8-2023

Understanding the Role of Interactivity and Explanation in Adaptive Experiences

Lijie Guo
lijieg@clemson.edu

Understanding the Role of Interactivity and Explanation in Adaptive Experiences

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Human-Centered Computing

---

by
Lijie Guo
August 2023

Accepted by:
Dr. Bart Knijnenburg, Committee Chair
Dr. Nathan McNeese
Dr. Guo Freeman
Dr. Marten Risius

Abstract

Adaptive experiences have been an active area of research in the past few decades, accompanied by advances in technology such as machine learning and artificial intelligence. Whether the currently ongoing research on adaptive experiences has focused on personalization algorithms, explainability, user engagement, or privacy and security, there is growing interest and resources in developing and improving these research focuses. Even though the research on adaptive experiences has been dynamic and rapidly evolving, achieving a high level of user engagement in adaptive experiences remains a challenge. This dissertation aims to uncover ways to engage users in adaptive experiences by incorporating interactivity and explanation through four studies.

Study I takes the first step to link the explanation and interactivity in machine learning systems to facilitate users' engagement with the underlying machine learning model with the Tic-Tac-Toe game as a use case. The results show that explainable machine learning (XML) systems (and arguably XAI systems in general) indeed benefit from mechanisms that allow users to interact with the system's internal decision rules.

Study II, III, and IV further focus on adaptive experiences in recommender systems in specific, exploring the role of interactivity and explanation to keep the user "in-the-loop" in recommender systems, trying to mitigate the "filter bubble" problem and help users in self-actualizing by supporting them in exploring and understanding their unique tastes.

Study II investigates the effect of recommendation source (a human expert vs. an AI algorithm) and justification method (needs-based vs. interest-based justification) on professional development recommendations in a scenario-based study setting. The results show an interaction effect between these two system aspects: users who are told that the recommendations are based on their interests have a better experience when the recommendations are presented as originating from an AI algorithm, while users who are told that the recommendations are based on their needs have

a better experience when the recommendations are presented as originating from a human expert. This work implies that while building the proposed novel movie recommender system covered in study IV, it would provide a better user experience if the movie recommendations are presented as originating from algorithms rather than from a human expert considering that movie preferences (which will be visualized by the movies' emotion feature) are usually based on users' interest.

Study III explores the effects of four novel alternative recommendation lists on participants' perceptions of recommendations and their satisfaction with the system. The four novel alternative recommendation lists (RSSA features) which have the potential to go beyond the traditional top N recommendations provide transparency from a different level — how much else does the system learn about users beyond the traditional top N recommendations, which in turn enable users to interact with these alternative lists by rating the initial recommendations so as to correct or confirm the system's estimates of the alternative recommendations. The subjective evaluation and behavioral analysis demonstrate that the proposed RSSA features had a significant effect on the user experience, surprisingly, two of the four RSSA features (the "controversial" and "hate" features) perform worse than the traditional top-N recommendations on the measured subjective dependent variables while the other two RSSA features (the "hipster" and "no clue" items) perform equally well and even slightly better than the traditional top-N (but this effect is not statistically significant). Moreover, the results indicate that individual differences, such as the need for novelty and domain knowledge, play a significant role in users' perception of and interaction with the system.

Study IV further combines diversification, visualization, and interactivity, aiming to encourage users to be more engaged with the system. The results show that introducing emotion as an item feature into recommender systems does help in personalization and individual taste exploration; these benefits are greatly optimized through the mechanisms that diversify recommendations by emotional signature, visualize recommendations on the emotional signature, and allow users to directly interact with the system by tweaking their tastes, which further contributes to both user experience and self-actualization.

This work has practical implications for designing adaptive experiences. Explanation solutions in adaptive experiences might not always lead to a positive user experience, it highly depends on the application domain and the context (as studied in all four studies); it is essential to carefully investigate a specific explanation solution in combination with other design elements in different fields. Introducing control by allowing for direct interactivity (vs. indirect interactivity) in adaptive

systems and providing feedback to users' input by integrating their input into the algorithms would create a more engaging and interactive user experience (as studied in Study I and IV). And cumulatively, appropriate *direct interaction* with the system along with *deliberate and thoughtful designs of explanation* (including visualization design with the application environment fully considered), which are able to arouse user reflection or resonance, would potentially promote both user experience and user self-actualization.

Dedication

To my parents, Chunyan and Shengcai, for their unconditional support and love.

To my little one and my husband, for their encouragement and support that keep me going and never giving up.

To my advisor Bart and my teammate Shahan, and other HATLab members for their invaluable support and help.

Acknowledgments

I would like to express my sincere gratitude to my advisor, Dr. Bart Knijnenburg, for his continuous encouragement, patience, and support of my Ph.D. research and my life. I really appreciate that he spent much time on guiding me to explore recommendation algorithms, conduct studies, and write good papers. He also gave me maximum of flexibility on balancing my Ph.D. pursuing and my life with my little one.

I also appreciate the support from my committee members (Dr. Nathan McNeese, Dr. Guo Freeman, and Dr. Marten Risius). I am thankful for the collaborations with Dr. Nathan McNeese and Dr. Marten Risius on two different projects, which helped me gain new knowledge and skills; I thank Dr. Guo Freeman for her encouragement when I took the class (HCC 8810 Selected Topic: Online Relationship) with her in Fall 2019.

I am grateful to Shahan for his patience and his expertise in user interface implementation, he helped me a lot with setting up the movie recommender systems during our collaboration. I thank Aminata for her professional comments on my dissertation. I am also grateful to Daricia, Moses, Reza, Christopher, Sushmita, Karishma, and Pratitee for their contributions to my research projects and papers. I would like to also thank all the HATLab members for their professional feedback and suggestions on my research studies.

Last but not the least, I would like to thank my family and my friends for their constant support, encouragement, and love.

Financial support was provided by National Science Foundation under grand no.1565809 and no.2045153, etc.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Motivation

With the growth of the Internet and the development of new technologies such as machine learning (ML), artificial intelligence (AI), and the Internet of Things (IoT), adaptive experience which seeks to create a personalized experience for each user by adapting to their individual characteristics, behaviors, and preferences [81] has been an active area of research in the past few decades. Personalization algorithms are among the most popular research trends and developments in adaptive experiences. There is a continued focus on developing and improving personalization algorithms that can provide a customized experience which creates relevant and personalized recommendations such as targeted advertising [77, 210, 239].

However, despite the potential benefits of adaptive experiences enabling personalization and context-aware interactions with users, achieving high levels of user engagement in such adaptive systems remains a challenge. For example, most users do not understand the rationale behind the recommendations made by adaptive systems, which can negatively impact their trust in the system [44, 2]; some users may even perceive adaptive systems as intrusive, which can lead to a decrease in their trust and overall satisfaction with the system [213].

The lack of transparency and control in these systems can have a significant impact on users' trust, engagement, and satisfaction with the systems. Thus, finding solutions to improve transparency and control into adaptive systems is beneficial to build positive user experience with these systems.Transparency refers to the degree to which system behavior is visible and under-

standable to users [214]. This includes providing clear explanations of how a system works, what data is being collected, and how it is being used [25, 14, 111]. Transparency can help users make informed decisions about technology use and build trust in the system [67, 247]. Control, on the other hand, refers to the ability of users to influence the system's behavior, such as providing users with customization options, settings and preferences, and the ability to adjust system behavior to suit their needs [13, 136]. Control can help users feel empowered and in charge of their interactions with technology [57, 150, 64, 185]. While both transparency and control have been identified as important factors for user trust and user satisfaction in other domains such as social media and online privacy [247, 246, 219, 21, 217, 130], their potential role in adaptive experience remains underexplored.

Providing explanation and allowing interactivity provide potential solutions to the lack of transparency and control in adaptive systems. Explanation promotes transparency by providing clear and comprehensive information. When explanations are provided, it increases transparency by shedding light on the underlying factors, processes, or decisions. This enables stakeholders to gain a deeper understanding of the subject matter.By incorporating interactivity into a system, designers and developers provide users with the means to actively engage with and exert control over the system. Interactivity enables users to provide input, manipulate elements, make decisions, explore, and actively participate in shaping the system's behavior. This interactive control enhances user experiences, promotes customization, and empowers users to achieve their desired outcomes within the system.

Therefore, this dissertation seeks to improve user experience by incorporating explanation and interactivity into the design of adaptive systems so as to help users better understand how the experience is catered to them and, in the meanwhile, allow users to provide feedback or adjust the personalization; ultimately leading to creating a better user experience with adaptive systems.

1.2 Research Objectives — Proposed Solutions

Providing explanation and accommodating interactivity individually have been shown to offer significant benefits in machine learning (ML) systems regarding supporting decision-making and building users' trust in systems [178, 229]. Explanations can help users better understand how a system works. When users have a clear understanding of how a system operates, they can make

more informed decisions about how to interact with it. Further, when users have a sense of control over their experiences, they are more likely to trust the system. This trust can increase the user's confidence in the system and lead to a better user experience.

Thus, my research endeavors to create a better user experience by incorporating explanation and interactivity into the design of adaptive systems. To achieve the research goal, I posit the following research questions:

Overall RQ1: How do interactivity designs contribute to better user experience in adaptive systems? (Chapter 3, 5 and 6)

Overall RQ2: How do explanation solutions influence user experience with adaptive systems? (Chapter 3, 4, 5 and 6)

Overall RQ3: How do the effects of explanation and interactivity on user experiences depend on personal (chapter 3, 5 and 6) and situational context? (chapter 4 and 6)

1.3 Summary of Studies

To answer the research questions, I conducted four studies focusing on different angles to understand the effects of explanation and interactivity on adaptive experiences.

