Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations

Ruoxi Shang
University of Washington
Seattle, USA
rxshang@uw.edu

K. J. Kevin Feng
University of Washington
Seattle, USA
kjfeng@uw.edu

Chirag Shah
University of Washington
Seattle, USA
chirags@uw.edu

ABSTRACT

Intelligent everyday applications typically rely on automated Recommender Systems (RS) to generate recommendations that help users make decisions among a large number of options. Due to the increasing complexity of RS and the lack of transparency in its algorithmic decision-making, researchers have recognized the need to support users with more explanations. While traditional explainability methods fall short in disclosing the internal intricacy of recommender systems, counterfactual explanations provide many desirable explainable features by offering human-like explanations that contrast an existing recommendation with alternatives. However, there is a lack of empirical research in understanding lay users' needs of counterfactual explanations in their usage of everyday intelligent applications. In this paper, we investigate whether and where should we provide counterfactual explanations in everyday recommender systems through a question-driven approach. We first conducted an interview study to understand how existing explanations might be insufficient to support lay users and elicit the triggers that prompt them to ask *why not* questions and seek additional explanations. The findings reveal that the utility of decision is a primary factor that may affect whether users want to consume counterfactual explanations. We then conducted an online scenario-based survey to quantify the correlation between utility and explanation needs and found significant correlations between the measured variables.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Explainable recommender system; Counterfactual explanations; User studies

ACM Reference Format:
Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. 2022. Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3531146.3533189

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533189

1 INTRODUCTION

In recent years, many online service platforms have adopted automated Recommender Systems (RS) to suggest various personalized content for users based on their prior activities and preferences [10]. While recommendation algorithms, often in the form of complex deep neural models, have become increasingly accurate in predicting the users' preferences, their decision-making rationale have become even more difficult to explain [34]. As this layer of opaqueness imposes challenges for users to control, understand, and trust the recommendations, one possible solution is to provide explanations alongside recommended content [46, 55]. In RS, an explanation is a description that justifies a recommendation and therefore helps users better understand whether a recommended item is relevant to their interests [45].

To provide more human-understandable explanations, researchers have been trying to draw parallels to the fundamentals of human reasoning. Based on findings in humanities and social sciences that show explanations are intrinsically contrastive, Miller identified contrastiveness as one of the characteristics of explanation that are not given sufficient attention in Explainable AI (XAI) [32]. As people explain "Why P", they would describe the cause of an event in contrast to some other event that did not occur; that is, an explanation would be of the form "Why P rather than Q?". In this example, P is the target event and Q is an implied counterfactual case that did not occur [32]. Contrastive explanations are used to answer *why not* questions about the system's behavior in which the consequences of the counterfactuals in question are pointed out [58]. This type of explanation is easily digestible and practically useful for understanding the reasoning behind a decision, challenging that reasoning, and altering future behavior in expectation of better results [50]. In the following narration, we will collapse the definition of contrastive and counterfactual explanations as done by many other researchers [43].

Previous work mainly studied its application in high-stake domains such as finance, healthcare, and criminal justice for algorithmic decision support [49, 51], but do not empirically study RS that mediate everyday tasks and practices [54]. Despite the large body of work in generating contrastive explanations through algorithmic methods[43], there is still insufficient understanding of when and how users really seek such explanations in RS, resulting in a gap between explainability research and actual user needs[8, 25, 32].

The lack of consideration of in-the-wild use contexts when evaluating counterfactual explanations led us to ground our research in real-world scenarios. To ensure that the scenarios are broad enough to be relatable to a wide range of users, we turn to everyday service applications—applications that provide an everyday service such as e-commerce and social media—for our study. More specifically, we focus on those everyday applications' adoption of intelligent systems that are capable of selecting relevant information, influencing user-facing content, and learning from users' behaviors to shape future recommendations [19]. Question-driven frameworks have been widely used in the space of understanding users' explanation needs, as an explanation is fundamentally an answer to a question [30, 31]. In this work, we want to focus on the question of *why not* that is under-addressed in RS that generate most content everyday intelligent applications. We note the synonymous relationship between ***why not* questions** and **counterfactual explanations** and refer to the former as a method of probing into the latter in this work. Moreover, since a counterfactual explanations serves to explain a *why not* question, we use ***why not* explanations** and **counterfactual explanations** interchangeably.

In order to better support users through counterfactual explanations in their interactions with everyday recommendations, we investigate the following research question:

> When do users need *why not* explanations when interacting with recommended content in intelligent everyday applications?

This work seeks to elicit users' implicit and explicit reasons behind the needs for explanations of alternative recommended content. That is, in addition to finding when users directly ask *why not* questions, we also look for situations when showing why something is missing from the recommendation might fulfill users' information needs. We do this by probing into users' decision-making processes in everyday service applications. Through this work, we make the following contributions:

- We examine and report how users interact with existing explanations provided for everyday recommendations and the triggers for them to seek *why not* explanations.
- We present quantitative findings from a survey on how users' perceived utility of making a decision is associated with their needs for *why not* explanations.

The remainder of the paper is structured as follows. We first review related work in this area. This is followed by descriptions and results from a preliminary survey, an interview study, and a survey study that are all approved for IRB exemption. We then discuss the implications of our findings from these studies, limitations, and future work in this research direction.

2 BACKGROUND

2.1 Explaining recommendations

Recommender systems (RS) decide what and how online information is presented to individuals for a wide variety of domains including restaurant search, video consumption, dating, shopping, and advertising by tracking and learning from the vast amount of data from user activities. While recommendations can make digital information more consumable and relevant for individuals [10], the decision-making processes of RS can be complicated and opaque in nature. Consequently, the personalized curation of content without proper justifications can make it difficult for users to assess the quality of the recommendations without consulting with external resources [18]. One particular line of research in this area is using explanations and justifications to support users in better understanding the decision-making rationale of RS [34]. Some classic examples of explanations in RS include item-based explanation (e.g., You are being recommended for this product because you have purchased a related product before) and user-based explanation (e.g., You are being recommended for this video because a person you follows have liked this video) [56]. While explanations can aid users in understanding the AI systems, it remains uncertain whether they can lead to improved trust and acceptance of the system [9, 30].

