. [106] had a male majority (39 male majority, 21 female, one other) while the participants in Abdulrahman et al. [2] were predominantly female (21 male, 48 female). Both of these aspects could explain the different results.

Most cases in which the explanation effects findings were inconclusive were independent studies evaluating different aspects. Looking at the impact of domain knowledge on transparency, Loepp et al. [83] found an effect, and Guesmi et al. [53] did not. We do not know a lot about the participants in Loepp et al. [83] other than that they are mainly female students with an age average about ten years lower than the participants in the study by Guesmi et al. [53], who recruited researchers with at least one scientific publication. Furthermore, both papers differ in the application domain, evaluated explanations, and research goals. Loepp et al. [83] propose a new explainable recommendation method, which they evaluate in the movie domain, while Guesmi et al. [53] investigate different detail levels of explanations in the document recommendation domain. Both studies evaluated different explanations with different modalities, complexities, and interaction features. All these aspects could have led to the differences in results, but the differences could also stem from factors that have not been considered yet.

We also found, for example, ambiguous results within the same study. This can occur when there are two different studies in one paper. Coba et al. [34], for example, conducted an online study in which they did not find an impact of the participants with a maximizing decision-making strategy on the perceived explanation quality. In their eye-tracking study, such an impact was found, though. The differences in results can be caused by the different participants and the change in study types or how the effect was measured (questionnaire compared to eye-tracking recording).

In other instances, the different results are due to different independent variables in the evaluation. We take this as an indication that other factors must be considered when interpreting the results.

In this survey, we looked at four potential factors, namely, the *explanation modality*, the *explanation generation method*, the *application domain* in which the explanation was evaluated, and the *type of recommender system*. Explanations might affect a user depending on their modalities or complexity, as different users might have other preferences. We could not observe any general patterns that certain explanation types are more likely to impact the results from our corpus. Possible reasons could be the small number of overall results or the necessity for a more fine-grained categorization of the explanation types. The impact of a user's agreeableness on the perceived explanation quality, for example, was investigated by the same research team in two separate publications [100, 101] focusing on privacy concerns regarding the information disclosed in the explanation. Both use text explanations in a group recommendation setting applied to the point-of-interest domain. The main difference in the explanations that arise from the articles is that the explanation in Najafian et al. [100] is static and discloses potentially private information, while the explanation in Najafian et al. [101] is presented in a chat interface that allows the user to interactively decide how much information is revealed to the group. Given the small data sample that we have for these results, applying a more fine-grained classification of explanation types would further increase the sparsity of the data.

The articles of our corpus often did not disclose information on the underlying recommender system and explanation generation approach. One possible reason for this could be the focus of the articles. Describing the setup of a user study and the characteristics of the participants in sufficient detail requires space, which is usually restricted, especially in conference publications. This might have led to less detailed descriptions of the explanation method and underlying recommender systems. We furthermore identified a set of articles in which the recommendations and corresponding explanations were not generated by an actual recommender system but instead simulated for the user study (see [34, 59, 80, 100, 121, 148], for example). These publications usually focus on investigating how different explanations impact the users and are often published in human-computer-interaction-centered venues such as CHI, UMAP, or IUI. In future work, it would be interesting to see if these proposed explanations and results can be replicated with existing explanation methods. In the articles that provided information on the recommendation approach, we found a dominance of content-based and hybrid methods. This could indicate a trend toward providing information in the explanations that go beyond a user's previous interactions or user-/item-similarity. Balog and Radlinski [13], for example, showed that their baseline consisting of a short movie description is often perceived en par, if not better, than human-generated explanations for movie recommendations, indicating that providing additional information on the recommended items can be useful to achieve certain explanation goals.

In terms of application domains, the majority carries a low risk when following recommendations without understanding them. The number of findings measured in a point-of-interest domain is higher than the number in movie recommendations, but overall, the distribution of domains could be correlated with the domains of the typical datasets used in recommender systems evaluation [29]. We further observe that only a few user characteristics were evaluated in three or more domains, namely gender, need for cognition, neuroticism, domain experience, technical expertise, and visualization literacy. For other characteristics, we either only have a small number of results or the characteristic was only evaluated in one or two domains. One reason why the diversity of domains is higher for experience characteristics than for personality could be that capturing the personality of the participants usually goes along with additional questionnaires such as the Big Five, which might increase the risk of survey fatigue. As pointed out in Section 4.1.3, while there are standardized questionnaires to measure the experience for some domains, it is mostly captured by one item in the questionnaire.

After categorizing the explanations, we still find different results for explanations with the same properties according to this classification, sometimes even from the same article. Kouki et al. [75], for example, compared 11 different explanations and found that the visualization literacy of the participants impacts the quality perception of some explanations while others do not. Both explanations are visual explanations generated by an intrinsically explainable hybrid recommender system. While an effect was found for their Venn diagram explanation, it was not found for their cluster dendrogram. We leave investigating the detailed differences of such occurrences to determine the cause of these differences for future work.

When examining the papers, we noticed differences in how the effects were defined and measured. For the effectiveness of explanations, for example, we found papers in which the participant was asked to rate an explanation's or system's effectiveness. In contrast, other papers measured it implicitly (e.g., by evaluating the participants' decision quality). Further specifying the effect that was measured might show a more precise picture in the cases where mixed results were found.

5.4 What Are the Implications of Not Measuring the Impact of User Characteristics on Explanations?

Fifteen papers in our corpus did not report any of the user characteristics we looked at in this work, which harms the reproducibility of the results as it is impossible to know whether the recruited participants are similar to the ones in the original study. Another aspect that makes reproducing and comparing the reported results between the individual papers challenging is how they are described. Figure 1 illustrates the example of the participant's age and how different the characteristics are reported in our corpus. This applies to demographics, participants' prior experience, and certain personality traits where the scales and calibration differ between studies.

In our survey, we only found a few studies in which the impact of user characteristics on explanation effects was evaluated. In this section, we discuss the possible implications of not measuring the impact of user characteristics on explanation effects.

We do not know which user characteristics can influence the effect of explanations on a user. Therefore, we also do not know what information we need about the users of an application to provide them with an explanation that has the desired effect. Does it matter what age group the users are? Do users in their 20s have a different perception of the explanations than users in their 30s? Does the perception change with a bigger age gap? Despite having the desired effect on most participants, we cannot rule out that the evaluated explanation negatively impacts the user experience of a substantial part of the target users by only assessing the effect of explanations on all participants.

We further noticed that the user characteristics are often evaluated in isolation. We only found rare occurrences in which multiple user characteristics are grouped. In fact, only one publication in our corpus disaggregated their results based on two demographic categories. Noorbehbahani and Zarein [107] grouped gender with age and evaluated if the explanation evokes a different quality perception of the recommender system between genders in different age groups and education levels. In all cases, they found an impact on the evaluated explanation effects. Still, it is unclear to what extent forming these subgroups based on multiple characteristics is meaningful and which characteristics should be grouped in the evaluation as we only have this one paper in our corpus.