Study I: Building trust in interactive machine learning via user contributed interpretable rules (in Chapter 3). Machine learning technologies are increasingly being applied in many different domains in the real world. Recommender systems are one of the most popular applications of machine learning. As autonomous machines and black-box algorithms began making decisions previously entrusted to humans, great academic and public interest has been spurred to provide explanations that allow users to understand the decision-making process of the machine learning model. Besides explanations, Interactive Machine Learning (IML) seeks to leverage user feedback to iterate on an ML solution to correct errors and align decisions with those of the users. Despite the rise in explainable AI (XAI) and Interactive Machine Learning (IML) research, the links between interactivity, explanations, and trust have not been comprehensively studied in the machine learning literature. Thus, in this study, we develop and evaluate an explanation-driven interactive machine learning (XIML) system with the Tic-Tac-Toe game as a use case, to understand how an XIML mechanism improves users' satisfaction with the machine learning system. We explore different modalities to support user feedback through visual or rules-based corrections. Our online

user study (n = 199) supports the hypothesis that allowing interactivity within this XIML system causes participants to be more satisfied with the system, while visual explanations play a less prominent (and somewhat unexpected) role. Finally, we leverage a user-centric evaluation framework to create a comprehensive structural model to clarify how subjective system aspects, which represent participants' perceptions of the implemented interaction and visualization mechanisms, mediate the influence of these mechanisms on the system's user experience.

This study takes the initial exploratory step in the investigation of the effect of interactivity and explanation in the context of a Tic-Tac-Toe XIML system. The results show that explainable machine learning (XML) systems (and arguably XAI systems in general) indeed benefit from mechanisms that allow users to interact with the system's internal decision rules. While the Tic-Tac-Toe game example system serves a relatively simple scenario (i.e., determining the outcome of a Tic-Tac-Toe game), I find that even in this simple scenario, explanation-driven interactive machine learning (XIML) systems have a better user experience, partially because they encourage users to engage in a mutual feedback loop that helps improve the system's performance. Specifically, XIML systems that allow users to edit the decision rules (as compared to only give feedback on the decision itself) make users feel more in control over the system, which increases the perceived quality of the system's feedback and, in turn, the overall system satisfaction.

The finding of this study demonstrates that introducing explanation and interactivity significantly increases users satisfaction with a machine learning system, especially when users perceive more control over the system. In the remaining studies in this work, I further explore this effect in recommender systems which are a common application of adaptive experiences.

Study II: Studying the effect of recommendation source and justification on professional development recommendations (in Chapter 4). This study was conducted in the process of building a recommender system that provides personalized professional development pathways for high school teachers seeking to increase their disciplinary knowledge and/or their teaching skills. A controlled experiment (N = 190) was conducted to study the effects of the presented justification for the recommendations (teachers' needs vs. their interests) and the presented source of the recommendations (a human expert vs. an AI algorithm) on users' perceptions of and experience with the system. Our results show an interaction effect between these two system aspects: users who are told that the recommendations are based on their interests have a better experience when the recommendations are presented as originating from an AI algorithm, while users who are told that

the recommendations are based on their needs have a better experience when the recommendations are presented as originating from a human expert.

This study focuses on explanation in a recommender system, it suggests that the presentation of recommendations should emphasize their algorithmic nature, and the justification of recommendations should relate back to users' interests over their needs. It implies that while building entertainment-oriented recommender systems, it would provide a better user experience if the entertainment-oriented recommendations are presented as originating from algorithms rather than from a human expert considering that entertainment-oriented preferences are usually based on users' interest.

Study III: Preference Exploration and Development: The Role of Individual Differences (in Chapter 5).

Traditional recommender systems typically align closely with users' current preferences with the sole purpose of appeasing users through easy to consume suggestions. However, this perspective can lead to complacency which hinders opportunities for taste development. To address this, we adopted a multidisciplinary approach by applying psychological insights surrounding self-actualization to the design of alternative exploration features that help users examine and understand their own tastes and preferences. In an online experiment (n=488), I investigated the effect of four novel alternative recommendation lists (RSSA features) on participants' perceptions of the recommendations and the system regarding perceived diversity, recommendation quality, recommendation conformity, taste coverage, system satisfaction, and choice satisfaction. The subjective evaluation and behavioral analysis demonstrate that the proposed RSSA features had a significant effect on the user experience, surprisingly, two of the four RSSA features perform worse than the traditional top-N recommendations on the measured subjective dependent variables while only one of the RSSA features performs slightly better than the traditional top-N, but this effect is not statistically significant. Moreover, the results indicate that individual differences, such as need for novelty and domain knowledge, play a significant role on users' perception of and interaction with the system.

This study considers explanation and interactivity from a different level: the RSSA features reflect transparency of the recommender system by considering four new alternative recommendation lists to reflect what the system learns from users' preferences, rather than just the traditional top-N recommendations; the design of allowing users to correct or confirm the system's estimates

of the alternative recommendations by rating the initial recommendations (so as to get updated recommendations from the system) enables users to indirectly interact with the system.

Even though the results is out of our expectation that the RSSA features do not perform significantly better on users' experience with the system, users do spend more time on interacting with the alternative recommendations and they prefer to select items from the alternative lists. Considering the effect of RSSA features on user experience with the system, I decided to shift away from these RSSA features and focus on using emotion (from the perspective of item feature) for diversification and visualization to investigate the associated effect on the user experience.

Study IV: Testing a diverse, transparent, and controllable movie recommender system (in Chapter 6). Based on the findings from the above three studies, I decided to integrate the explanation-driven (in the form of visualization) interactive mechanisms into a movie recommender system, and present the movie recommendations as originating from algorithms rather than from a human expert (in terms of the findings from study II), and introduce emotion as the movie feature for diversification, which performs as a prerequisite for incorporating visualization (visualized explanation) and interactivity. Online reviews for products and services provide a representation of the emotions that the product/service evoked. Dr. Mokryn has leveraged the content of such reviews to develop an eight-dimensional emotional vector describing every product/service on each of the eight emotions of Plutchik's wheel of emotions [191]. This vector represents the "emotional signature" of the item, which can be considered as a feature of the item. Therefore, I explored whether emotional signatures can be used as a novel selection criteria for users to find, evaluate, and select products and services that meet their preferences in this study. I do this by integrating the emotional signatures into a movie recommender system for diversification, visualization, and user interaction.

While considering emotions for diversification, visualization, and interactivity with the primary goal to optimize user experience, this study also seeks to help users on self-actualizing by supporting them in exploring and understanding their unique tastes by combining three distinct directions into a novel emotion-based recommender systems[1], which can be a potential solution to the "filter bubble" problem.

The results show that introducing emotions for diversification, visualization, and interactiv-

[1]The *emotion-based recommender* here refers to a recommender that re-ranks and diversifies the recommendations based on their emotional signatures

ity indeed benefits recommender systems in individual taste exploration which further contributes to both user experience and self-actualization.

Overall Contribution. I conducted four studies to explore ways to understand the role of explanation and interactivity in adaptive experience. The overall contributions of this dissertation are fourfold: 1) I built different forms of explanations both in a simplified and objective ML application (the Tic-Tac-Toe game) and in complicated and subjective recommender systems; 2) I implemented three interaction elements to encourage user engagement with systems: allowing users to edit (direct control) the system provided rules and responding to their inputs on the fly, enabling users to rate (indirect control) on the initial recommendations to correct or conform the recommendations, and designing a panel to allow users to specify (direct control) their emotion preferences on movies and get the updated recommendations immediately; 3) I show that integrating users' input into adaptive systems and providing feedback to the user accordingly increase users' engagement with the system, which finally contributes to their overall satisfaction with the system; 4) I argue that appropriate direct interactivity along with deliberate and thoughtful designs of explanation would potentially promote both user experience and user self-actualization in adaptive experiences.

Chapter 2

Background and Related Work

2.1 Explanation in Adaptive Systems

An explanation provided by the system can be presented in the form of visualized explanation, textual explanation, or a hybrid explanation design to provide transparency, convey additional information, or justify the system's prediction to the decision maker [184].

2.1.1 Explainable Machine Learning

Given the reliance on Artificial Intelligence (AI) for important decision making problems, the area of explainability has become a crucial subject of research in this area. Explainable Artificial Intelligence (XAI) [90] aspires that an interpretable explanation is available to support any prediction provided by the system. Although AI algorithms often cannot be directly explained [4], XAI methods aim to provide human-readable and interpretable explanations of the decisions taken by these algorithms. Research has shown that, both explainability and interpretability can increase user trust in the system [245]. In recent research, a variety of concrete XAI methods and implementations have been proposed, some of which involve new predictive algorithms where explainability is built in, and others which focus on post-hoc explanations agnostic to the underlying algorithms where the aim is to approximate a supervised ML algorithm by a simpler and more interpretable model. These techniques include calculating feature importance [87, 202], finding similar data instances from past predictions [94], identifying what features are present or missing to support the prediction

for building contrastive explanations [53], and generating interpretable rule-based representations [47, 203]. Rule sets are regarded as an interpretable model class [71] and in the solution of Study I, I have chosen to use the BRCG algorithm [47] as a rule-based explainer, which aims to optimize the trade-off between accuracy and rule set complexity. I go beyond this work in Study I (Chapter 3), though, by explicitly testing the effect of allowing users to *interact* with the rule-based system, i.e. by creating new rules or editing or removing existing rules.