Explanations are essentially an answer to a *why* question [32]. Prior work has adopted the question-driven framework to generate explanations. In particular, Lim and Dey [31] used scenarios to elicit types of information demands and questions that users have when using context-aware applications. Liao et al. [30] developed an explainable AI (XAI) question bank to represent users' needs for explainability as questions that they might ask when using various AI products. Gregor and Banbasat found that when interacting with intelligent systems, users are not likely to ask for or access explanations without a specific trigger, such as anomalous system behavior [20].

In our study, we elicit user interview and survey data to understand perceptions of regular explanations as well as needs for counterfactual explanations. This way, we can better identify pain points across both and see how design modifications in one type of explanation can potentially help solve pain points in the other.

2.2 *Why not* explanations

One particular type of question people may want ask an intelligent system is the *why not* question, which closely relates to the line of research in XAI that focuses on generating contrastive explanations [43]. A contrastive explanation explains not only why event A happened, but why A happened as opposed to some alternative event B. Miller [32] provides an extensive survey of social science literature in relation to XAI. He concluded that when people ask "Why did event P happen?" questions, they are essentially asking "Why did event P happen rather than event Q?", where Q is often implicit in the context. According to social science researchers, contrastive explanations are appealing for two reasons. First, people ask contrastive questions when they are surprised by an event and expected something different. The contrasting case identifies what they expected to happen. This provides a 'window' into the questioner's mental model, identifying what they do not know [29]. Second, offering contrastive explanations is "simpler, more feasible, and cognitively less demanding to both questioner and explainer" [33], as a contrastive question surfaces characteristics that differentiate the actual causal history from its counterfactual alternative [29].

Researchers have proposed to use contrastive explanations for personalizing human-AI interactions via an explanatory dialogue by interactively adjusting their conditional statements and extracting additional explanations [42]. Wilkinson et al.[53] explored the effects of "why" and *why not* justifications on users' perceptions of explainability and trust for a movie recommender. By only providing a *why not* explanation between a recommended movie with a non-recommended movie chosen by the system, they found that such *why not* justifications negatively influenced users' willingness to depend on and follow advice from the system. This further suggests the need to elicit when, where, and how users want explanations for *why not* questions.

Previous work has also underscored the general importance of counterfactual explanations. Elzein argues that the ability to explain a decision contrastively is indispensable for responsible decision-making [15]. In a recent proposal, Attolou also pointed out that for end-users, it is important to understand why the system does not recommend other items that the user might expect (*why not* question) in addition to why the system recommends certain items (*why* question) [4]. Despite the importance of counterfactual explanations, there is little empirical work investigating user needs of such explanations in practice. Our work strives to bridge this gap.

2.3 Supporting lay users in everyday recommendations

Lay users have different explanation needs than AI researchers, developers, and domain experts when interacting with AI-generated content [38]. Average users of intelligent everyday applications may not be aware of algorithmic decision-making [16]. The uncertainty and lack of understanding towards how the RS generates the contents can also lead to "algorithmic anxiety" [22] and "algorithmic aversion" [12]. Eiband [14] investigated and categorized different problems and coping strategies reported by users for three everyday intelligent applications (Facebook, Netflix, and Google Maps). Their findings call for a reconsideration of how and what to explain for recommendations in order to better support users in making decisions across options.

Though explanations may bring desiderata such as transparency, trust, and satisfaction [45], research has shown the importance of presenting only relevant and crucial information. For everyday recommendations, whether *why not* explanations can be useful, and when to provide them to users with everyday recommended content remain understudied. Complicated explanations may also decrease users' acceptance of a system [21] and negatively impact user confidence and enjoyment [39]. Bunt et al. [6] also found that when users are interacting with everyday low-cost applications, the perceived cost of consuming explanations tended to outweigh its benefits. The ever-growing amount of time spent on these applications [7] motivated us to determine whether this is still the case.

While information overload is the main challenge in using explanations to support users' decision-making, over-simplification can also lead to negative effects. In a study that compares personal and impersonal sources of movie recommendations, Kunkel et al. [28] found that simplistic explanations (conventional similarity-based explanations) fall short in explaining the decision of RS when compared to more sophisticated human-like explanations. Pu and Chen [36] found that when users interact with RS, the desirable level of detail in an explanation is associated with the perceived level of risk in executing a decision. For example, users would prefer more detailed explanations for products that involve a high level of financial and emotional risks such as cars and houses, while they would prefer a short and concise explanation on low-risk products such as movies and books.

If explanation detail is impacted by decision risk, then perhaps so is the desire to explore alternative options. Currently, there is little work examining the relationship between counterfactual explanations and cost of decision-making. We engage with this topic through our exploration of decision utility as it pertains to counterfactual explanations.

3 PRELIMINARY SURVEY

To begin our investigation, we conducted an exploratory survey to get some insights about people's general explanation needs for recommended content in everyday applications.

3.1 Survey Procedure

In this short survey, we offered participants a list of seven types of recommendations based on the categorization of recommender systems proposed by Zhang and Chen [57]. For each type of recommendation, we asked how frequently people use it and only analyzed the responses in which the participants indicate that they have at least seen it "infrequently". We recruited participants through Amazon MTurk who reside in the US and are aged 18 or above in August 2021, and we received a total of 91 valid responses. Each participant received $1.5 after completing the survey. Of the 89 participants, 67 identified as male and 24 as female. The average age was 36.2 (std = 9.91). All completed their high school education, 48 have at least a bachelor's degree, and 15 have professional and/or Master's degrees.