Knowing which user characteristics to pay attention to could also lead to a more effective and efficient evaluation. The participants can be targeted more precisely, and the data that needs to be collected about the participants could be narrowed down to the essential aspects.

Ultimately, finding out which user characteristics matter in the evaluation process of explanations requires a bigger focus on reproducibility studies. In our corpus of 124 papers, only Nelekar et al. [106] reproduced a study. Comparing the setups of both the original study and the reproduced one might lead to insights on aspects that explain the differences in results. However, to get to this point, effort needs to be put into increasing the comparability of the studies.

Given our findings, we provide a set of recommendations for the field to be able to move in this direction.

5.5 Recommendations

Finally, we derive the following recommendations from the results regarding the participant recruiting, reporting of their data, and the evaluation of explanation effects in recommender systems.

5.5.1 Participant Recruiting. In terms of the recruitment process, we recommend a *greater awareness of who is being recruited and whether this sample is representative of the target audience of the application domain.* Furthermore, the *shortcomings of the participant pool in crowdsourcing platforms need to be considered and addressed when opting for this type of recruiting.* We suggest including a statement in the publication on what efforts were undertaken to ensure a representative participant sample with respect to the application domain. While the majority of application domains in our corpus have a low risk when a user follows the recommendations without understanding them, they usually target a diverse audience, which should be represented in the evaluation process. Domains such as the medical domain involve a higher risk compared to, e.g., POI recommendations. We, therefore, suggest paying special attention to the representativeness of the participant sample in these domains to make sure that the explanations are beneficial to all users. We particularly recommend paying attention to the participants' location so as to not exclude users from the Global South.

5.5.2 Reporting of Participant Data. The differences in what and how participant data is reported, pointed out in Section 4, make it impossible to reproduce experiments or draw conclusions about the external validity of the results. Therefore, we recommend *working toward and applying reporting standards such as the APA style guideline* [11]. They suggest, for example, avoiding reporting the age of participants with open-ended bins and including options for non-binary genders. A unified way of describing participants in human-subject experiments would not only improve the assessability of the external validity of the experiment results but also the reproducibility of the study.

Regarding *reporting the participants' personality traits, we want to encourage a more fine-grained division of the participants instead of a median split.* A comparison of, e.g., people with a very high or low NFC score with the average would be interesting to see. Furthermore, it would improve the reproducibility of the study if the participant distributions were reported for the personality traits instead of the distribution across two groups after a median split.

5.5.3 Evaluation of Explanations. We recommend *more research efforts to investigate the impact of user characteristics on the effects of different explanations.* Knowing how the characteristics of users and explanations are connected is vital to moving toward explanation designs that benefit the entire user base of an application. As can be seen in Table 3, the majority of articles in our corpus recorded demographic data about their participants, but the results were only disaggregated for a few studies. Analyzing the impact of demographic characteristics on the measured effects would, therefore, not require additional data collection efforts and could serve as a starting point to learn more about how explanations impact different user groups. We further want to encourage research teams to include explanations from related work as baselines in their evaluation to learn more about the external validity of the previously reported results. Table 4 can serve as a guide to the kinds of questions that require answering and to identify promising baseline explanations.

5.5.4 Reproducibility. Several Information Retrieval and Recommender Systems conferences (e.g., RecSys, SIGIR, ECIR, UMAP) have designated reproducibility tracks or include reproducibility efforts in their call for papers. Despite the challenges that we identified in this work, we want to encourage the community to reproduce existing work and submit to these venues to discuss the difficulties they are facing with the community and to increase awareness of the issues that still need to be solved. In our analysis of context variables, we identified that many explanations were simulated or not enough information was provided about the explanation generation process. Trying to generate the evaluated explanations with existing recommender systems and explainable AI methods could make for an interesting starting point.

5.5.5 Pre-Registration of User Studies. Pre-registering a study protocol is common practice in medicine and is increasingly promoted in social sciences to reduce reporting and publication bias [95]. We recommend fostering a discussion among the research community on how to establish and adopt the pre-registration practice for human subject evaluation.

5.5.6 FAIR Data Sharing. Lastly, we argue that *the community should adopt a privacy-preserving approach to* ***Findable, Accessible, Interoperable, and Reusable (FAIR)*** *data-sharing practices* [150] to enable both meta-analyses and comparisons across papers. We acknowledge that given that the data collected is about human subjects, special care should be given to preserve the subjects' privacy, but we believe that privacy-preserving solutions can be found.

6 Limitations

One of the main limitations of this work is that we are bound to what has been reported in the papers of our corpus. Some papers might have recorded and evaluated user characteristics but not reported them. It is more likely that statistically significant results are reported than results where the evaluated effect was not found. This might affect the results for RQ 2. However, our work highlights what is known in the research community, so we think the reporting bias did not significantly influence our contribution.

Smaller limitations come from the way certain information is reported in the papers. In publications with multiple user studies, the information on the participants is occasionally summarized and collectively reported for all studies. In these cases, we only considered the information once when analyzing the participants for RQ 1. Regarding the location data, it is sometimes unclear if the country of residence or the participants' nationality is meant. As most papers collected information about the country of residence, we opted to count these cases to this category.

When creating the corpus for a survey, a selection bias might limit the scope of the papers. However, we mitigated this by querying library databases covering multiple disciplines and not selecting papers based on venue or number of citations. Furthermore, we only analyzed the papers published between 2017 and 2022. This timespan showed the largest increase in explainability research [145], demonstrating that the sample of papers we analyzed covers a significant time period.

Regarding the validity of the results of RQ2, we only considered the impact of user characteristics on explanation effects and the application domain, explanation modality, explainability type, and recommender type as context variables. However, as indicated in Section 5.3, other factors might also play a role, such as the evaluation approach and methodology. Another factor limiting these results is how we categorized the evaluated explanation effects. Our goal was to provide an overview of which user characteristics have been evaluated on which effects. Therefore, we kept the effect groups on a broad level. However, the publications evaluated different aspects of user experience or other notions of trust. We plan to extend our analyses of the results in this corpus in future work to include the aforementioned context variables and a more fine-grained categorization of the explanation effects.

Lastly, the publication period of our corpus selection includes the years in which the Covid-19 pandemic made in-person user studies impossible in most parts of the world. This could have led to an over-proportional representation of online- and crowdsourced studies. However, these study types were already popular before the start of the pandemic, which is why we do not believe that this had an impact on our results.

7 Conclusions and Future Work

In this literature survey, we examined recent publications in which recommender systems explanations were evaluated on users and investigated the question of who these explanations actually serve.

Analyzing the surveyed study participants and answering the first research question, it was brought to light that *most studies do not reflect the average population and exclude larger user groups* such as people from the Global South, users without academic degrees, or older people. Overall, the information on what characteristics are being recorded and how they are reported can vary greatly between publications. In terms of the second research question, we cannot definitely say if these characteristics impact the effect that explanations have on the users due to the small number of studies in which this was evaluated. Our analysis only shows possible indications that should be further investigated by future work in order to find an answer to this research question.