2.1.2 Visualization in Machine Learning

Visualization is an important aspect of ML systems. Hohman et al. presented a survey of the role of visual analytics in deep learning research [102]. Importantly, visualization helps with interpretation and promotes trust. For instance, Kaur et al. [118] mentioned that visualizations may help data scientists uncover issues with datasets or models, although the existence of visualizations can also lead to cases of over-trust. Furthermore, Hohman et al. [101] argued that data scientists have different reasons to interpret models, often balancing competing concerns of simplicity and completeness. Herlocker et al. presented an evaluation of a "white box" conceptual model of recommendations, as opposed to the typical black box approach [98]. Their results demonstrate that explaining recommendations through some form of visualized interfaces improves users' acceptance of a predicted rating, which is in line with Middleton et al.'s findings [166].

Visualization supports explainability and enables interactivity with the underlying ML model. In this light, Gretarsson et al. introduced an interactive graph-based interface for a movie recommender system and conducted a user study focusing on the interactive visualization design. Their findings highlight that the visual interactive interface helps to produce recommendations with higher accuracy and make the predictions more acceptable [83]. Bostandjiev et al. conducted an evaluation to compare different interactive and non-interactive hybrid strategies for computing recommendations over diverse semantic and social web APIs [33]. They found [120] that integrating explanation and interaction in a visual representation of the hybrid system improves the relevance of predicted content, thereby increases users' satisfaction.

In comparison with existing work on visualization, the solution presented in Chapter 3 (Study I) uses visualization not only to *explain* the ML model's decisions, but also for allowing users to *interact* with the underlying model through Boolean rules. In Study I, I aim to investigate if intuitively visualizing the decision-making process of an ML model and allowing users to interact

with it increases their engagement with the underlying system and improves their overall satisfaction.

2.1.3 Justifying Recommendations

Most recommender systems are black boxes, giving users little insight into how the system has modeled and acted upon their preferences. Providing explanations in recommender systems can increase users' perception of transparency and trust [231, 232]. Initial efforts to explain recommendations can be traced back to recommender systems for news articles, books, and movies [28, 173, 99]. In the two decades following these works in [28, 173, 99], a wide variety of different types of explanations has been developed and tested [72, 73].

An important distinction must be made between *explanations* and *justifications* of recommendations: explanations describe the mechanism by which the recommender system arrived at the recommendations, while justifications provide a reason for the recommendation that is independent of the underlying algorithm [225]. Another important distinction considers the goal of the justifications: they could be employed in an attempt to promote the recommendations (i.e., convincing users to adopt the recommendations), or they could be designed to increase users' knowledge about the recommendations (i.e., allowing them to make more informed decisions about the recommendations) [27]. The idea that the goal of "good explanation should not be to 'sell' the user on a recommendation, but rather, to enable the user to make a more accurate judgment of the true quality of an item" [27] is more in line with the goal of recommendations for self-actualization.

In Study I (Chapter 3), I investigated the effect of explainability together with interactivity in the context of machine learning; in Study II (Chapter 4), I have examined the effect of justification together with recommendation source in recommending personalized professional pathways in a scenario-based online user study. I consider justifications in that study since it is neither realistic nor necessary for the end-user to understand the exact mechanism by which the recommendations are derived. I specifically aim to help the target users (i.e., high school teachers) make a more accurate judgment of the quality of the recommended pathways.

2.1.4 Visualization in Recommender Systems

Visualization leverages visual representations to facilitate human perception [96], this section specifically focuses on the application of visualization as explanation in recommender systems.

Past research have investigated to increase transparency by explaining the recommendations [98, 27, 244, 233, 6]. Explanations have proven to be helpful in users' decision making and even building trust into the recommendation systems. Zhang and Chen categorized the explainable recommendations into six types: user or item explanation (or example-based explanation), feature-based explanation, opinion-based explanation, sentence explanation, visual explanation, and social explanation [260]. Among this six categories of the explanations in recommender systems, feature-based explanations are explanations of recommendations focused on features relevant to a user or an item [260], Herlocker et.al. showed that a feature-based explanation (an explanation referring to a particular movie feature) can strongly convince the movie consumption [98]. While visual explanations focus on the presentation style of recommendations [39], such as using graphs or other visual elements to explain the recommendations. The presentation style of explanation plays a key role in system credibility, a graph can be interpreted by human eyes much faster than plain text since human eyes can process many visual cues [8]. Al-Taie et. al argued that visualizing explanations with simple graphs performs better than textual explanations [8].

As mentioned above, the work covered in Study IV built a recommender system by introducing movie emotions as a key feature, and subsequently implement the diversification and interaction functions based on this new feature. Also, I consider combining the feature-based explanation and visual explanation to present the recommendations to the end user.

2.2 Interactivity in Adaptive Systems

2.2.1 Interactive Machine Learning

The advent of interpretability mechanisms into machine learning (ML) solutions has opened the opportunity for users to interact with the system in order to provide feedback. Although humans have their own limitations, human expertise can provide complementary perspectives. Therefore, researchers have explored interactivity in ML in different domains.

Fails and Olson presented an interactive ML framework [63, 13], where the ML model was intentionally trained quickly and the results were presented to the user, allowing the user to give feedback, explore the impact of their changes, and then tune their feedback accordingly. This work first popularized the phrase interactive machine learning (IML). The work in [115] allowed users to interact with a classifier's confusion matrix in order to support users to specify their preferences for

classifications at the decision boundaries. However, the level of influence the user has on the ML model is limited to the data points that fall along the decision boundaries. The approach in [196] considered predictive variance for the human and machine at each point to allocate human effect. The findings demonstrate how the role of algorithmic triage in allocating human and computational effort has the potential to yield substantial benefits for the task of automation. ML systems can be optimized to complement humans via the use of discriminative and decision-theoretic modeling methodologies, and the work presented at [251] provides the first systematic investigation of how ML systems can be trained to complement human reasoning.

2.2.2 Combining Interactivity and Explainability in Adaptive Systems

Interpretable and interactive systems have been shown to offer significant benefits for users with respect to factors such as transparency, control, decision support, and trust [178, 229]. In order for a user to correct a system, they must first understand it. The study of integrating interactivity into XAI systems is gaining increased attention. Sangdeh et al. presented the results of a sequence of pre-registered experiments that focus on the number of features and the transparency of the model [192]. They discussed that people can better simulate the predictions of a clear model with few features compared to the predictions of a clear model with more features or the predictions of a black-box model. Madumal et. al. developed a framework to allow a user to interrogate and probe an explanation [151], where their goal is to allow the user to interact with the explanation to ensure the user understands the ML prediction. However, these two solutions focus on comprehensive model understanding, but do not allow the user to provide modifications or corrections to the system. In [139], authors built an explainable version of a model and allowed users to interact with the explanations. The explanations are in the form of key words leading to an email being predicted as spam, and users are given the opportunity to add or remove keywords or adjust the weight each word has on the predicted label. This work demonstrates the potential of explanations to increase the understandability of the model behavior and to serve as a vehicle for interaction.

In comparison with the research detailed above, the solution presented in Study I (Chapter 3) combines interpretability and interactivity through a rule-based, explanation-driven interactive ML system, which allows its users to *inspect* the reasoning behind the predictions made by the underlying ML model in the form of Boolean rules, and allows them to *update* these rules so as to align the system's predictions with the user's decision making process. Study II (Chapter 4)

addresses explainability through needs- vs. interest-based explanations. Study III (Chapter 5) offers transparency through reflection upon the alternative lists that motivated by self-actualization, as well as (indirect) interactivity through interaction with the alternative lists. Study VI (Chapter 6), on the other hand, incorporates explainability and interactivity in a movie recommender system. The visualized explanation and interactivity were able to implemented through the introduction of emotion, these implements thus allow users to understand the novel feature of movies (i.e., emotion) and recognize the diversification nature of the algorithm, as well as enabling users to actively specify their individual emotion tastes on movies.

2.3 Application Context in Recommender Systems

2.3.1 The Source of Recommendations

Past research has found conflicting results on whether it would be better to present the recommendations as originating from an AI algorithm or from a human expert—some have carefully documented cases of "algorithm aversion", where users tend to prefer to receive recommendations from a human rather than an algorithm [34, 114], while others have found situations where algorithmic suggestions are preferred to human suggestions [89, 146, 215]. Research does not reach a consensus regarding the existence of algorithm aversion. On the one hand, a sizable body of work has shown or acknowledged the existence of algorithm aversion [193, 62, 12, 142, 34, 114]. This work shows that individuals are more likely to delegate strategic decisions to other humans rather than to an AI system [142]. A potential reason for this phenomenon may be that the emotional responses to the outcomes of delegated decisions are more intense when responsibility is delegated to another human being rather than to an AI-enabled system. For example, Promberger et al.[193] compared computer-generated recommendations against physician-generated recommendations in the context of physical health. They found that patients trust a physician more, and consequently, are more likely to follow the physician's advice. On the other hand, the same researchers found that patients feel less responsible when following a physician's recommendation rather than an AI agent's recommendation. Such inconsistency makes algorithm aversion an important topic for research.

Contrasting the work on algorithm aversion is a growing body of empirical evidence suggesting that users actually *prefer* algorithmic advice [54, 89, 146, 215]. For example, Dijkstra et al. found that individuals find expert systems more rational than human advisors [54]. Gunaratne et

al. studied human decisions in the context of an online retirement saving system [89], they found that while both types of advice increase users' saving performance, users are more likely to follow advice coming from an algorithmic source rather than crowd-source advice.

More recent studies suggest that the occurrence of algorithm aversion or algorithm seeking behaviors crucially depends on external factors. For example, Beger et al. revealed that users do not prefer human advice to algorithmic advice when they are unfamiliar with the human advisor [24]. Similarly, Castelo et al. explored the 'algorithm aversion' phenomenon in 6 studies with different tasks [35] and found that the algorithm aversion phenomenon is task-dependent. The findings of their 6 studies were demonstrated in a conceptual model suggesting that the *objectivity* of the tasks decreases users' discomfort with the algorithms and increases their perceptions of algorithm effectiveness and willingness to rely on algorithms—for objective tasks, users prefer the algorithm, while for subjective tasks, they prefer human advice. Castelo et al. note, though, that if the algorithm exhibits high affective human-likeness, this reduces the effect.