The survey started with three sets of scenario-based questions (online shopping, online dating app, online apartment searching). After that, participants were asked to answer some questions about their general explanation needs for different types of recommendations. These questions include: How frequent do you see the following types of recommendations in your everyday life? When the system offers you a particular recommendation, how much would you be interested to get an answer for a why not question such as why not suggest A instead of B?

3.2 Survey Findings

Among the most common types of applications (e-commerce, point-of-interest, social media, and multimedia), social media had the highest percentage (15.5%) of participants who indicated "Not at all interested" with regards to recommendation explanations, while the other three types of applications had less than 9% of participants completely not interested in explanations. For e-commerce, point-of-interest, social media, and multimedia recommendations the percentage of participants who indicated "Very interested" or "Extremely interested" were more than 62%, 40%, 43%, and 48%, respectively. For the scenario-based questions, across all reasons (including **efficiency, effectiveness, persuasiveness, transparency, satisfaction, scrutability,** and **trust** as the seven possible goals of using expalnations in recommender systems identified by Tintarev and Masthoff [45]) for why they need explanations, effectiveness (i.e., The explanation might help me to make good decisions) was chosen by the highest percentage of participants 44% averaged over the three scenarios.

We also looked at the responses to the open question "Please give us a short description of a time when you noticed an explanation for any type of recommendation." All responses fell into the following four categories of applications.

(1) Social media applications (Facebook, Instagram, Twitter) recommending users to follow based on mutual connections
(2) Multi-media applications (YouTube, Spotify) recommending images, songs, movies, or videos based on user history
(3) E-commerce platforms (Amazon) recommending products based on user history and similar products
(4) Point-of-interest applications (Google Maps) recommending restaurants and routes based on users' preferences and past activity

These preliminary findings gave us the impression that users of everyday online service applications generally have a high interest in explanations, and that the commonly noticed explanations are primarily coming from social media, multi-media, e-commerce, and point-of-interest applications. Based on the finding that the most commonly chosen reason for explanation needs is effectiveness (making better decisions), we focus on understanding how *why not* explanations can better aid decision-making going forward.

4 INTERVIEW STUDY METHOD

We chose semi-structured interviews as the study method for its exploratory and versatile nature. We were constrained by a lack of prior empirical work to deductively approach this problem, as everyday user-facing *why not* explanation needs is still under-explored. An inductive approach helped us better delineate people's transparency needs without being influenced by the dominant assumptions on the technical usefulness of counterfactual explanations.

4.1 Recruitment and Participants

We screened and recruited participants who regularly use online service platforms such as e-commerce sites, social media applications, and multimedia streaming services who have noticed and interacted with recommended content. We distributed a message across social media platforms and also leveraged snowball sampling to broaden our reach. To understand how average users seek explanations, we recruited participants who had no extensive knowledge of the algorithmic side of RS. Based on responses to the recruitment survey, we filtered and selected participants based on the quality of their answers to the open question, their knowledge level of how the RS works, and their general awareness of recommended content. We interviewed a total of 12 participants. The interviews were semi-structured and conducted on Zoom, and typically lasted from 45 to 60 minutes. All participants consented to being recorded, and participants were compensated with a $10 Amazon gift card at the end of the interview.

We began by asking interviewees to recall when they last interacted with recommended content on any of the online service applications they use. To provide participants with some references, we complied a list of recommendations including e-commerce, social media, multimedia, point-of-interest, health, academic, navigation, and search engines based on the categorization proposed by Zhang and Chen [57]. We also provided a few application examples (e.g., YouTube, Spotify, Yelp etc.) for each type of recommendation. If participants struggled to think of relevant examples, we would then show them the list as a reference. They were also encouraged to think of examples outside the list. While showing the examples might induce biases, the influence on responses is minimal, as list covers a wide range of commonly used everyday applications that are representative of most everyday interactions with recommendations and explanations.

4.2 Analysis

Each interview audio was recorded and transcribed with some light editing for readability (e.g., removed pauses, repetitions, filler speech, etc.). The lead author then used open coding on the transcripts to generate a set of codes. After the first round of coding, that author revisited all the transcripts to collate the codes, and two authors then imported the codes to an online application, Miro, and created affinity diagrams to identify themes and patterns in participants' responses. The codes were clustered based on similarity using an inductive approach. These clusters were then used to identify emerging themes. During this process, we re-examined the raw data, the coded quotes, and the themes in an iterative manner.

5 INTERVIEW STUDY FINDINGS

5.1 Users' perceptions of existing explanations

5.1.1 *General need for more explanations.*

When being asked if they would prefer to view more or less explanations in their recommendations, most participants express a strong interest and need for additional information.

> I need explanation. Just because most of these websites are not transparent enough. I really value being informed. And for me, I just want to make informed decisions whether I go a lot. (P11)

> Because like, there are so many algorithms that go on behind the scenes or just even showing one little explanation, it just kind of helps just better experience. (P4)

In the above quotes, P4 enjoyed the general sense of transparency when she saw explanatory details alongside the recommendations. Several participants, on the other hand, mentioned that although they preferred to have the information available, they would not want to access explanations every time they were presented with recommended content. For example, P5 wanted more details to be available on demand.