We close our discussion with recommendations for participant recruiting, reporting of participant data, and evaluation of explanation effects. The *recruiting process should be more mindful of reflecting the targeted user group regarding demographics and characteristics*. A description of the targeted user groups and a statement of how the recruited participants reflect them should be explicitly mentioned. *The way participant data is reported in the papers should follow standards to ensure comparability across papers and improve their reproducibility*. Ideally, the data would be shared in a privacy-preserving, FAIR manner to enable subsequent analyses and data re-use. In order to know which user characteristics matter, we recommend putting more effort into investigating the impact of user characteristics on explanation effects.

In future work, we plan to extend this analysis to address the limitations and gain further insights into the effects of explanations on users and aspects that can impact them. We hope this will pave the way to a better understanding of the interaction between user characteristics and explanations in recommender systems — a topic in dire need of elucidation.

References

[1] Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. 2020. Complementing educational recommender systems with open learner models. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge* (Frankfurt, Germany) *(LAK'20)*. ACM, New York, NY, USA, 360–365. https://doi.org/10.1145/3375462.3375520

[2] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. 2022. Changing users' health behaviour intentions through an embodied conversational agent delivering explanations based on users' beliefs and goals. *Behaviour & Information Technology* 42, 9 (May 2022), 1338–1356. https://doi.org/10.1080/0144929x.2022.2073269

[3] Amal Abdulrahman, Deborah Richards, Hedieh Ranjbartabar, and Samuel Mascarenhas. 2019. Belief-based agent explanations to encourage behaviour change. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA'19)*. ACM, New York, NY, USA, 176–178. https://doi.org/10.1145/3308532.3329444

[4] Ahmad Hassan Afridi. 2018. Visualizing serendipitous recommendations in user controlled recommender system for learning. *Procedia Computer Science* 141 (2018), 496–502. https://doi.org/10.1016/j.procs.2018.10.136

[5] Ahmad Hassan Afridi. 2019. Transparency for beyond-accuracy experiences: A novel user interface for recommender systems. *Procedia Computer Science* 151 (2019), 335–344. https://doi.org/10.1016/j.procs.2019.04.047

[6] Ahmad Hassan Afridi and Fatma Outay. 2020. Triggers and connection-making for serendipity via user interface in recommender systems. *Personal and Ubiquitous Computing* 25, 1 (Feb. 2020), 77–92. https://doi.org/10.1007/s00779-020-01371-w

[7] Mohammed Alshammari, Olfa Nasraoui, and Scott Sanders. 2019. Mining semantic knowledge graphs to add explainability to black box recommender systems. *IEEE Access* 7 (2019), 110563–110579. https://doi.org/10.1109/ACCESS.2019.2934633

[8] Alaa Alslaity and Thomas Tran. 2020. The effect of personality traits on persuading recommender system users. In *Proceedings of the 7th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with the 14th ACM Conference on Recommender Systems (RecSys 2020), Online Event (CEUR Workshop Proceedings, Vol. 2682)*. CEUR-WS.org, Aachen, Germany, 48–56. https://ceur-ws.org/Vol-2682/paper5.pdf

[9] Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2019. MoodPlay: Interactive music recommendation based on artists' mood similarity. *International Journal of Human-Computer Studies* 121 (Jan. 2019), 142–159. https://doi.org/10.1016/j.ijhcs.2018.04.004

[10] Alejandro Ariza-Casabona, Ludovico Boratto, and Maria Salamó. 2024. A comparative analysis of text-based explainable recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) *(RecSys'24)*. Association for Computing Machinery, New York, NY, USA, 105–115. https://doi.org/10.1145/3640457.3688069

[11] American Psychological Association. 2022. *Publication Manual of the American Psychological Association.* American Psychological Association, Washington, DC.

[12] Z. K. A. Baizal, Dwi H. Widyantoro, and Nur Ulfa Maulidevi. 2020. Computational model for generating interactions in conversational recommender system based on product functional requirements. *Data & Knowledge Engineering* 128 (Jul. 2020), 101813. https://doi.org/10.1016/j.datak.2020.101813

[13] Krisztian Balog and Filip Radlinski. 2020. Measuring recommendation explanation quality. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR'20)*. ACM, New York, NY, USA, 329–338. https://doi.org/10.1145/3397271.3401032

[14] Natã M. Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI'19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300773

[15] Catalin-Mihai Barbu, Guillermo Carbonell, and Jürgen Ziegler. 2019. The influence of trust cues on the trustworthiness of online reviews for recommendations. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (Limassol, Cyprus) *(SAC'19)*. ACM, New York, NY, USA, 1687–1689. https://doi.org/10.1145/3297280.3297603

[16] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[17] Jordan Barria-Pineda, Kamil Akhuseyinoglu, Peter Brusilovsky, Kerttu Pollari-Malmi, Teemu Sirkiä, and Lauri Malmi. 2020. Personalized remedial recommendations for SQL programming practice system. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP'20 Adjunct)*. ACM, New York, NY, USA, 135–142. https://doi.org/10.1145/3386392.3399312

[18] Jordan Barria Pineda and Peter Brusilovsky. 2019. Making educational recommendations transparent through a fine-grained open learner model. In *Proceedings of the Workshop on Intelligent User Interfaces for Algorithmic Transparency in Emerging Technologies co-located with the 24th ACM Conference on Intelligent User Interfaces, IUI 2019, Los Angeles, CA, USA (CEUR Workshop Proceedings, Vol. 2327)*. CEUR-WS.org, Aachen, Germany, 5 pages.

[19] 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* (Limassol, Cyprus) *(IUI'17)*. ACM, New York, NY, USA, 287–300. https://doi.org/10.1145/3025171.3025209

[20] 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* (Singapore, Singapore) *(UMAP'18)*. ACM, New York, NY, USA, 285–289. https://doi.org/10.1145/3209219.3209251

[21] Émilie Bigras, Pierre-Majorique Léger, and Sylvain Sénécal. 2019. Recommendation agent adoption: How recommendation presentation influences employees' perceptions, behaviors, and decision quality. *Applied Sciences* 9, 20 (Oct. 2019), 14 pages. https://doi.org/10.3390/app9204244

[22] Veronika Bogina, Julia Sheidin, Tsvi Kuflik, and Shlomo Berkovsky. 2020. Visualizing program genres' temporal-based similarity in linear TV recommendations. In *Proceedings of the 2020 International Conference on Advanced Visual Interfaces* (Salerno, Italy) *(AVI'20)*. ACM, New York, NY, USA, Article 69, 5 pages. https://doi.org/10.1145/3399715.3399813

[23] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI'20)*. ACM, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[24] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To trust or to think. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (Apr. 2021), 1–21. https://doi.org/10.1145/3449287

[25] John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of Personality Assessment* 48, 3 (1984), 306–307.