As a result of a literature review, Jussupow et al. [114] concluded that the existence of algorithm aversion may relate to other parameters such as performance, perceived capabilities, human involvement, human agents' expertise, as well as social distance. Therefore, research needs to investigate these various different parameters and domains to better understand the algorithm aversion phenomenon. Indeed, one of the goals of Study II (Chapter 4) is to understand how users' perceptions of AI algorithm-based recommendations differentiate from those of human expert-suggested recommendations in an educational setting, with either interest-driven justification or needs-driven justification for the recommendations. As Logg et al. pointed out, "algorithm appreciation" may appear in some domains where the algorithm has been historically used and popularly accepted by most people such as weather forecasts [146], which means the prior experience with the algorithmic or human advice could be a confounding factor in the phenomenon of algorithm aversion. Since Study II (Chapter 4) considers such a system recommending personalized professional development pathways that specifically targets high school teachers, it avoids the possible additional effects that come from prior experiences. Arguably, Study II (Chapter 4) reveals implications for the design of recommender systems for professional development and beyond.

2.3.2 Diversifying Recommendations

Diversity has been proven to be one of the effective ways to address the "filter bubble" problem (see more details in 2.4.2.1) in recommender systems [238, 171, 74, 222]. Given that recommender systems have been applied in different context and different algorithms have been developed in recommender systems, the effects of diversification cannot be generalized.

Treviranus et al. studied the effect of artificially incorporating popularity into recommender systems on diversity, the results showed that emphasizing popularity contributes to user homogenization increase [238]. Chaney et al. demonstrated that algorithmic confounding can lead to homogenization of user behavior without increase utility [37]. Gharahighehi and Vens empirically studied balancing diversity and accuracy in session-based recommendation systems and validated this performance on music recommender systems [75]. Their results demonstrated a personalizing diversification idea that hybridizing diversity and accuracy is effective in music recommendations where emphasizing accuracy for the focused sessions (i.e. the user is interested in more focused content) and emphasizing diversity for the broader sessions (i.e., the user is interested in broader content). However, this idea is domain-dependant, it does not apply to news recommender systems [74]. The authors further conducted another study making neighborhood-based session-based recommender systems diversity-aware, aiming to address the "filter bubble" phenomenon in the context of news [74]. They did see evidence of improvement of the diversity measures, however, they did not perform a user study to measure the actual users' perceived diversity since this might differ from the diversity that they measured [60]. Similarly, an empirical study was conducted to examine the effect of multiple news recommender systems on different diversity dimensions. The results demonstrated that the personalized news recommendations did not reduce the diversity [171]. However, the study relies on specific diversity measure that only associates limited users behavior data such as number of articles read or the time spent on, and it was tested on simulated data, which makes the finding less convincing.

While there are evidence showing the positive effects of diversification on "pricking" the "filter bubbles", some research shows signs of negative effect of diversification. Lunardi et al. checked the effects of collaborative filtering algorithms on generating "filter bubbles" in the domain of news content by comparing different diversification strategies [149]. The results reveal that diversification approaches would not always decrease homogeneity of items, one of the reasons relies on the

items' features which have a detrimental role in the diversification process. Sun et al. proposed a cross-domain matrix factorization [133] model leveraging social tags to address the "filter bubble" problem by balancing the recommendation accuracy and diversity and alleviating the recommendation polarity [222]. However, their diversity dimension is complicated for users without expertise.

Although diversity has been extensively studied in recommendation systems with applications in different contexts, limited attention has been put in diversifying on dimensions (such as emotions of items) that are easy to understand by users (since they are the end objective to consume the recommendations) and users' actual perception of the diversity.

In comparison with existing work on diversity, my solution in Study IV (Chapter 6) diversifies the recommendations by the item emotions, which is an intuitive feature for users to understand the diversification mechanism. In the work presented in Study IV (Chapter 6), I aim to investigate if diversifying recommendations by item emotions increases users perceived diversity of the provided options and thus mitigate the feeling of being in the trap of the "filter bubble".

2.3.3 Emotions in Recommender Systems

2.3.3.1 Emotions in Recommender Systems

Some studies have introduced emotions into recommender systems. Moshfeghi et al. considered two emotion spaces extracted from the movie synopsis and the movie reviews together with three semantic spaces to predict the rating of a movie of a given user, aiming to tackle the cold start problem where there is no past rating for an item [174]. Their results shows significant improvement in the accuracy of prediction. Instead of extracting emotions from movie information, Ho et al. consider emotions to recommend movies in a different way. Based on some research findings that colors are strongly correlated to emotions [97] and human emotions can be represented in a natural form of color [152], they built a emotion-based recommender system capturing user emotions by using a sequence of three colours [100]. They introduced the emotions from the perspective of user profile.

Different from the above work, I consider emotions from the perspective of item features extracted from online reviews and only using the emotion feature for diversifying the recommendations instead of using it for predicting.

2.3.3.2 Emotions Evoked from Online Reviews

Online reviews for products or services have become an important source for users to refer to when making a decision as the internet technology permeates our everyday life [175, 68, 56]. Online reviews not only convey information about products' or services' attributes but also users' actual experience with the products or services [9], which reflects users' opinion about the products or services. Ullah et al. argued that online reviews is likely to convey emotional experiences (the evoked emotions during the experience) of the users with the products or services [241]. Compared with face-to-face communication, users can express emotions of their experiences with the products or services in online reviews more freely and completely [230]. Thus, emotions evoked in the online reviews of products or services can be treated as a feature of these products or services.

Mokryn et al. adopted a conceptualization that considers the emotions that the item evokes in users in the form of online review of the item. This conceptualization of emotions as an attribute of the item is used in affective information retrieval, especially multimedia retrieval [169]. They further extracted emotions elicited by the movie from the online reviews, and created an emotional signature — consisting of the eight emotions of the Plutchik's wheel [191] — of the movie by using NRC lexicon for emotion detection in a text [19]. They carefully conducted an online experiment with real participants to validate the convergent validity of the emotional signature (i.e., if the emotional signature actually captures what users perceive it to be from the perspective of emotions), they did so by asking participants the emotions elicited in certain movies, the results demonstrated a significant correlation between the emotional signatures and the participants-stated explicit emotions of the tested movies [169]. Cohen-Kalaf et al. further developed a novel Movie Emotion Map system that enables to view and browse through a large collection of movies according to the movies' emotional characteristics [42]. They performed a qualitative evaluation with 18 target users to examine the effectiveness, efficiency, and users' satisfaction associated with browsing and exploring through movies according to emotions. Their results indicate that users could easily browse through movies according to the visualized landscape in the system and that the tool enabled them to search, filter, and find movies based on their emotional characteristics.

2.3.3.3 Introducing Emotion as an Item Attribute

Based on this prior work of extracting emotional signature from online reviews of movies and verifying it being able to represent one feature of movies [169], Dr. Mokryn's team has sought to collaborate with us to apply emotional signature to recommender systems. We had proposed to build an actual movie recommender system and conduct an online user experiment to test the effect of emotion-based diversification and visualization on user experience with this system. The proposal I have written is available at this link[1]. In this project, I took the lead of designing the user interfaces of the systems and implementing all the algorithms for generating different recommendations. The work of this collaborated study was submitted to RecSys 2023.

Partially motivated by this collaborated project, I was further interested in introducing emotions for diversification, visualization, and interactivity in a recommender system, with the primary goal to optimize user experience. To the best of my knowledge, little research has investigate the combination of these three components in recommender systems. Thus, I also expected experimental designs in the final study (Study IV) would help users on self-actualizing by supporting them in exploring and understanding their unique tastes by combining three distinct directions into a novel emotion-based recommender systems[2], which can be a potential solution to the "filter bubble" problem.

2.4 User Experience with Adaptive Systems

2.4.1 User Experience and User-centric Evaluation in Adaptive Systems

While there is ample work in ML on improving algorithms, few studies investigate ways to make interactive ML systems easy, effortless, and enjoyable to use. This is problematic, as poor user experience can diminish the use of an ML system, even if it provides highly accurate ML results. The user experience model proposed in [95] describes how the user perceives certain objective aspects of the system, such as its interaction and presentation style, from the perspective of pragmatic attributes and hedonic attributes. These user perceptions in turn cause an empirical evaluation in terms of satisfaction. Forlizzi's work proposed a framework that can be applied in different domains

[1] https://docs.google.com/document/d/1vNvhtSODiT5rrcsVq0Ew9gSOiooPUcZPJpc4ZiPMBNE/edit?usp=sharing

[2] The *emotion-based recommender* here refers to a recommender that re-ranks and diversifies the recommendations based on their emotional signatures

to benefit new products and system designs in terms of interactions and user experiences [69].

In this dissertation, however, I use the validated user-centric evaluation framework developed by Knijnenburg et al. [128] to evaluate the user experience with the proposed systems covered in all the four studies. This framework explains the user experience of recommender systems—a typical application area of machine learning—and provides an excellent platform for studies of interactive ML systems. I extend this framework from recommender systems to the XIML context covered in Study I (Chapter 3), and apply it to the results of the user study to understand the user experience of the XIML Tic-Tac-Toe game system. The Tic-Tac-Toe game has been considered in previous works studying AI application due to its simplicity [116, 3]. In Study I, the Tic-Tac-Toe game is used because it is easy to learn how to play the game in real-time, hence no "domain expertise" is needed to understand the basic operations of the implemented system. This setup allows us to target the evaluation towards non-expert XAI users. That said, users' analytical skills, their familiarity with the Tic-Tac-Toe game, and/or their familiarity with XAI systems may very well still have an effect on the system's user experience (cf. [123]).