> I don't want to know every single time. But if I have the thought that I want to be able to know everything. [...] Because right now I know that I won't be able to find that information. (P5)

5.1.2 *Lack of relevant details.*

When asked whether they were satisfied with the information provided in existing explanations, participants expressed once again a desire for more details. Many participants found the information currently provided to be too generic and simple. One participant provided an example where the explanations for Facebook advertisements did not appear to be telling the whole truth regarding targeting based on her specific interests:

> So I'm still getting those like tailored recommendations and ads when I click on "Why are you seeing this ad?" All it does is say you're seeing this because the brand is trying to reach females ages 18 and up and people who live in the United States. [...] The explanation is super vague. I'm like, okay, this is a lie. They're not targeting any random person who lives in the United States. (P8)

As P8 suggested, the lack of details in explanations can potentially erode users' trust in the recommender system. P5 also indicated a similar distrust, saying "The less detailed it is, the less I trust that they're actually telling me the truth". Apart from explanations failing to provide desired details, some participants also pointed out that the explanations may not be relevant to their interactions with the application. P9 mentioned that it is unclear how the existing explanations are useful besides offering a sense of transparency:

> I don't know how helpful it would be if it ended up on my, like, on my screen, if I were actually look at it. If they were to say, Oh, we recommended this to you because of this. I'm not sure if I'll like put too much thought into it. (P9)

5.1.3 Existing explanations are not noticed. During the interview, participants were given a few examples of existing explanations. One example frequently noticed by participants was LinkedIn suggesting new connections based on the number of mutual connections. Another less noticed example was on the Instagram explore page, where one small line of explanation that reads "Based on posts you saved" shows up under every recommended image. While most participants noted that they had seen these, several of them did not despite being users of the platform. Participants had high-level awareness that explanations are often displayed with recommended content, yet they did not seem to observe specific instances of such explanations.

> It's like now that you've pointed it out, I'm just realizing that there are already explanations on that I just usually just skip over. [...] There are actual explanations, but I never paid attention to. (P10)

> That's exactly how I always see them just passive. It's just kind of built in. You almost don't even notice it. (P4)

Additionally, P3 and P5 mentioned that they did not actively seek out explanations because they do not think the information is even available in the first place. P9 explained that this lack of notice may be due to the limited attention allocated towards explanatory information when skimming through recommended content.

5.2 Triggers for asking questions about alternative suggestions

Throughout the interview, we observed signals of users wanting to learn what hasn't been shown to them and the reason for that, as well as the triggers. Participants may not be actively aware of missing recommendations (i.e. items that are not shown to them but still have some relevance). Noting this, we ask: *what triggers users to question the existence of potentially relevant and interesting content outside of their current recommendations?* Below, we elaborate on two triggers we identified.

5.2.1 Trigger 1: General skepticism. Some participants tend to be more skeptical toward algorithmic recommendations than others. Their primary concern is whether the algorithm is really giving them suggestions based on their interests. Building on this skepticism, some of them wonder why they have been only given these particular contents instead of others. For example, P3 found himself wondering whether Netflix is only recommending their popular shows when he tries to find something to watch:

> What if I go on Netflix and it only let me watch whatever is the most popular show every time? It feels like it spirals me into this hole of watching whatever is very popular. (P3)

P2 expresses a similar concern when reading news articles on Facebook:

> Why didn't I get this article, whether it's like, 'Oh, we found this one was not factual' or was it biased? What part of it is considered to be biased then?"(P2)

When making a choice based on recommended content, some participants would like to know why an item is recommended over alternatives. P3 mentioned that counterfactual explanations can potentially help him to calibrate the amount of trust he can put into a particular recommended item.

> I guess it would just be useful in understanding like, how much trust I should put in, like how specific the recommendation actually is. (P3)

5.2.2 Trigger 2: Seeking novelty. While recommender systems attempt to predict users' preferences based on their past activities, the recommended content can inevitably become homogeneous and similar. Several participants mentioned that they constantly got tired of recommendations on their feed page and wished to seek fresher alternative content.

> There have definitely been times in the past couple months, when I'm like, you know, bingeing YouTube and I just get bored. And like, I don't know what to I don't know what to watch next. [...] I don't know what to switch to. (P10)

In addition to boredom, some participants expressed their interest in knowing how to acquire alternative suggestions from the system due to the fear of missing out (FOMO).

> So like, a great example is my boyfriend uses TikTok and Twitter, and so do I. [. . .] I don't see as many like, funny jokes and memes that my boyfriend sees. And then he sends it to me. And I'm like, Why? Why don't my timeline show me that? I want that, too. (P5)

Several participants mentioned that they got interested in alternative recommendations when they accidentally saw something new and interesting. For example, P7 expressed an interest in being informed of "things that I don't even know that I don't know." After watching an unexpected video about user experience research that showed up in her YouTube feed, she became interested in knowing what could be some other informative and educational video suggestions that are not showing up in her recommendation feed.

> The term UX was very new to me. And I don't know how it came up in my feeds, but one video I got triggered off a path of like other videos coming in. Then it was a very nice surprise for me that I came across very randomly. I was wondering like how many of these things am I missing out on? [...] It's not so much the element of surprise that I'm looking for in these in these platforms, it's that I want to go down within a narrow field. (P7)

5.3 Cost of decision making as a factor for explanation needs

While we found several motivations and reasons for participants to use counterfactual reasoning, they did not express the desire to use such reasoning in all scenarios. This is partly attributed to the additional information complexity inherent in counterfactual explanations due to the need for comparisons with other items. For example, P12 mentioned that casually ordering food for a quick dinner, he would rely on "gut reaction" and chose a recommended item without referring to explanations. In the same vein, a few participants also mentioned that counterfactual information would appear desirable for a decision if there were significant perceived benefits associated with it.

> The process of looking at different why not explanations feels like it would take a long time. I would care more if I'm making like a pretty big decision out of it. Like buying stuff or like watching a really like long running show, more consequential than like listening to a single song. I don't think I've cared too much about exploring that for Spotify necessarily. It would be more useful for making decisions on Amazon or Netflix. (P3)

P10 mentioned concerns about how recommended content might negatively affect her personal beliefs and mental health. She would like to see a *why not* explanation to quickly grasp a suggested item's uniqueness if that item is "loaded with information". On the other hand, P2 found a balanced perspective to be essential when reading news. Hence, she identified a need for *why not* explanations when she received one-sided news. She wanted to understand what makes a particular side of the perspective suggested to her but not others.