[26] Mohamed Amine Chatti, Mouadh Guesmi, Laura Vorgerd, Thao Ngo, Shoeb Joarder, Qurat Ul Ain, and Arham Muslim. 2022. Is more always better? The effects of personal characteristics and level of detail on the perception of explanations in a recommender system. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP'22)*. ACM, New York, NY, USA, 254–264. https://doi.org/10.1145/3503252.3531304

[27] Li Chen and Feng Wang. 2017. Explaining recommendations based on feature sentiments in product reviews. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI'17)*. ACM, New York, NY, USA, 17–28. https://doi.org/10.1145/3025171.3025173

[28] Li Chen, Dongning Yan, and Feng Wang. 2019. User evaluations on sentiment-based recommendation explanations. *ACM Transactions on Interactive Intelligent Systems* 9, 4, Article 20 (Aug. 2019), 38 pages. https://doi.org/10.1145/3282878

[29] Jin Yao Chin, Yile Chen, and Gao Cong. 2022. The datasets dilemma: How much do we really know about recommendation datasets?. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM'22)*. Association for Computing Machinery, New York, NY, USA, 141–149. https://doi.org/10.1145/3488560.3498519

[30] Michael Chmielewski and Sarah C. Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11, 4 (2020), 464–473. https://doi.org/10.1177/1948550619875149

[31] Jaewon Choi, Hong Joo Lee, and Hee-Woong Kim. 2017. Examining the effects of personalized app recommender systems on purchase intention: A self and social-interaction perspective. *Journal of Electronic Commerce Research* 18, 1 (2017), 73 – 102. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85021167503&partnerID=40&md5=b59bc6cba88be85a5bfaad0778c97654 Cited by: 20.

[32] Michael Chromik and Martin Schuessler. 2020. A taxonomy for human subject evaluation of black-box explanations in XAI. In *Proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies co-located with the 25th International Conference on Intelligent User Interfaces (IUI 2020), Cagliari, Italy (CEUR Workshop Proceedings, Vol. 2582)*. CEUR-WS.org, Aachen, Germany, 7 pages. https://ceur-ws.org/Vol-2582/paper9.pdf

[33] Robert B. Cialdini. 2001. *Influence: Science and Practice*. Vol. 4. Allyn and Bacon, Boston, MA, USA.

[34] Ludovik Coba, Laurens Rook, Markus Zanker, and Panagiotis Symeonidis. 2019. Decision making strategies differ in the presence of collaborative explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI'19)*. ACM, New York, NY, USA, 291–302. https://doi.org/10.1145/3301275.3302304

[35] Ludovik Coba, Markus Zanker, Laurens Rook, and Panagiotis Symeonidis. 2018. Exploring users' perception of collaborative explanation styles. In *Proceedings of the 2018 IEEE 20th Conference on Business Informatics (CBI)* (Vienna, Austria), Vol. 01. Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 70–78. https://doi.org/10.1109/CBI.2018.00017

[36] C. Conati, G. Carenini, E. Hoque, B. Steichen, and D. Toker. 2014. Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. *Computer Graphics Forum* 33, 3 (June 2014), 371–380. https://doi.org/10.1111/cgf.12393

[37] Cristina Conati and Heather Maclaren. 2008. Exploring the role of individual differences in information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Napoli, Italy) *(AVI'08)*. Association for Computing Machinery, New York, NY, USA, 199–206. https://doi.org/10.1145/1385569.1385602

[38] Rodrigo Fernandes Gomes da Silva, Chanchal K. Roy, Mohammad Masudur Rahman, Kevin A. Schneider, Klérisson Paixão, Carlos Eduardo de Carvalho Dantas, and Marcelo de Almeida Maia. 2020. CROKAGE: Effective solution recommendation for programming tasks by leveraging crowd knowledge. *Empirical Software Engineering* 25, 6 (Sep. 2020), 4707–4758. https://doi.org/10.1007/s10664-020-09863-2

[39] Khalil Damak, Olfa Nasraoui, and William Scott Sanders. 2021. Sequence-based explainable hybrid song recommendation. *Frontiers in Big Data* 4 (Jul. 2021), 13 pages. https://doi.org/10.3389/fdata.2021.693494

[40] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of Mechanical Turk workers. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM'18)*. Association for Computing Machinery, New York, NY, USA, 135–143. https://doi.org/10.1145/3159652.3159661

[41] Vicente Dominguez, Ivania Donoso-Guzmán, Pablo Messina, and Denis Parra. 2020. Algorithmic and HCI aspects for explaining recommendations of artistic images. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (Nov. 2020), 1–31. https://doi.org/10.1145/3369396

[42] Vicente Dominguez and Pablo Messina. 2018. Towards explanations for visual recommender systems of artistic images. In *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, IntRS 2018, co-located with the 12th ACMConference on Recommender Systems (RecSys 2018) (CEUR Workshop Proceedings, Vol. 2225)*. CEUR-WS.org, Aachen, Germany, 69–73.

[43] Zhenhua Dong, Chuan Shi, Shilad Sen, Loren Terveen, and John Riedl. 2012. War versus inspirational in Forrest Gump: Cultural effects in tagging communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 6. The AAAI Press, Washington, DC, USA, 82–89. https://doi.org/10.1609/icwsm.v6i1.14258

[44] Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2020. Explaining recommendations by means of aspect-based transparent memories. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI'20)*. ACM, New York, NY, USA, 166–176. https://doi.org/10.1145/3377325.3377520

[45] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS One* 18, 3 (Mar. 2023), 1–17. https://doi.org/10.1371/journal.pone.0279720

[46] Fan Du, Sana Malik, Georgios Theocharous, and Eunyee Koh. 2018. Personalizable and interactive sequence recommender system. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA'18)*. ACM, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188506

[47] Dennis E. Egan. 1988. Chapter 24 - Individual differences in human-computer interaction. In *Handbook of Human-Computer Interaction*, Martin Helander (Ed.). North-Holland, Amsterdam, 543–568. https://doi.org/10.1016/B978-0-444-70536-5.50029-4

[48] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, Cambridge, MA, USA, 172–186. https://proceedings.mlr.press/v81/ekstrand18b.html

[49] Somayeh Fatahi, Mina Mousavifar, and Julita Vassileva. 2023. Investigating the effectiveness of persuasive justification messages in fair music recommender systems for users with different personality traits. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP'23)*. Association for Computing Machinery, New York, NY, USA, 66–77. https://doi.org/10.1145/3565472.3592958

[50] Phyliss Jia Gai and Anne-Kathrin Klesse. 2019. Making recommendations more effective through framings: Impacts of user- versus item-based framings on recommendation click-throughs. *Journal of Marketing* 83, 6 (Sep. 2019), 61–75. https://doi.org/10.1177/0022242919873901

[51] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI'22)*. ACM, New York, NY, USA, 794–806. https://doi.org/10.1145/3490099.3511138