2.4.2 Filter Bubble and Self-actualization

2.4.2.1 The Filter Bubble Problem in Recommender Systems

Recommender systems have penetrated into every aspect of our daily activities online, especially e-commerce, social networking, and search engines, through which people experience most of their day-to-day online activities. For example, Amazon has increased the number of products that the customer has never purchased before up to 40% with the Amazon Personalize service [216]; Facebook heavily relies on algorithms with their News Feed to predict what the user wants to see [1]; Google offered the personalized search for sign-in users when they have the web history enabled on their accounts, they even have expanded this service for signed-out users since around 2009 [104]. Algorithms have been implemented into recommender systems to learn from users' behavior data and browser history about relevant content of their preference, so as to personalize information to individual taste [5, 131].

With the increasing permeation of personalized recommendation, a potential problem of recommendation systems has gradually surfaced: online users get stuck in a "filter bubble" [187] which isolates users from a diversity of viewpoints, content, and experiences, and thus prevents them

from discovering new and unknown areas of their own taste [124]. This issue has attracted widespread attention from popular opinion [36, 186], and a growing number of researches have investigated in it in different domains [182, 105, 66, 171, 86].

Nguyen et. al validated the occurrence of "filter bubble" in the domain of movie recommender systems through measuring the longitudinal influence of collaborative filtering algorithms, they found that the collaborative filtering algorithms did narrow set of movies over time slightly; however, users who actually consumed the recommended movies experienced less of the narrowing effects [182]. Aridor et al. went further and explained the empirical results of Nguyen et. al's findings by analyzing a model of user decision-making in recommender systems with four central components [15]. Grossetti and Mouza proposed a community-aware model based on the similarities between communities on a large Twitter data set to generate re-ranked lists of recommendations, aiming to weaken the "filter bubble" effect for the affected users [86]. Flaxman et al. examined web browsing histories of 50 thousand online users located in the US in the news domain, they found that the mean ideological distance between individual users did increase in news consumption in social networks and search engines [66]. Another research has investigated the effect of recommendation algorithms on amplification of extremist content, which supports the policy concerns regarding "filter bubbles" in the context of extremist content online [250]. The findings suggest that the recommendation algorithms on YouTube showed signs of promoting far-right materials via the provided recommendations after users view the corresponding videos.

The existence of the "filter bubble" problem has attracted the attention of many researchers. In view of the fact that the recommender system has been applied in different domains, and the negative impact it produces also varies greatly depending on the domain. Therefore, I provide a possible solution to the "filter bubble" problem in the domain of movie recommender systems.

2.4.2.2 Self-Actualization in Recommender System

Theories of Self-Actualization. The concept of self-Actualization was originally introduced by Kurt Goldstein, a German Neurologist and Psychiatrist in the 20th Century. He defined it as "man's' desire for self-fulfillment, and the propensity of an individual to become actualized in his potential" [80]. This concept of Self-Actualization adopted in the context of recommender system is inspired by Abraham Maslow's hierarchy of needs that details five phases of personal development: the physiological needs, the safety needs, the love needs, the esteem needs, and the need for

self-actualization [157]. In his 1965 publication "Self-Actualization and Beyond", Maslow defined self-actualizing people as individuals who "learn through the process of intrinsic learning". He pursues that they "listen to their own voices, take responsibility, are honest, and who work" [156]. In this publication, Maslow further pointed out that of the two types of learning: intrinsic and extrinsic, self-actualizing individuals usually rely of the former. This claim, especially pertaining to the high information retention associated with intrinsic motivation in learning, has since been supported by theories in all areas of human goal pursuit, including the goal content theory (GCT) [49] and the self-determination theory (SDT) [48, 50]. Maslow further pursues that the self-actualizing process is one of finding out who one is, what one is, what one likes and dislikes, what one deems good or bad to them, where one is going and what is one's mission [156]. Recommender Systems, as informational navigation systems par excellence, constitute the perfect tool to assist individuals in successfully completing the above described self-actualization process.

Recommender Systems for Self-actualization (RSSA). As most residents of the developed world are at the highest levels in Maslow's hierarchy of needs [80], the goal of recommender systems for self-actualization (RSSA) is to help users with what Maslow calls their "metamotivation": their motivation to go beyond the need itself and create a situation of constant betterment. Besides, according to Rogers, individuals have an innate drive towards self-actualization, which he defined as the inherent motivation to grow, develop, and strive for personal fulfillment. The actualizing tendency emphasizes the importance of self-awareness, authenticity, and the alignment of one's thoughts, feelings, and actions [206].Thus, by leveraging recommendation algorithms to encompass all aspects of Jameson's ARCADE model — a model of choice support strategies suggests that systems can help users to access (A), represent (R), combine and compute (C), advise about (A), design (D), or evaluate (E) a choice situation [110], RSSA can help users construct preferences based on their long-term goals, which provides them with a plan to more confidently make their life decisions. Moreover, RSSA may help users develop long-term goals that reflect their unique personal tastes by departing from the status quo of turning recommenders into traps.

Gaining insight into users' perceptions of recommender systems for self-actualization (RSSA) will enable us to better support users' preference exploration and development. To this effect, I have completed another research study to develop a validated instrument — the OPAD Scales — that measures user's perceptions of recommendation in the context of online personalized advertising including the perceived self-actualization(see Section 2.5). Particularly, *self-actualization* refers to

the perception of the extent to which personalized ads are attuned to help the consumer meaningfully improve their own lives [124]. A focus on self-actualization encourages ad personalization algorithms to go beyond tailoring the ads to users' short-term likes by simply optimizing the "click-through rate" (CTR) of the presented ads [61, 256]. Among the validated OPAD Scales, some of them were finally used in the final study (Study IV), see more details in Section 2.5.

2.5 The OPAD-Perception Scales

2.5.1 Study Overview

Gaining insight into users' perceptions of online personalized advertising will enable advertisers and social media platforms to better support users' privacy expectations and provide user-friendly interfaces for controlling the ad personalization process. To this effect, I have developed a reliable and validated measurement instrument (the OPAD-Perception scales) to understand users' perceptions online personalized advertising. The OPAD-Perception scales provide useful tools to measure users' experience with mechanisms that specifically allow them to inspect and control the ad personalization process, some of the scales were also used in Study IV (Chapter 6).

2.5.2 Study Background

2.5.2.1 Challenges with Online Personalized Ads

Online advertising has become central to the business model of virtually all social media platforms. To increase the relevance of presented ads, social media platforms tailor the advertised content to consumers' desires. This process is, in essence, an application of *recommender systems* that find appealing items for each user among an abundance of available options—the difference being that the recommended content is sponsored by an advertiser and pushed by the platform rather than requested by the user. Prior work has shown that consumers assess the appeal of an advertised product by conducting a comparison of its fit with their own needs or desires [7, 172, 226]. Thus, in theory, personalized ads should be more representative of consumers' needs which may result in increased ad effectiveness. Indeed, personalization has been shown to increase ad click-through rates [255] and purchase intentions [20] considerably when compared with non-personalized advertisements.

While personalization can increase ad effectiveness, it simultaneously causes privacy concerns about the substantial amount of tracking and targeting needed to drive the personalization process [159, 195, 240]. For one, researchers have shown that extensive data collection and personalization creates an experience that is "creepy" [242, 20, 258, 237]. Moreover, social media users could become trapped in echo chambers and filter bubbles—once they engage with specific ads, the personalization process may create a self-perpetuating tailored experience [205, 257, 187]. Recent work has argued for the inclusion of epistemic goals when considering users' desires [61]. As such, incorporating "self-actualization" would surface advertisements that are more aligned with consumers' long-term ambitions, which could support them in developing, exploring, and understanding their own unique tastes and preferences [124]. Furthermore, recent media attention to the prevalence of data breaches, and the ability of targeted ads to spread misinformation, discriminate, and impact politics, has increased consumer awareness of the risks associated with online advertising [55, 41, 52, 259, 220]. These developments require further academic investigation, and this current study provides researchers with useful tools to measure both the negative and positive perceptions consumers may have of online personalized ads.

2.5.2.2 Providing Transparency and Control

Online personalized advertising is basically a recommender system that dynamically takes input details about an advertisement and the past website viewing behavior of an individual user [243]. To avoid the downsides of personalized ads without losing their benefits, businesses must commit themselves to *transparency* in how they gather customer data and how they use it to provide value. Indeed, regulators have started to enforce policies that force companies to disclose how they collect and use consumer information [180]. Moreover, businesses must provide consumers with adequate mechanisms to *manage* the ad personalization process. Indeed, a research study shows that 85% of consumers want more information and greater control over the data that companies collect on them [207].

To address this desire, social media companies have created transparency tools that give insight into the recommendation process. Platforms like Twitter, Facebook, and Instagram have a "why am I seeing this ad?" feature which offers an explanation for being shown a specific ad. On the research front, a number of studies has investigated bringing transparency into online advertising [188, 144], but research on ad transparency and control is still in its infancy [252], as is

transparency and control in movie recommendations.

2.5.2.3 Measuring Users' Experience with Personalized Ads

While most HCI research on personalized ads has employed single-item measures, researchers in other fields have investigated the underlying factors that contribute to acceptable ad experiences in a more systematic manner. For example, Soh et al. presented the ADTRUST Scale [218], which measures trust in advertising with four distinct factors: reliability, usefulness, affect, and willingness to rely on. They argued that within the context of advertising, trust is composed of beliefs that significantly influence consumers' willingness to act on ad-conveyed information. Trust is an important element of users' perceptions of online ads. For example, John et al. found that trust enhances users' experience with personalized ads [113]. Moreover, Kim et. al found that platform trust and acceptable information flow played a major role in engagement with ads [119].