> Like for counterfactuals, if that helps a lot for politics, especially because the one thing I get worried about with recommendation systems is if I only get one side of news, although certain side of news is not usually reputable most of the time, but I still like to try to keep them as balanced as I can. (P2)

6 SURVEY STUDY METHOD

Based on the interview findings, we found that the potential costs and benefits associated with selecting a recommendation (e.g. finding a movie to watch, choosing a product to buy) may affect the user's interest in understanding why alternative options are not being suggested. Behavioral economics literature has introduced the concepts of experienced utility and decision utility, where experienced utility refers to the endpoint of a decision process after the person attains the outcome of that decision, and decision utility refers to the utility signal perceived at the point of choice to guide the decisions [5]. Hence, we hypothesized that **the utility associated with decision-making is correlated with users' needs for *why not* explanations when choosing from a list of recommended items**. To elicit how information demand changes with the potential costs and benefits of a decision outcome, we conducted a second survey study using two different scenarios to test and quantify potential correlation effects between the utility of decision-making and users' *why not* explanation needs for the missing alternative recommendations.

6.1 Survey Design

To test and find evidence for the hypothesis, we designed a survey that uses two different scenarios to elicit users' information needs for *why not* explanations. Using described scenarios instead of deployment of an actual system allow us to more efficiently study and understand the potential association between decision-making utility and explanation needs. We also wanted to focus on the type and content of the explanation (*what to explain*) rather than the presentation format of an explanation (*how to explain*) [13]. Deploying an actual interface might introduce additional factors that are independent of our study focus. The premise of both scenarios is using an everyday Point-of-interest (POI) application to find a restaurant. We intentionally make the two scenarios differ in the perceived utility associated with decision-making. The first scenario is constructed as follows:

> Imagine a time when you are using an application with point-of-interest recommendations (e.g. Yelp, Google Maps, OpenTable, etc.) to ***find a place to eat casually without any special purposes***. You search on the application and scroll down to choose from the list of recommended restaurants.

We call this "casual" scenario S_c. The second scenario differs only in the bold portion, which is replaced with ***find a restaurant for a special event (e.g., group dinner, birthday, anniversary, etc.)***. We call this "special" scenario S_s. For both S_c and S_s, participants were asked to indicate how often they had encountered a similar situation (Likert scale) as well as the questions shown in Table 1.

1) Decision Utility: How much thought would you put into making a choice? 2)

The questions for the utility variables were created based on the definition of experience and decision utility in behavioral economics [5]. The three dependent variables *Understand What*, *Understand Why*, and *Explanation Need* are constructed based on the three main dimensions of *why not* explanation needs observed in the interview findings: interest in knowing what are the alternative recommendations, interest in understanding why they do not show up, and the willingness to read an explanation for alternative recommendations. In addition to having information needs, participants would sometimes have less of interest in knowing why but directly refining the search results to see alternative options. So we added a fourth variable, *User Action* to make a comparison with explanation needs.

One example to illustrate this scenario would be when a user looks through the recommended restaurants nearby, they might want to know why the suggestions are mostly Italian restaurants but not Asian restaurants. By knowing *why not Asian restaurants* in contrast with Italian restaurant suggestions, they can easily deduce what specific aspects lead to those suggestions to determine the quality and relevancy of the suggestions. To reduce order effects [44] caused by the order in which the two scenarios appear, we launched two versions of the survey that reverse the order of S_c and S_s. We conducted three rounds of pilots to modify the phrasing of the question, the structure of the survey, and the construction of the scenarios.

6.2 Participants

We recruited a total of 101 participants from Amazon MTurk in January 2022. All participants were based in the United States and had both overall HIT approval rates of at least 95% and the total number of HITs approved over 5,000. Each participant received $1.80 as compensation for finishing the survey. We did not implement attention check questions in our survey based on a set of studies that advise against inserting such questions, as they can lower data quality later on in the survey [48]. Instead, we removed responses based on the question *How often have you encountered scenarios similar to the one mentioned in this scenario?* If the participant's answer is "Never", their responses might not accurately reflect their actual decision making. We used this as a filtering criterion to exclude their responses from the analysis. In the end, we used a total of 89 valid responses in our analysis. Of the 89 participants, 50 identified as male and 39 as female. The average age was 38.0 (std = 10.1) and all were from North America/Central America. All but one completed their high school education, 56 have at least a bachelor's degree, and 8 have professional and/or Master's degrees.

7 SURVEY STUDY FINDINGS

7.1 t-test Analysis

In this analysis, we wanted to test our intervention by looking at how participants' perceived utility varies between the two scenarios. To identify a statistically significant difference between our independent variables (S_s and S_c) and dependent variables (listed in Table 2), we ran a paired t-test across all variables. The variables exhibit some violations from normality based on Shapiro-Wilk test for normality [41], so we used the Wilcoxon Signed Rank Test [52] for our analysis. The results are shown in Table 2 alongside the means for each variable.

The p-values for all the variables are less than 0.05, suggesting that the participants' responses are significantly different across these two scenarios. Furthermore, means are higher in S_s than S_c in all variables. However, we also observed that there are nontrivial variances across the responses. That is, even though there is a significant difference in the two scenarios for all variables, we cannot yet conclude that there is a correlation between perceived utility and explanation needs. We conducted a regression analysis in the following section to provide further evidence on the variable relationships.

7.2 Regression Analysis

The results of the first analysis demonstrate that participants perceive the utility associated with the decision-making to be different between S_c and S_s, and their likelihoods in exploring, questioning, and seeking an explanation for alternative recommended content are also significantly different across these two scenarios.

Based on this result, to further establish the relationships between the three exogenous variables for utility (Decision Utility, Experience Utility (negative), Experience Utility (positive)) and the four endogenous variables for the different facets of *why not* explanation needs (User Action, Understand What, Understand Why, Explanation Need), we ran an ordinal logistic regression on each pair of variables. Note that in this analysis, we do not analyze responses for S_c and S_s separately, since we observed variances across participants' perceived utility for the two scenarios. In this regression analysis, we combined responses from the two scenarios and looked at the variable-wise correlations. We also ran a linear regression which matched the results of our ordinal regression. The ordinal (and by courtesy, linear) regression coefficients are shown in the heatmap in Figure 1.