[52] Mouadh Guesmi, Mohamed Amine Chatti, Laura Vorgerd, Shoeb Ahmed Joarder, Qurat Ul Ain, Thao Ngo, Shadi Zumor, Yiqi Sun, Fangzheng Ji, and Arham Muslim. 2021. Input or output: Effects of explanation focus on the perception of explainable recommendation with varying level of details. In *Proceedings of the 8th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with the 15th ACM Conference on Recommender Systems (RecSys 2021), Online Event (CEUR Workshop Proceedings, Vol. 2948)*. CEUR-WS.org, Aachen, Germany, 55–72. https://ceur-ws.org/Vol-2948/paper4.pdf

[53] Mouadh Guesmi, Mohamed Amine Chatti, Laura Vorgerd, Thao Ngo, Shoeb Joarder, Qurat Ul Ain, and Arham Muslim. 2022. 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* (Barcelona, Spain) *(UMAP'22 Adjunct)*. ACM, New York, NY, USA, 175–183. https://doi.org/10.1145/3511047.3537685

[54] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *Comput. Surveys* 51, 5 (Sept. 2019), 1–42. https://doi.org/10.1145/3236009

[55] Junius Gunaratne, Lior Zalmanson, and Oded Nov. 2018. The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems* 35, 4 (Oct. 2018), 1092–1120. https://doi.org/10.1080/07421222.2018.1523534

[56] Sophia Hadash, Martijn C. Willemsen, Chris Snijders, and Wijnand A. IJsselsteijn. 2022. Improving understandability of feature contributions in model-agnostic explainable AI tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI'22)*. ACM, New York, NY, USA, Article 487, 9 pages. https://doi.org/10.1145/3491102.3517650

[57] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33, 2-3 (2010), 61–83. https://doi.org/10.1017/S0140525X0999152X

[58] Stella Heras, Javier Palanca, Paula Rodriguez, Néstor Duque-Méndez, and Vicente Julian. 2020. Recommending learning objects with arguments and explanations. *Applied Sciences* 10, 10 (May 2020), 18 pages. https://doi.org/10.3390/app10103341

[59] Diana C. Hernandez-Bocanegra, Tim Donkers, and Jürgen Ziegler. 2020. Effects of argumentative explanation types on the perception of review-based recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP'20 Adjunct)*. ACM, New York, NY, USA, 219–225. https://doi.org/10.1145/3386392.3399302

[60] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2020. Explaining review-based recommendations: Effects of profile transparency, presentation style and user characteristics. *i-com* 19, 3 (Dec. 2020), 181–200. https://doi.org/10.1515/icom-2020-0021

[61] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Effects of interactivity and presentation on review-based explanations for recommendations. In *Human-Computer Interaction – INTERACT 2021*. Springer International Publishing, Cham, 597–618. https://doi.org/10.1007/978-3-030-85616-8_35

[62] Minsung Hong, Sojung An, Rajendra Akerkar, David Camacho, and Jason J. Jung. 2019. Cross-cultural contextualisation for recommender systems. *Journal of Ambient Intelligence and Humanized Computing* 15, 2 (2019), 1–12. https://doi.org/10.1007/s12652-019-01479-9

[63] Nan Hu, Paul A. Pavlou, and Jie Zhang. 2017. On self-selection biases in online product reviews. *MIS Quarterly* 41, 2 (2017), pp. 449–475. https://www.jstor.org/stable/26629722

[64] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. *Translational Psychiatry* 11, 1 (Feb. 2021), 108. https://doi.org/10.1038/s41398-021-01224-x

[65] Michael Jugovac, Ingrid Nunes, and Dietmar Jannach. 2018. Investigating the decision-making behavior of maximizers and satisficers in the presence of recommendations. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP'18)*. ACM, New York, NY, USA, 279–283. https://doi.org/10.1145/3209219.3209252

[66] Marius Kaminskas, Frederico Durao, and Derek Bridge. 2017. Item-based explanations for user-based recommendations. In *Proceedings of 9th International Conference on Information, Process, and Knowledge Management*. IARIA Press, Wilmington, DE, USA, 65–70.

[67] Shivani Kapania, Alex S. Taylor, and Ding Wang. 2023. A hunt for the Snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI'23)*. Association for Computing Machinery, New York, NY, USA, Article 133, 15 pages. https://doi.org/10.1145/3544548.3580645

[68] Soultana Karga and Maya Satratzemi. 2019. Using explanations for recommender systems in learning design settings to enhance teachers' acceptance and perceived experience. *Educ. Inf. Technol.* 24, 3 (Apr. 2019), 2953—-2974. https://doi.org/10.1007/s10639-019-09909-z

[69] Byung Hyung Kim, Seunghun Koh, Sejoon Huh, Sungho Jo, and Sunghee Choi. 2020. Improved explanatory efficacy on human affect and workload through interactive process in artificial intelligence. *IEEE Access* 8 (2020), 189013–189024. https://doi.org/10.1109/access.2020.3032056

[70] Akiva Kleinerman, Ariel Rosenfeld, and Sarit Kraus. 2018. Providing explanations for recommendations in reciprocal environments. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys'18)*. ACM, New York, NY, USA, 22–30. https://doi.org/10.1145/3240323.3240362

[71] Akiva Kleinerman, Ariel Rosenfeld, Francesco Ricci, and Sarit Kraus. 2020. Supporting users in finding successful matches in reciprocal recommender systems. *User Modeling and User-Adapted Interaction* 31, 3 (Oct. 2020), 541–589. https://doi.org/10.1007/s11257-020-09279-z

[72] Akihiro Kokubo and Kazunari Sugiyama. 2022. Explainable recommendation enhancing review properties and PPLM. In *Proceedings of the 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, Piscataway, NJ, USA, 151–158. https://doi.org/10.1109/WI-IAT55865.2022.00030

[73] Sherrie Yi Xiao Komiak. 2003. *The Impact of Internalization and Familiarity on Trust and Adoption of Recommendation Agents*. Ph. D. Dissertation. University of British Columbia. https://doi.org/10.14288/1.0091325

[74] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User preferences for hybrid explanations. In *Proceedings of the 11th ACM Conference on Recommender Systems* (Como, Italy) *(RecSys'17)*. ACM, New York, NY, USA, 84–88. https://doi.org/10.1145/3109859.3109915

[75] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2020. Generating and understanding personalized explanations in hybrid recommender systems. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (Nov. 2020), 1–40. https://doi.org/10.1145/3365843

[76] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) *(CHI'19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[77] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from explainable artificial intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[78] Yu Liang and Martijn C. Willemsen. 2021. Interactive music genre exploration with visualization and mood control. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI'21)*. ACM, New York, NY, USA, 175–185. https://doi.org/10.1145/3397481.3450700

[79] Mengqi Liao and S. Shyam Sundar. 2021. When E-commerce personalization systems show and tell: Investigating the relative persuasive appeal of content-based versus collaborative filtering. *Journal of Advertising* 51, 2 (Mar. 2021), 256–267. https://doi.org/10.1080/00913367.2021.1887013