Taking trust into account may thus contribute to the creation of ads that inform consumers of items of personal interest in a manner they deem acceptable. But a thorough understanding of consumers' perceptions of online personalized ads requires a measurement instrument that moves beyond trust in the ads themselves, and explicitly covers users' perceptions of the ad personalization process. In an effort to reach that goal, this study builds on prior work by acknowledging different aspects of the interaction with online ads that impact the balance between privacy and personalization.

2.5.3 Methodologies

Inspired by prior work in Human-Computer Interaction (HCI) that has developed survey scales as a practical resource for researchers and practitioners (e.g. [58]), in this work I employed a three-stage scale development procedure outlined in Figure 2.1.

2.5.3.1 Initial Qualitative Evaluation of the Scale Items

At the first stage of the study, we reviewed the related literature to identify the dimensions and aspects related to users' perceptions of online personalized ads addressed by others to develop a set of candidate items, generating 58 initial items. Then we proceed to a card sorting study where 35 social media users were asked to sort the proposed initial item set with 58 candidate items into appropriate groups (i.e., factors) so as to replace and restructure ambiguous items. We

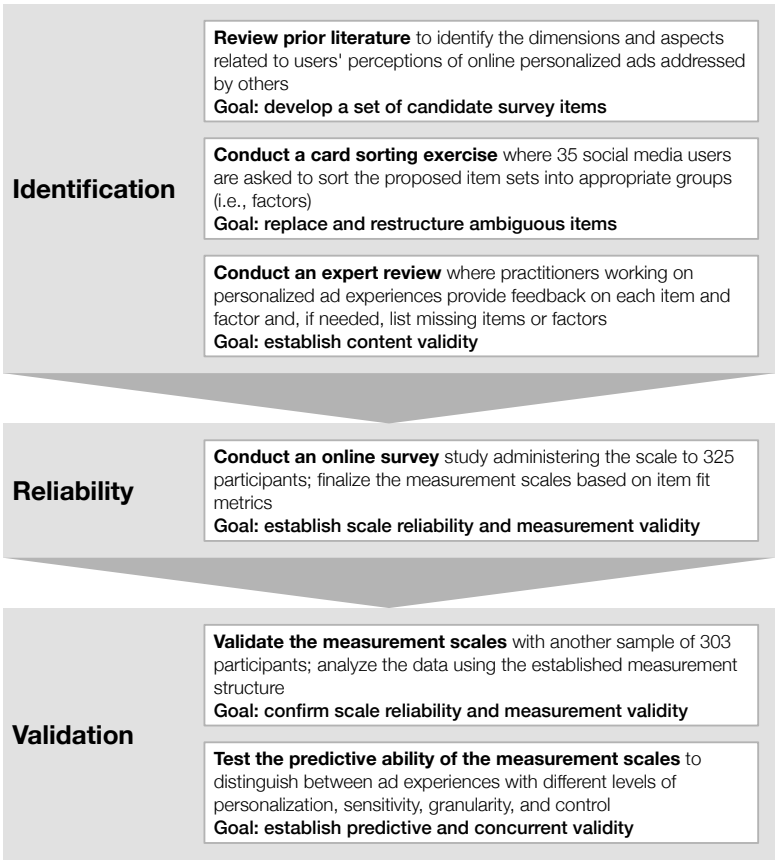

Figure 2.1: Employed three-stage scale development procedure (based on Soh et al. [218]).

proposed three metrics (i.e., misclassification rate, discriminant ratio, and misfit index) to analyze classification results. With the fine-analysed resulting classifications, we further conducted an expert review where practitioners working on personalized ad experiences provide feedback on each item and factor and, if needed, list missing items or factors, in order to establish content validity of the item set.

This stage ended up with 60 items representing ten dimensions/factors describing users' perceptions of online personalized advertising that are ready for the next stage of building reliability.

2.5.3.2 Online User Experiment

At the reliability stage, we conducted an online survey study by administering the scales to 325 participants and finalized the measurement scales based on item fit metrics, aiming to establish the scale reliability and measurement validity. This stage generated a final 10-factor 49-item measurement instrument.

To validate the final measurement instrument resulting from the above reliability stage, we conducted a CFA on the validation sample with 303 online participants, using the exact same measurement structure as the final outcome of the reliability stage. With the CFA analysis on the validation data sample, we confirmed the scale reliability and measurement validity; concurrently, we also tested the predictive ability of the measurement scales to distinguish between ad experiences with different levels of personalization, sensitivity, granularity, and control on the same validation sample, to establish predictive and concurrent validity.

2.5.3.3 The Validated/Resulting Scales

This study resulted in the **Online Personalized Advertising-Perception (OPAD-Perception) Scales**. These scales consist of 49 items that measure users' self-reported perceptions across 10 dimensions (i.e., constructs): *OPAD Reliability*, *OPAD Usefulness*, *OPAD Transparency*, *OPAD Interactivity*, *OPAD Targeting accuracy*, *OPAD Accountability*, *OPAD Creepiness*, *Willingness to rely on OPADs*, *OPAD Self-Actualization*, and *OPAD Persuasion*.

2.5.4 Scales to Apply to the Final Study

The contributions of this measurement study are threefold: 1) we developed a reliable instrument based on qualitative and quantitative input from end-users and experts; 2) we showed that these OPAD-Perception Scales can robustly capture users' multi-faceted perceptions of online personalized ads and their delivery mechanisms; 3) the OPAD-Perception Scales can help researchers, advertisers, and social media platform developers evaluate solutions designed to better meet users' privacy expectations and improve their understanding of and engagement with the ad personalization process on social media platforms. The *transparency*, *interactivity*, *persuasion*, and *self-actualization* scales were also used in the Study IV in Chapter 6 (the *transparency* and *persuasion* scales were integrated to the *understandability* and *choice-satisfaction* scales, respectively) to measure users' perceptions of these dimensions on the movie recommender system.

Chapter 3

Study I: Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules

(Note: This work has been published in the 27th International Conference on Intelligent User Interfaces (IUI '22) [92].)

This work takes the initial exploratory step in the investigation of the effect of the combining explanation (Overall RQ2) and interactivity (Overall RQ1) in Machine Learning systems on user experience in the context of a Tic-Tac-Toe XIML system.

3.1 Introduction

Machine learning (ML) is increasingly being applied in domains like data mining, computer vision, natural language processing, biometric recognition, search engines, medical diagnosis, detection of credit card fraud, securities market analysis, DNA sequencing, speech and handwriting recognition, strategic games, and robotics. Despite the broad application of ML and its apparent

efficiency and high accuracy, ML often remains a "black box", hiding its inner workings from its users. This is not surprising: after all, most end users are not experts in probability theory, statistics, approximation theory, convex analysis, computational complexity theory, and other disciplines that are involved in machine learning, thereby making it difficult to explain how a machine learner works. This lack of explanation can be problematic, though, in situations where human users are expected to give feedback to the ML system—human decision-making is driven by forms of observation, experience and logical thinking, and explanation is at the core of these cognitive processes [103, 147]. Thus, in these situations, it becomes necessary for ML mechanisms to explain their operation to the end user. There are some successful applications of explanatory machine learning (XML) mechanisms, including in the area of music and movie recommendation, mortgage qualification, visual cues to find the "focus" of deep neural networks in image recognition, and proxy methods to simplify the output of complex systems [78].

While these improvements in explainability and interpretability may assist in allowing the user to trust an ML system [79, 29], they do not resolve the often limited ability of ML systems to consume user feedback. When ML systems use data to learn a decision-making process such as a classification task, the accuracy of these systems is dependent on the quality of knowledge captured in the input training dataset. In many cases, though, the available data only have a partial view of the domain. When this is the case, user feedback may be used to improve the underlying solution. Typical mechanisms for supporting user feedback involve allowing users to provide feedback through relabelling instances via active learning, or by adjusting feature importance. However, the knock-on effect and expected impact of such feedback is less transparent to the user, and thus does not allow the user to correct errors or add domain logic. Rule-based models [138, 47] provide an excellent opportunity for domain logic feedback, but their logic is usually static—either crafted by the data scientist into the solution through data selection or as a set of post processing logic rules[1]. While such rule-based models have the advantage of being *interpretable*, in order to achieve coverage, the model must also provide *control* to allow end-users to inspect add rules.

Teso et al. argued that interaction and understandability are central to trust in machine learners [229]. However, few researches have investigated the combination of explanation and interaction in machine learning systems. In this paper, we seek to develop and evaluate a machine learning

[1]One can of course add a new rule for each newly provided ground truth value, but this solution would cover increasingly narrow slices of the realm of possible input parameters, which negatively impacts interpretability.

solution that is *both interactive and interpretable*. We draw on existing work by Fails and Olson, who were the first to introduce the term Interactive Machine Learning (IML) to better integrate the user into the ML solution [63, 13]. IML seeks to include the domain expert in the model generation process by providing a model to the user and allowing them to give feedback on this model, explore the impact of their changes to the model, and then tune their feedback accordingly [13]. We further draw on existing research on eXplainable Machine Learning (XML) mechanisms that present the machine learning model to the user in an intuitively understandable format.

In many XML solutions, visualization plays a crucial role. An analogy can be drawn with how people create a mental representation of their reading when they read a text—this practice helps them improve their critical comprehension, and provides them a chance to interact with the text and make it their own. Likewise, when considering explainability in rule-based machine learning models, intuitive visualizations can improve not only users' basic comprehension of the machine learning rules but also their understanding of how the rules work together, allowing them to make connections between rules. This ability to make connections is extremely important in providing users comprehensive tools to interact with and make changes to machine learning models, i.e. by creating new rules or editing or removing existing rules [22].