From the correlation coefficients, we can see that *Experience Utility (positive)* and *User Action* have the strongest correlation (0.65) across all the pairs. We adopted the range of correlation coefficient values from prior literature [17] to determine the strength of correlation. *Explanation Need* had moderate correlation with the endogenous variables, while the exogenous variables were moderately correlated with *Decision Utility*.

For the variable *User Action*, we found that it is most strongly correlated with *Experience Utility (positive)*. That is, making a better choice could potentially be the strongest motivation for users to take action in refining the results and looking for alternative suggestions that did not show up. Both *Experience Utility (positive)* and *Decision Utility* demonstrate weak to moderate positive correlations with the pair of variables that measure users' likelihood to ask what and why are certain items absent from their recommendation list. Interestingly, *Experience Utility (negative)* and *Understand What* displayed no significant correlation, and Experience Utility (negative) and Understand Why displayed very week correlation. That is, dissatisfaction from making a bad choice was not noticeably linked to users wondering if there were other options besides ones recommended to them. The last variable *Explanation Need* has moderate correlation *Experience Utility (positive)* and *Decision Utility*. That is, the expected utility at the point of decision making and the utility of making a good choice based on experiences have a significant moderate correlation to how much users would want to see a *why not* explanation.

8 DISCUSSION

8.1 Discussion of Interview Findings

In RS, explanations function as an aid for users to make action-oriented decisions, performing as a support mechanism for task goals [37]. In particular, *why not* explanations can shed light on why the algorithm does not suggest certain items, enhancing the level of transparency in algorithmic profiling [3] without disclosing any of the RS's inner workings. We note that we have not yet seen examples of counterfactual explanations deployed in everyday RS;

Table 1: Variable names and corresponding survey questions

Variable	Survey Question	Range
Decision Utility	How much thought would you put into making a choice?	1–5
Experience Utility (negative)	How dissatisfied would you be if you make a bad choice?	1–4
Experience Utility (positive)	How satisfied would you be if you make a good choice?	1–4
User action	How likely would you try to refine the search results to see better options?	1–4
Understand What	How likely would you question for the absence of potential options from your recommendation list?	1–4
Understand Why	How likely would you question the reason for the absence of potential options from your recommendation list?	1–4
Explanation Need	If the application can provide you with an explanation for why certain potential options are not recommended to you how much do you want to see such an explanation?	1–5

Table 2: t-test results for each variable

Variable	Mean in S_s (Variance)	Mean in S_c (Variance)	Difference in Mean
Decision Utility	4.11 (0.72)	3.35 (0.78)	0.76 ***
Experience Utility (negative)	3.10 (1.0)	2.40 (0.77)	0.70 ***
Experience Utility (positive)	3.64 (0.28)	3.37 (0.42)	0.27 ***
User Action	3.52 (0.46)	3.18 (0.60)	0.34 ***
Understand What	2.93 (0.97)	2.61 (0.95)	0.32 **
Understand Why	2.76 (0.84)	2.47 (0.87)	0.29 **
Explanation Need	3.61 (1.4)	3.15 (1.5)	0.46 ***

** for p-value < 0.01, *** for p-value < 0.001 .

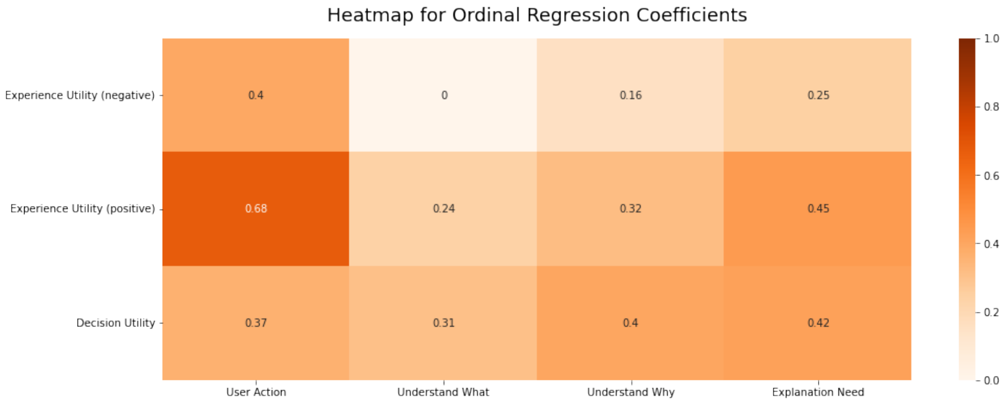

Figure 1: Heatmap showing the ordinal regression correlation coefficients between variables on the y-axis (exogenous variable) and variables on the x-axis (endogenous variable). Note that the cells with 0 indicates that the regression coefficients are not statistically significant (i.e., p-value > 0.05).

however, even without concrete examples, participants identified scenarios in which they would find counterfactual explanations helpful. These insights may be valuable in envisioning what these explanations can look like in the real world.

We first explored participants' experiences with existing explanations. Our participants generally expressed a need for better explanations to aid their daily interactions with recommendations. Kulesza et al. proposed the concepts of soundness and completeness of explanations in intelligent systems [26], where soundness refers to explaining nothing but the truth, and completeness refers to explaining the entire truth. We found that participants tended to trust the truthfulness of the recommendation system's explanations, but they were more skeptical of whether the explanation was disclosing the key reasons that influenced the recommended content. Most participants indicated that they want to see more detailed explanations instead of high-level transparency statements. In addition to that, we also observed a discrepancy between participants' strong need for explanations and how little they actually notice and interact with existing explanations. While there is a near consensus among participants' responses on needing more details from current explanations, participants typically did not usually notice explanations. This shows that while we need better explanations with more actionable details, understanding the specific triggers that prompt users to seek for *why not* explanations will aid designers and engineers in finding the right time and place to offer such explanatory features.