[80] Mengqi Liao, S. Shyam Sundar, and Joseph B. Walther. 2022. User trust in recommendation systems: A comparison of content-based, collaborative and demographic filtering. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI'22)*. ACM, New York, NY, USA, Article 486, 14 pages. https://doi.org/10.1145/3491102.3501936

[81] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI'20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[82] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI'21)*. Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. https://doi.org/10.1145/3411764.3445488

[83] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2019. Interactive recommending with tag-enhanced matrix factorization (TagMF). *International Journal of Human-Computer Studies* 121 (Jan. 2019), 21–41. https://doi.org/10.1016/j.ijhcs.2018.05.002

[84] Boxuan Ma, Min Lu, Yuta Taniguchi, and Shin'ichi Konomi. 2021. CourseQ: The impact of visual and interactive course recommendation in university environments. *Research and Practice in Technology Enhanced Learning* 16, 1 (June 2021), 18 pages. https://doi.org/10.1186/s41039-021-00167-7

[85] Ayoub El Majjodi, Alain D. Starke, and Christoph Trattner. 2022. Nudging towards health? Examining the merits of nutrition labels and personalization in a recipe recommender system. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP'22)*. ACM, New York, NY, USA, 48–56. https://doi.org/10.1145/3503252.3531312

[86] Avleen Malhi, Samanta Knapic, and Kary Främling. 2020. Explainable agents for less bias in human-agent decision making. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer International Publishing, Cham, 129–146. https://doi.org/10.1007/978-3-030-51924-7_8

[87] Christopher King Manner and Wilburn C. Lane. 2017. Who posts online customer reviews? The role of sociodemographics and personality traits. *Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior* 30 (2017), 19–42. https://www.jcsdcb.com/index.php/JCSDCB/article/view/226

[88] Millecamp Martijn, Cristina Conati, and Katrien Verbert. 2022. "Knowing me, knowing you": Personalized explanations for a music recommender system. *User Modeling and User-Adapted Interaction* 32, 1-2 (Jan. 2022), 215–252. https://doi.org/10.1007/s11257-021-09304-9

[89] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. 2022. Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management* 39, 1 (Jan. 2022), 53–63. https://doi.org/10.1080/10580530.2020.1849465

[90] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. 2020. Recommender systems and their ethical challenges. *AI & Society* 35, 4 (Dec. 2020), 957–967. https://doi.org/10.1007/s00146-020-00950-y

[91] 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* (Marina del Ray, California) *(IUI'19)*. ACM, New York, NY, USA, 397–407. https://doi.org/10.1145/3301275.3302313

[92] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2020. What's in a user? Towards personalising transparency for music recommender interfaces. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP'20)*. ACM, New York, NY, USA, 173–182. https://doi.org/10.1145/3340631.3394844

[93] Aditi Mishra, Utkarsh Soni, Jinbin Huang, and Chris Bryan. 2022. Why? Why not? When? Visual explanations of agent behaviour in reinforcement learning. In *Proceedings of the 2022 IEEE 15th Pacific Visualization Symposium (PacificVis)* (Tsukuba, Japan). Institute of Electrical and Electronics Engineers (IEEE), Piscataway, NJ, USA, 111–120. https://doi.org/10.1109/PacificVis53943.2022.00020

[94] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems* 11, 3–4, Article 24 (Sept. 2021), 45 pages. https://doi.org/10.1145/3387166

[95] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1, 1 (2017), 0021. https://doi.org/10.1038/s41562-016-0021

[96] Cataldo Musto, Fedelucio Narducci, Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2019. Linked open data-based explanations for transparent recommender systems. *International Journal of Human-Computer Studies* 121 (Jan. 2019), 93–107. https://doi.org/10.1016/j.ijhcs.2018.03.003

[97] Cataldo Musto, Alain D. Starke, Christoph Trattner, Amon Rapp, and Giovanni Semeraro. 2021. Exploring the effects of natural language justifications in food recommender systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP'21)*. ACM, New York, NY, USA, 147–157. https://doi.org/10.1145/3450613.3456827

[98] Daniel Müllensiefen, Bruno Gingras, Jason Musil, and Lauren Stewart. 2014. The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLOS One* 9, 2 (Feb. 2014), 1–23. https://doi.org/10.1371/journal.pone.0089642

[99] Mohammad Naiseh, Nan Jiang, Jianbing Ma, and Raian Ali. 2020. Personalising explainable recommendations: Literature and conceptualisation. In *Trends and Innovations in Information Systems and Technologies*, Álvaro Rocha, Hojjat Adeli, Luís Paulo Reis, Sandra Costanzo, Irena Orovic, and Fernando Moreira (Eds.). Springer International Publishing, Cham, 518–533. https://doi.org/10.1007/978-3-030-45691-7_49

[100] Shabnam Najafian, Amra Delic, Marko Tkalcic, and Nava Tintarev. 2021. Factors influencing privacy concern for explanations of group recommendation. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP'21)*. ACM, New York, NY, USA, 14–23. https://doi.org/10.1145/3450613.3456845

[101] Shabnam Najafian, Tim Draws, Francesco Barile, Marko Tkalcic, Jie Yang, and Nava Tintarev. 2021. Exploring user concerns about disclosing location and emotion information in group recommendations. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (Virtual Event, USA) *(HT'21)*. ACM, New York, NY, USA, 155–164. https://doi.org/10.1145/3465336.3475104

[102] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *Comput. Surveys* 55, 13s, Article 295 (Feb. 2023), 42 pages. https://doi.org/10.1145/3583558

[103] Sidra Naveed, Tim Donkers, and Jürgen Ziegler. 2018. Argumentation-based explanations in recommender systems. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP'18)*. ACM, New York, NY, USA, 293–298. https://doi.org/10.1145/3213586.3225240

[104] Sidra Naveed, Benedikt Loepp, and Jürgen Ziegler. 2020. On the use of feature-based collaborative explanations: An empirical comparison of explanation styles. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP'20 Adjunct)*. ACM, New York, NY, USA, 226–232. https://doi.org/10.1145/3386392.3399303

[105] Sidra Naveed and Jürgen Ziegler. 2020. Featuristic: An interactive hybrid system for generating explainable recommendations-beyond system accuracy. In *Proceedings of the 7th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with the 14th ACM Conference on Recommender Systems (RecSys 2020), Online Event (CEUR Workshop Proceedings, Vol. 2682)*. CEUR-WS.org, Aachen, Germany, 14–25.