In this chapter, I aim to provide such comprehensive means to interact with machine learning models by combining (and thereby taking a step beyond the existing work on) XML and IML. I build upon the user editable AI solution proposed in [45, 11], which enables user feedback through boolean rules, by explicitly testing the effect of allowing users to interact with an intuitively visualized machine learning mechanism. This study makes the following contributions:

1. I develop an eXplanation-driven Interactive Machine Learning (XIML) system that adjudicates the outcome of a Tic-Tac-Toe game which includes:

   (a) an intuitive visualization of the learned rules (as opposed to a textual description of the rules), and

   (b) a mechanism that allows users to interact with the system by editing existing rules (as opposed to only allowing them to indicate the correctness of the rules). More specifically, our eXplanation-driven Interactive Machine Learning (XIML) system allow users to **correct** or **remove** the system provided rules, and even **add** their own rules to the system. Our system will then take these user-provided rules and evaluate how much im-

provement/decreases the user have made to the system's performance and provide this feedback to the use (see more details in Section 3.3.

2. I conduct an online user experiment that independently manipulates the two XIML components outlined above in a between-subjects manner.

3. I evaluate the effect of the visualization and the interaction mechanism on the user's experience by extending the user-centric evaluation framework proposed by Knijnenburg et al. [128] to the domain of XIML.

3.2 User-Centric Evaluation

While there is ample work in ML on improving algorithms, few studies investigate ways to make interactive ML systems easy, effortless, and enjoyable to use. This is problematic, as poor user experience can diminish the use of an ML system, even if it provides highly accurate ML results. The user experience model proposed in [95] describes how the user perceives certain objective aspects of the system, such as its interaction and presentation style, from the perspective of pragmatic attributes and hedonic attributes. These user perceptions in turn cause an empirical evaluation in terms of satisfaction. Forlizzi's work proposed a framework that can be applied in different domains to benefit new products and system designs in terms of interactions and user experiences [69]. In this work, however, I use the validated user-centric evaluation framework developed by Knijnenburg et al. [128]. This framework explains the user experience of recommender systems—a typical application area of machine learning—and provides an excellent platform for studies of interactive ML systems. I extend this framework from recommender systems to the XIML context, and apply it to the results of my user study to understand the user experience of this XIML Tic-Tac-Toe game system. The Tic-Tac-Toe game has been considered in previous works studying AI application due to its simplicity [116, 3]. In our case, the Tic-Tac-Toe game is used because it is easy to learn how to play the game in real-time, hence no "domain expertise" is needed to understand the basic operations of this system. This setup allows me to target the evaluation towards non-expert XAI users. That said, users' analytical skills, their familiarity with the Tic-Tac-Toe game, and/or their familiarity with XAI systems may very well still have an effect on the system's user experience (cf. [123]).

3.3 Online User Experiment

3.3.1 System

This section outlines the system, its algorithm, and its interaction and visualization mechanisms. The system designed for running the user study is based on an ML algorithm that learns to adjudicate the end states of a basic Tic-Tac-Toe game. It consists of two main components: the algorithm used to generate the Boolean rules and the Web application the users interact with.

3.3.1.1 Algorithm

The initial set of rules of the Tic-Tac-Toe algorithm is generated using the Boolean Decision Rules via Column Generation (BRCG) approach [47]. More precisely, I used the open-source implementation available in the AI Explainability 360 library[2]. The algorithm generates a Boolean rule in disjunctive normal form (DNF, i.e., an OR of ANDs) for binary classification, where a data instance that satisfies the DNF rule belongs to the positive class. As DNF classification rules are equivalent to decision rule sets, where each conjunction within the DNF constitutes an individual rule of the rule set, the terms "clause", "conjunction", and (single) "rule" (within a rule set) are used interchangeably.

BRCG was trained on the UCI Tic-Tac-Toe dataset[3]. For the initial training phase I only used 20% of the dataset, so as to produce a rule set with an ample room for improvement. The full dataset consists of the complete set of possible board configurations at the end of Tic-Tac-Toe games, where each configuration is labelled either as "X wins" or "X does not win". A rule set was generated for each one of these two labels. The two rule sets together form the rule-based model of this interactive ML system; users of the system are asked to improve the model by providing feedback on its classifications and/or by tweaking the rules themselves.

Table 3.1 lists the rules generated for the two classes. The accuracy score, which provides the users with a measure on how much they were able to improve the AI system at each turn, is computed by replacing the original rule with the user-modified version and evaluating the new rule set against the entire Tic-Tac-Toe dataset.

[2] `https://github.com/Trusted-AI/AIX360`
[3] `https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame`

Table 3.1: DNF rules generated with the Light BRCG method implemented in the AI Explainability 360 toolkit[2]. Each variable in the rule represents a cell of the board (e.g. mm for the middle-middle cell, tl for the top-left cell, br for the bottom-right cell etc.). The possible values that a variable can assume are b (blank), x and o.

Class	DNF rules
X does not win	$(mm = \text{`}o\text{'} \wedge ml \neq \text{`}o\text{'} \wedge br = \text{`}o\text{'} \wedge bm \neq \text{`}b\text{'} \wedge bl \neq \text{`}x\text{'}) \vee$ $(tl \neq \text{`}x\text{'} \wedge mm = \text{`}o\text{'} \wedge br = \text{`}o\text{'}) \vee$ $(tl = \text{`}x\text{'} \wedge mm = \text{`}b\text{'} \wedge ml \neq \text{`}b\text{'} \wedge bm = \text{`}x\text{'}) \vee$ $(tm = \text{`}b\text{'} \wedge mr = \text{`}x\text{'} \wedge mm \neq \text{`}b\text{'} \wedge ml = \text{`}o\text{'}) \vee$ $(tm = \text{`}b\text{'} \wedge tl \neq \text{`}b\text{'} \wedge mm \neq \text{`}b\text{'} \wedge br = \text{`}o\text{'} \wedge bm = \text{`}x\text{'}) \vee$ $(tm = \text{`}o\text{'} \wedge mr \neq \text{`}b\text{'} \wedge mm = \text{`}o\text{'} \wedge ml = \text{`}o\text{'}) \vee$ $(tm = \text{`}o\text{'} \wedge mr = \text{`}o\text{'} \wedge bl = \text{`}b\text{'}) \vee$ $(tr \neq \text{`}x\text{'} \wedge mm = \text{`}o\text{'} \wedge bl \neq \text{`}x\text{'}) \vee$ $(tr \neq \text{`}x\text{'} \wedge tl = \text{`}o\text{'} \wedge mm = \text{`}o\text{'}) \vee$ $(tr = \text{`}b\text{'} \wedge mr \neq \text{`}o\text{'} \wedge ml = \text{`}o\text{'} \wedge br = \text{`}x\text{'} \wedge bm \neq \text{`}x\text{'}) \vee$ $(tr \neq \text{`}o\text{'} \wedge tl \neq \text{`}o\text{'} \wedge br \neq \text{`}x\text{'} \wedge bl \neq \text{`}b\text{'}) \vee$ $(tr = \text{`}o\text{'} \wedge tm = \text{`}o\text{'} \wedge tl = \text{`}o\text{'} \wedge bm \neq \text{`}o\text{'})$
X wins	$(mm \neq \text{`}o\text{'}) \vee$ $(tr = \text{`}o\text{'} \wedge tl \neq \text{`}o\text{'} \wedge br \neq \text{`}o\text{'} \wedge bl = \text{`}x\text{'}) \vee$ $(tr = \text{`}x\text{'} \wedge br \neq \text{`}o\text{'}) \vee$ $(tr = \text{`}x\text{'} \wedge tm = \text{`}x\text{'} \wedge tl = \text{`}x\text{'})$

3.3.1.2 Web Application: Interaction and Visualization

The front-end Web application of this XIML system presents users with an end-state of a Tic-Tac-Toe game (the "input"), the outcome that follows from the rule set (the "prediction") and the rule that triggered the predicted outcome (the "explainer"). The Web application has two optional features (manipulated between-subjects in this study, see section 3.3.3.3): interaction and visualization.

The "interaction" feature allows users to correct the system-estimated predictions by allowing the user to modify the triggered rule if they disagree with the prediction provided by the system—without interaction, the system only asks the user whether they agree with the rule or not. The system will take users' feedback and evaluate how the updated rule increases or decreases the accuracy of the model. Users can then further modify the rule based on this feedback, or skip to the next instance to be evaluated.

The "visualization" feature shows users an intuitive grid-based version of the applied DNF rule (right side of Figure 3.1). Without the visualization feature, the DNF rule is displayed as text. Note that the display of the interaction feature also depends on the visualization feature: the interactive version of the grid-based visualization allows the user to edit the rule in the grid, while

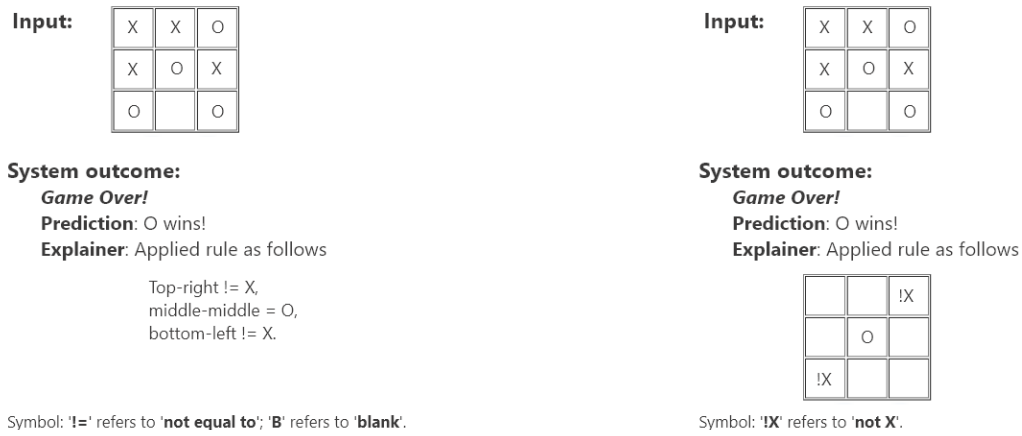

Figure 3.1: The visualization manipulation of textural explained rules (left) and grid-visualized explained rules (right) conditions.

the interactive version of the text-based visualization asks the user to construct the rule using text (see Figure 3.4). More details are provided in section 3.3.3.3.