We then identified triggers that prompt users to ask *why not* questions. We found that participants do not always want *why not* explanations in everyday applications, but they would ask *why not* questions and showed interest in seeing *why not* explanations. Participants' needs for *why not* explanations depends on if their perceive the decision-making as relatively consequential. Gregor and Bensabat [20] argue that users are not likely to access explanations without a specific trigger, as the act of accessing indicates a need to learn how a system works. For a relatively low-stakes decision-making domain like interacting with recommendations in everyday applications, learning how the recommender system works may not be top-of-mind for the user. Our observations also echoed with Bunt et al.'s finding that the trade-off between undesirable costs and uncertain benefits associated with reading explanations is the main reason why users do not want more information in certain cases [6]. Our findings point to a major distinction between the interest in knowing more about missing recommendations and the actual need in reading an explanation to know why. That is, while participants might be generally interested in understanding missing recommendations, they might not always want to afford the cognitive effort to to read *why not* explanations. We further found that utility associated with decision-making among various recommended items might be a crucial factor for how much they would want to see a *why not* explanation.

We anticipate that counterfactual explanations will be consumed at a slower rate than regular explanations in scenarios with high decision utility, partially owing to higher cognitive load [11]. This is an acceptable tradeoff given that the resulting decision may have higher consequences. Cognitive load management in counterfactual explanations, however, should be carefully addressed, given that previous work has shown that the presence of uncertainty under high load conditions leads to decreased trust in an RS [59]. In the world of regular explanations, works like COGAM [1] allows for cognitive load to be decreased without decreasing model accuracy, and/or model accuracy to be increased without increasing cognitive load. Envisioning what similar techniques could look like in a counterfactual environment may be a promising area of future work.

8.2 Discussion of Survey Findings

In our survey study, we focused on exploring if correlations exist between people's perceived cost of decision-making and their explanation needs for items outside of the list of recommendations. Our survey showed that the scenario with a higher perceived cost of decision-making (S_s) garnered higher needs for counterfactual explanations and the probing of alternative options. This resonates with the work of Kulesza et al.[27] and Bunt et al.[6] on users' perception of the cost-benefit trade-off of attending to explanations.

Moreover, the statistically significant differences between the two scenarios point to users' desires to potentially see different explanations within the same app, depending on the usage scenario. Explicitly, this has implications on the design of recommendation interfaces and underlying algorithms in "premium-tier" services of apps, such as Luxe from Airbnb [2] and Uber Black from Uber [47]. More implicitly, this calls for a deeper understanding of users' decision-making costs before providing them with recommendation explanations.

We see that *User Action* and *Experience Utility (positive)* has the highest correlation out of all other variable combinations. One possible reason for this observation is that the potential of obtaining higher satisfaction with a recommended item will motivate further action in the form of choice refinement. Correlation between *User Action* and *Experience Utility (negative)* is noticeably lower, suggesting that disappointment from selecting a non-ideal recommended item does not outweigh satisfaction from selecting an ideal one. In fact, the positive correlation difference between *Experience Utility (positive)* and *Experience Utility (negative)* can be seen across all endogenous variables. This may be true for decision-making in everyday apps, particularly on social media, where many interactions are optimized for short-term euphoria [35]. The same cannot be said for high-stakes applications of recommender systems in domains such as disaster response [23] without further study.

We observed no correlation between *Understand What* and *Experience Utility (negative)*. Questioning the absence of potential options from recommended items usually requires at least some existing intuition of other available options. Therefore, if a user is armed with such intuition and would be dissatisfied with a poor choice, it is sensible for them to actively explore other options instead of questioning the items' absence. Interestingly, this is not symmetrical with *Experience Utility (positive)*. Since ratings of *Experience Utility (negative)* and *Experience Utility (positive)* can be made analogous to the risk-reward ratio [40], risk—*Experience Utility (negative)*—may motivate users to act differently than reward—*Experience Utility (positive)*—due to humans' risk-averse nature [24]. Further study in risk-averse behaviors in the context of recommender systems is necessary to confirm this.

9 LIMITATIONS AND FUTURE WORK

In our interview study recruitment, we selected our participants based on their self-reported computer science background, technical literacy, and knowledge of RS, but we did not implement a method (such as a short technical quiz) to verify this. We also had a relatively small sample of participants as we stopped interviewing after we observed data saturation.

While our survey study allowed us to identify strong evidence for correlation, we cannot conclude with certainty any causal effects of decision-making utility on users' *why not* explanation needs. Future work can implement a functional interface with real data to use in our survey to test stronger hypotheses. We defer investigating the relationship between recommendation presentation techniques, such as those suggested by Eiband et al. [13], and counterfactual reasoning to future work.

Additionally, we only considered one type of scenario in our survey. Decision-making associated with restaurant selection may be generalized to other scenarios, but may also be polluted by variations in affordances in different applications (e.g. customizability of item filters in Yelp vs. Netflix). The generalizability of our findings is also limited by demographic distributions we sampled. Future work may be interested in exploring different demographics, such as college students, on an everyday service platform of a different nature, such as an e-commerce website.

Finally, our survey focused exclusively on *why not* explanations and did not make a direct comparison with conventional *why* explanations. Therefore, we cannot determine whether perceived decision-making utility affected different types of explanations similarly or or uniquely affected *why not* explanations.