[106] Shreeya Nelekar, Amal Abdulrahman, Manik Gupta, and Deborah Richards. 2021. Effectiveness of embodied conversational agents for managing academic stress at an Indian university (ARU) during COVID-19. *British Journal of Educational Technology* 53, 3 (Dec. 2021), 491–511. https://doi.org/10.1111/bjet.13174

[107] Fakhroddin Noorbehbahani and Zeinab Zarein. 2018. The impact of demographic factors on persuasion strategies in personalized recommender system. In *Proceedings of the 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE)* (Mashhad, Iran). Institute of Electrical and Electronics Engineers (IEEE), Piscataway, New Jersey, USA, 104–109. https://doi.org/10.1109/ICCKE.2018.8566550

[108] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27 (Dec. 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[109] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining recommendations in e-learning: Effects on adolescents' trust. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI'22)*. ACM, New York, NY, USA, 93–105. https://doi.org/10.1145/3490099.3511140

[110] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: A user study for AI-based clinical decision support systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI'22)*. ACM, New York, NY, USA, Article 568, 9 pages. https://doi.org/10.1145/3491102.3502104

[111] Florian Pecune, Lucile Callebert, and Stacy Marsella. 2022. Designing persuasive food conversational recommender systems with nudging and socially-aware conversational strategies. *Frontiers in Robotics and AI* 8 (Jan. 2022), 22 pages. https://doi.org/10.3389/frobt.2021.733835

[112] Lara Quijano-Sanchez, Christian Sauer, Juan A. Recio-Garcia, and Belen Diaz-Agudo. 2017. Make it personal: A social explanation system applied to group recommendations. *Expert Systems with Applications* 76 (June 2017), 36–48. https://doi.org/10.1016/j.eswa.2017.01.045

[113] Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popovic, and Daniel S. Weld. 2022. Exploring the role of local and global explanations in recommender systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems Extended Abstracts* (New Orleans, LA, USA) *(CHI EA'22)*. ACM, New York, NY, USA, Article 290, 7 pages. https://doi.org/10.1145/3491101.3519795

[114] Arpit Rana and Derek Bridge. 2018. Explanations that are intrinsic to recommendations. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP'18)*. ACM, New York, NY, USA, 187–195. https://doi.org/10.1145/3209219.3209230

[115] Arpit Rana, Rafael M. D'Addio, Marcelo G. Manzato, and Derek Bridge. 2022. Extended recommendation-by-explanation. *User Modeling and User-Adapted Interaction* 32, 1-2 (Mar. 2022), 91–131. https://doi.org/10.1007/s11257-021-09317-4

[116] Katharina Reinecke and Abraham Bernstein. 2011. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Transactions on Computer-Human Interaction* 18, 2, Article 8 (July 2011), 29 pages. https://doi.org/10.1145/1970378.1970382

[117] Masahiro Sato, Budrul Ahsan, Koki Nagatani, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. 2018. Explaining recommendations using contexts. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI'18)*. ACM, New York, NY, USA, 659–664. https://doi.org/10.1145/3172944.3173012

[118] Masahiro Sato, Shin Kawai, and Hajime Nobuhara. 2019. Action-triggering recommenders: Uplift optimization and persuasive explanation. In *Proceedings of the International Conference on Data Mining Workshops (ICDMW)* (Beijing, China). Institute of Electrical and Electronics Engineers (IEEE), Piscataway, New Jersey,USA, 1060–1069. https://doi.org/10.1109/ICDMW.2019.00155

[119] Julia Sheidin, Joel Lanir, Cristina Conati, Dereck Toker, and Tsvi Kuflik. 2020. The effect of user characteristics in time series visualizations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI'20)*. Association for Computing Machinery, New York, NY, USA, 380–389. https://doi.org/10.1145/3377325.3377502

[120] Dorin Shmaryahu, Guy Shani, and Bracha Shapira. 2020. Post-hoc explanations for complex model recommendations using simple methods. In *Proceedings of the 7th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with the 14th ACM Conference on Recommender Systems (RecSys 2020), Online Event (CEUR Workshop Proceedings, Vol. 2682)*. CEUR-WS.org, Aachen, Germany, 26–36. https://ceur-ws.org/Vol-2682/paper3.pdf

[121] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology* 24, 1 (Jan. 2022), 2 pages. https://doi.org/10.1007/s10676-022-09623-4

[122] Gkika Sofia, Skiada Marianna, Lekakos George, and Kourouthanasis Panos. 2016. Investigating the role of personality traits and influence strategies on the persuasive effect of personalized recommendations. In *Proceedings of the 4th Workshop on Emotions and Personality in Personalized Systems co-located with the ACM Conference on Recommender Systems (RecSys 2016), Boston, MA, USA (CEUR Workshop Proceedings, Vol. 1680)*. CEUR-WS.org, Aachen, Germany, 9–17. https://ceur-ws.org/Vol-1680/paper2.pdf

[123] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: Empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI'19)*. Association for Computing Machinery, New York, NY, USA, 107–120. https://doi.org/10.1145/3301275.3302322

[124] Alain Starke. 2019. The effectiveness of advice solicitation and social peers in an energy recommender system, In *Proceedings of the 6th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with the 13th ACM Conference on Recommender Systems(RecSys 2019), Copenhagen, Denmark. CEUR Workshop Proceedings* 2450, 65 – 71. https://ceur-ws.org/Vol-2450/short3.pdf

[125] Alain Starke, Edis Asotic, and Christoph Trattner. 2021. "Serving each user": Supporting different eating goals through a multi-list recommender interface. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys'21)*. ACM, New York, NY, USA, 124–132. https://doi.org/10.1145/3460231.3474232

[126] Alain Starke, Martijn Willemsen, and Chris Snijders. 2021. Promoting energy-efficient behavior by depicting social norms in a recommender interface. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Sep. 2021), 1–32. https://doi.org/10.1145/3460005

[127] Christian Sturm, Alice Oh, Sebastian Linxen, Jose Abdelnour Nocera, Susan Dray, and Katharina Reinecke. 2015. How WEIRD is HCI? Extending HCI principles to other countries and cultures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI EA'15)*. Association for Computing Machinery, New York, NY, USA, 2425–2428. https://doi.org/10.1145/2702613.2702656

[128] 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. In *Adjunct Proceedings of the 27th Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19 Adjunct)*. ACM, New York, NY, USA, 241–245. https://doi.org/10.1145/3314183.3323456

[129] Yuan Sun and S. Shyam Sundar. 2022. Exploring the effects of interactive dialogue in improving user control for explainable online symptom checkers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems Extended Abstracts* (New Orleans, LA, USA) *(CHI EA'22)*. ACM, New York, NY, USA, Article 417, 7 pages. https://doi.org/10.1145/3491101.3519668

[130] Kyosuke Takami, Yiling Dai, Brendan Flanagan, and Hiroaki Ogata. 2022. Educational explainable recommender usage and its effectiveness in high school summer vacation assignment. In *Proceedings of the 12th International Learning Analytics and Knowledge Conference* (Online, USA) *(LAK22)*. ACM, New York, NY, USA, 458–464. https://doi.org/10.1145/3506860.3506882

[131] Yiyi Tao, Yiling Jia, Nan Wang, and Hongning Wang. 2019. The FacT. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. ACM, New York, NY, USA, 295–304. https://doi.org/10.1145/3331184.3331244

[132] Serge Thill, Maria Riveiro, Erik Lagerstedt, Mikael Lebram, Paul Hemeren, Azra Habibovic, and Maria Klingegård. 2018. Driver adherence to recommendations from support systems improves if the systems explain why they are given: A simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour* 56 (July 2018), 420–435. https://doi.org/10.1016/j.trf.2018.05.009