10 CONCLUSION

In this work, we elicited people's needs for counterfactual (*why not*) explanations when interacting with recommended content in everyday service applications. Through an interview study, we identified two triggers—general skepticism towards recommendations and a desire for novelty—that prompt participants to ask *why not* questions as well as one main factor—perceived decision utility—that positively influences participants' need to use counterfactual explanations if they were provided with one. Based on the findings, we deployed a scenario-based survey study to quantify the correlation between users' decision utility and different facets of their demand for counterfactual explanations. We found significant positive correlations between several variables that provide evidence in support of our hypothesis that users' perceived utility of decision making and corresponding outcomes are associated with their needs for *why not* explanations. Our findings have implications on the future design of decision utility-aware, user-facing recommendation systems in everyday applications.

ACKNOWLEDGMENTS

A part of this project was funded through the National Science Foundation (NSF) award IIS-1910154.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. *COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376615

[2] Airbnb, Inc. [n.d.]. Every home is a destination. https://www.airbnb.com/luxury. Accessed: 2022-01-20.

[3] Oscar Alvarado and Annika Waern. 2018. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 chi conference on human factors in computing systems.* 1–12.

[4] Hervé-Madelein Attolou. 2021. Why-Not explanations for recommenders. In *Actes de la conférence BDA 2021.* 99.

[5] Kent C Berridge and John P O'Doherty. 2014. From experienced utility to decision utility. In *Neuroeconomics.* Elsevier, 335–351.

[6] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces.* 169–178.

[7] L. Ceci. 2021. Average time spent daily on a smartphone in the United States 2021. https://www.statista.com/statistics/1224510/time-spent-per-day-on-smartphone-us/. Accessed: 2022-01-20.

[8] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems.* 1–12.

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

[10] Sarah Dean, Sarah Rich, and Benjamin Recht. 2020. Recommendations and user agency: the reachability of collaboratively-filtered information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 436–445.

[11] Cary Deck and Salar Jahedi. 2015. The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review* 78 (2015), 97–119.

[12] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

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

[14] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When people and algorithms meet: User-reported problems in intelligent everyday applications. In *Proceedings of the 24th international conference on intelligent user interfaces.* 96–106.

[15] Nadine Elzein. 2019. The demand for contrastive explanations. *Philosophical Studies* 176, 5 (2019), 1325–1339.

[16] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. " I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems.* 153–162.

[17] James D Evans. 1996. *Straightforward statistics for the behavioral sciences.* Thomson Brooks/Cole Publishing Co.

[18] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.

[19] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167, 2014 (2014), 167.

[20] Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.

[21] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work.* 241–250.

[22] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic anxiety and coping strategies of Airbnb hosts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–12.

[23] Nan Jing, Yijun Li, and Zhao Wang. 2014. A context-aware disaster response system using mobile software technologies and collaborative filtering approach. In *Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design (CSCWD).* IEEE, 516–522.

[24] Daniel Kahneman and Amos Tversky. 2013. Choices, values, and frames. In *Handbook of the fundamentals of financial decision making: Part I.* World Scientific, 269–278.

[25] Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth. 2021. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. *arXiv preprint arXiv:2103.01035* (2021).

[26] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 1–10.

[27] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing.* IEEE, 3–10.

[28] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–12.

[29] David K Lewis. 1986. Causal explanation. (1986).

[30] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

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

[32] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[33] Tim Miller. 2021. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review* 36 (2021).

[34] 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, 3 (2017), 393–444.

[35] Elliot Panek. 2014. Left to Their Own Devices: College Students' "Guilty Pleasure" Media Use and Time Management. *Communication Research* 41, 4 (2014), 561–577. https://doi.org/10.1177/0093650213499657

[36] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*. 93–100.

[37] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[38] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI.. In *IUI Workshops*, Vol. 2327. 38.

[39] James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek Abdelzaher, and John O'donovan. 2015. Getting the message? A study of explanation interfaces for microblog data analysis. In *Proceedings of the 20th international conference on intelligent user interfaces*. 345–356.

[40] Wolfram Schultz. 2015. Neuronal reward and decision signals: from theories to data. *Physiological reviews* 95, 3 (2015), 853–951.

[41] S. S. SHAPIRO and M. B. WILK. 1965. An analysis of variance test for normality (complete samples)†. *Biometrika* 52, 3-4 (12 1965), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

[42] Kacper Sokol and Peter Flach. 2020. One explanation does not fit all. *KI-Künstliche Intelligenz* 34, 2 (2020), 235–250.

[43] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access* 9 (2021), 11974–12001.

[44] Fritz Strack. 1992. "Order effects" in survey research: Activation and information functions of preceding questions. In *Context effects in social and psychological research*. Springer, 23–34.

[45] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 399–439.

[46] Nava Tintarev and Judith Masthoff. 2015. Explaining recommendations: Design and evaluation. In *Recommender systems handbook*. Springer, 353–382.

[47] Uber Technologies, Inc. [n.d.]. Uber Black. https://www.uber.com/us/en/ride/uberblack/. Accessed: 2022-01-20.

[48] Dave Vannette. [n.d.]. Using Attention Checks in Your Surveys May Harm Data Quality. https://www.qualtrics.com/blog/using-attention-checks-in-your-surveys-may-harm-data-quality/. Accessed: 2022-01-02.

[49] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

[50] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

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

[52] Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6 (1945), 80–83. http://www.jstor.org/stable/3001968

[53] 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 (TOIS)* 39, 4 (2021), 1–21.

[54] Michele Willson. 2017. Algorithms (and the) everyday. *Information, Communication & Society* 20, 1 (2017), 137–150.

[55] Jingjing Zhang and Shawn P Curley. 2018. Exploring explanation effects on consumers' trust in online recommender agents. *International Journal of Human–Computer Interaction* 34, 5 (2018), 421–432.

[56] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192* (2018).

[57] Yongfeng Zhang, Xu Chen, et al. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.

[58] Ellin Zhao and Roykrong Sukkerd. 2019. Interactive explanation for planning-based systems: WIP abstract. In *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*. 322–323.

[59] Jianlong Zhou, Syed Z Arshad, Simon Luo, and Fang Chen. 2017. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP conference on human-computer interaction*. Springer, 23–39.