[133] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*. IEEE, Piscataway, NJ, USA, 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[134] Nava Tintarev and Judith Masthoff. 2015. *Explaining Recommendations: Design and Evaluation*. Springer US, Boston, MA, 353–382. https://doi.org/10.1007/978-1-4899-7637-6_10

[135] Nava Tintarev and Judith Masthoff. 2016. Effects of individual differences in working memory on plan presentational choices. *Frontiers in Psychology* 7 (Nov. 2016), 22 pages. https://doi.org/10.3389/fpsyg.2016.01793

[136] Nava Tintarev, Shahin Rostami, and Barry Smyth. 2018. Knowing the unknown. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing* (Pau, France) *(SAC'18)*. ACM, New York, NY, USA, 1396–1399. https://doi.org/10.1145/3167132.3167419

[137] Marko Tkalcic and Li Chen. 2015. *Personality and Recommender Systems*. Springer US, Boston, MA, 715–739. https://doi.org/10.1007/978-1-4899-7637-6_21

[138] Helma Torkamaan and Jürgen Ziegler. 2019. Rating-based preference elicitation for recommendation of stress intervention. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19)*. ACM, New York, NY, USA, 46–50. https://doi.org/10.1145/3320435.3324990

[139] Thi Ngoc Trang Tran, Muesluem Atas, Man Le, Ralph Samer, and Martin Stettinger. 2020. Social choice-based explanations: An approach to enhancing fairness and consensus aspects. *JUCS - Journal of Universal Computer Science* 26, 3 (Mar. 2020), 402–431. https://doi.org/10.3897/jucs.2020.021

[140] Thi Ngoc Trang Tran, Viet Man Le, Muesluem Atas, Alexander Felfernig, Martin Stettinger, and Andrei Popescu. 2021. Do users appreciate explanations of recommendations? An analysis in the movie domain. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys'21)*. ACM, New York, NY, USA, 645–650. https://doi.org/10.1145/3460231.3478859

[141] Chun-Hua Tsai and Peter Brusilovsky. 2019. Evaluating visual explanations for similarity-based recommendations. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19)*. ACM, New York, NY, USA, 22–30. https://doi.org/10.1145/3320435.3320465

[142] Chun-Hua Tsai and Peter Brusilovsky. 2020. The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction* 31, 3 (Oct. 2020), 591–627. https://doi.org/10.1007/s11257-020-09281-5

[143] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. 2021. Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI'21)*. ACM, New York, NY, USA, Article 152, 17 pages. https://doi.org/10.1145/3411764.3445101

[144] Kosetsu Tsukuda and Masataka Goto. 2020. Explainable recommendation for repeat consumption. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys'20)*. ACM, New York, NY, USA, 462–467. https://doi.org/10.1145/3383313.3412230

[145] Alexandra Vultureanu-Albişi and Costin Bădică. 2022. A survey on effects of adding explanations to recommender systems. *Concurrency and Computation: Practice and Experience* 34, 20 (Jan. 2022), 15 pages. https://doi.org/10.1002/cpe.6834

[146] GaoShan Wang, XiQuan Liu, ZhongGuo Wang, and XueLan Yang. 2020. Research on the influence of interpretability of artificial intelligence recommendation system on users' behavior intention. In *Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering* (Xiamen, China) *(EITCE'20)*. ACM, New York, NY, USA, 762–766. https://doi.org/10.1145/3443467.3443850

[147] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable recommendation via multi-task learning in opinionated text data. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR'18)*. ACM, New York, NY, USA, 165–174. https://doi.org/10.1145/3209978.3210010

[148] Agung Toto Wibowo, Advaith Siddharthan, Judith Masthoff, and Chenghua Lin. 2018. Understanding how to explain package recommendations in the clothes domain. In *Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, IntRS 2018, co-located with the 12th ACM Conference on Recommender Systems (RecSys 2018), Vancouver, Canada (CEUR Workshop Proceedings, Vol. 2225)*. CEUR-WS.org, Aachen, Germany, 74–78. https://ceur-ws.org/Vol-2225/paper11.pdf

[149] 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 Transactions on Information Systems* 39, 4 (Oct. 2021), 1–21. https://doi.org/10.1145/3441715

[150] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, 1 (Mar. 2016), 1–9. https://doi.org/10.1038/sdata.2016.18

[151] Zekun Yang and Zhijie Lin. 2021. Interpretable video tag recommendation with multimedia deep learning framework. *Internet Research* 32, 2 (July 2021), 518–535. https://doi.org/10.1108/intr-08-2020-0471

[152] Run Yu, Zach Pardos, Hung Chau, and Peter Brusilovsky. 2021. Orienting students to course recommendations using three types of explanation. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP'21)*. ACM, New York, NY, USA, 238–245. https://doi.org/10.1145/3450614.3464483

[153] Yi Yu, Kazunari Sugiyama, and Adam Jatowt. 2023. AdaReX: Cross-domain, adaptive, and explainable recommender system. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Beijing, China) *(SIGIR-AP'23)*. Association for Computing Machinery, New York, NY, USA, 272–281. https://doi.org/10.1145/3624918.3625331

[154] Yi Yu, Kazunari Sugiyama, and Adam Jatowt. 2024. Sequential recommendation with collaborative explanation via mutual information maximization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR'24)*. Association for Computing Machinery, New York, NY, USA, 1062–1072. https://doi.org/10.1145/3626772.3657770

[155] Jingjing Zhang and Shawn P. Curley. 2017. Exploring explanation effects on consumers' trust in online recommender agents. *International Journal of Human–Computer Interaction* 34, 5 (Sep. 2017), 421–432. https://doi.org/10.1080/10447318.2017.1357904

[156] Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101. https://doi.org/10.1561/1500000066

[157] Zhirun Zhang, Li Chen, Tonglin Jiang, Yutong Li, and Lei Li. 2022. Effects of feature-based explanation and its output modality on user satisfaction with service recommender systems. *Frontiers in Big Data* 5 (May 2022), 17 pages. https://doi.org/10.3389/fdata.2022.897381

[158] Yong Zheng and Juan Ruiz Toribio. 2021. The role of transparency in multi-stakeholder educational recommendations. *User Modeling and User-Adapted Interaction* 31, 3 (Apr. 2021), 513–540. https://doi.org/10.1007/s11257-021-09291-x

[159] Robert Zimmermann, Daniel Mora, Douglas Cirqueira, Markus Helfert, Marija Bezbradica, Dirk Werth, Wolfgang Jonas Weitzl, René Riedl, and Andreas Auinger. 2022. Enhancing brick-and-mortar store shopping experience with an augmented reality shopping assistant application using personalized recommendations and explainable artificial intelligence. *Journal of Research in Interactive Marketing* 17, 2 (Apr. 2022), 273–298. https://doi.org/10.1108/jrim-09-2021-0237

Received 29 April 2024; revised 21 January 2025; accepted 22 January 2025