3.3.2 User-centric Evaluation Framework

Knijnenburg et al. [128] proposed a user-centric evaluation framework to evaluate how users experience recommender systems. I adopt this framework to set up a user experiment to evaluate this XIML system from the perspective of the user. The framework provides a set of structurally related concepts measuring the user experience of the system, which are subdivided into five categories: objective system aspects (OSA), subjective system aspect (SSA), user experience with the system(EXP), personal characteristics (PC), and situational characteristics (SC; not present in this work). I propose to instantiate the framework with the following concepts for the user-centric evaluation of XIML systems:

- In terms of objective system aspects (OSA), I consider the visual design of the explanation and the interaction mechanism by which the user gives feedback to the system.
- I posit that the OSAs influence the understandability of the system, the user's perception of control, the difficulty of giving feedback, and the quality of said feedback—these act as the subjective system aspects (SSA) in the model.
- The user experience (UX) goal of the XIML system is to build the user's trust in the system and increase their satisfaction with the system, hence I propose this as the main user experience

(EXP) construct.

- As noted by Knijnenburg and Willemsen [126, 123], the effectiveness of interaction mechanisms for adaptive systems may crucially depend on the users' personal characteristics (PC), such as age, gender, education level, familiarity with ML/AI, familiarity with XAI in particular, and familiarity with the Tic-Tac-Toe game.

The resulting framework (Figure 3.2) allows us to conduct an empirical evaluation in a more integrative fashion than most existing evaluations of explainable machine learning or artificial intelligence systems. The next subsection describes how I set up and conducted this online user experiment based on the user-centric evaluation framework.

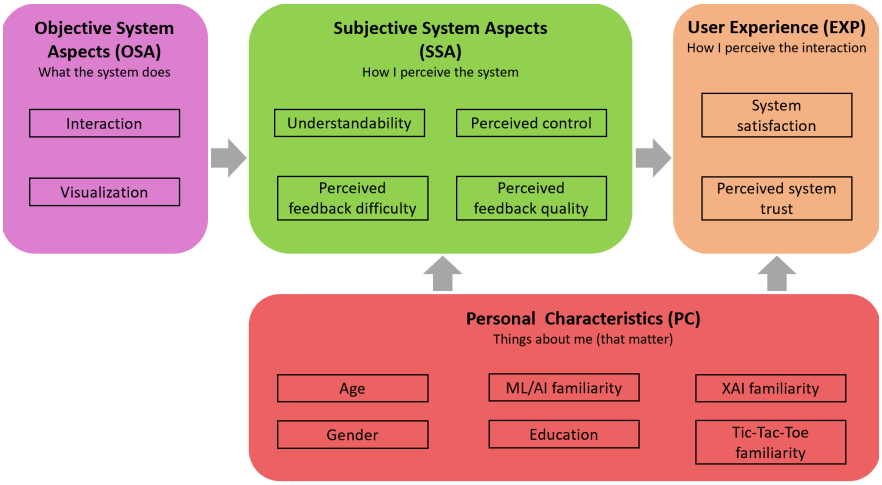

Figure 3.2: The user-centric evaluation framework, based on [128].

3.3.3 Study Setup

In this study, 237 Amazon Mechanical Turk participants answered 6 questions measuring their personal characteristics, interacted with ten instances of the Tic-Tac-Toe game, and answered 34 questions about their user experience and 3 attention check questions. The study took 10-20 minutes to complete.

3.3.3.1 Participants

I conducted the study on Amazon Mechanical Turk, limiting the recruitment to adult users living in the US. Each participant received 3 US dollars as compensation at the end of the study.

Three attention check questions in the style of "Please select 'agree' for this question" were randomly placed among the 34 user experience questions to track if participants were actually paying attention to what they are doing during the completion of the study. I used these questions plus the time taken to complete the study to filter out participants who clearly rushed through the study. Of the 237 recruited participants, 199 were used in the data analysis. 12 participants were between the ages of 18 and 24, 84 between 25 and 34, 60 between 35 and 44, 26 between 45 and 54, and 17 older than 54. The sample included 122 female participants and 115 male participants.

3.3.3.2 Procedure

I randomly assigned the participants to one of the four conditions as mentioned below. Participants were presented a welcome page with an introduction to the study and a consent form containing statements of possible risks, discomforts, and incentives. Participants were then asked to answer six questions about their personal characteristics: gender, age, education level, familiarity with machine learning methods or artificial intelligence (ML/AI familiarity), familiarity with explanatory artificial intelligence (XAI familiarity), and familiarity with the Tic-Tac-Toe game (Tic-Tac-Toe familiarity). The ML/AI familiarity and the XAI familiarity are both on a 5-point familiarity scale[4] while the Tic-Tac-Toe familiarity is on a 4-point familiarity scale[5].

Next, the system presented each participant with ten different simulated Tic-Tac-Toe game instances (as shown in Figure 3.3) with a corresponding classification outcome (the system prediction of the game instance with explained rules) provided by the machine learner. Eight of the ten instances had either an incorrect prediction or an explanation with an incorrect rule, while the other two instances have both a correct prediction and a correct explanation. Following each game instance, participants were asked to answer to correct the prediction and/or to modify the proposed rule, depending on the experimental condition (see below). The system used the participant's corrected prediction and (where applicable) their updated rule to evaluate the impact on the overall correctness of the system. This impact was subsequently shown to the participant as feedback (e.g. "Awesome! Your input has improved the system's correctness by 8.6% on providing the correct prediction and explanation!"), and the participant was then given the opportunity to either edit their input or to

[4]Scale: None, Some familiarity (aware of basic concepts but no hands-on experience), Familiar (some hands-on experience, taken an introduction class, and/or read some literature), Very familiar (published papers on these topics, active ongoing projects), Expert (developed new algorithms, wrote a book chapter, wrote numerous papers)

[5]Scale: None, Some familiarity (aware of it, no playing experience), Familiar (have played at it), Very familiar(good at playing it)

skip to the next instance, until all ten instances were completed.

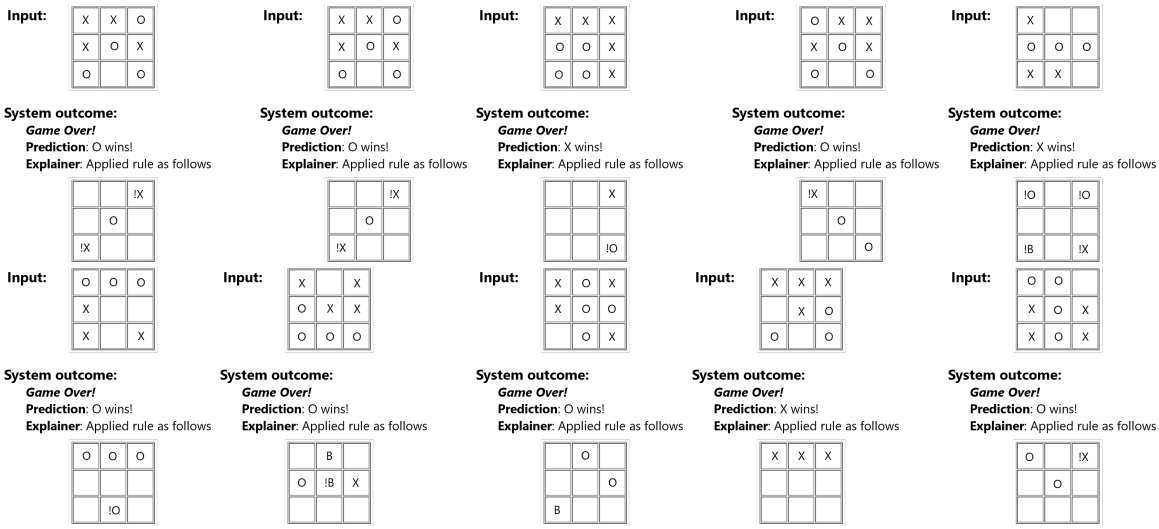

Figure 3.3: Ten different simulated Tic-Tac-Toe game instances with the system-provided grid-visualized explanations.

After each participant completed the action steps for each instance, they were asked to answer the final survey containing 34 questions reflecting on their user experience, plus the three attention check questions.

3.3.3.3 Experimental Conditions

My goal is to evaluate the two inventions described in Section 3.3.1.2: an intuitive grid-based **visualization** of the rule applied to determine the outcome prediction for the current instance and an **interaction** mechanism that allows users to modify the triggered rule if they disagree with the prediction provided by the system. I test these two inventions against reasonable baseline conditions: the grid-based explanation is tested against a textual explanation of the applied rule, as shown in Figure 3.1; the "interaction" version, where participants are allowed to correct the system-provided prediction of the instance and edit the applied rule (displayed in Figure 3.4), is tested against a version where participants are only allowed to correct the prediction.

This experiment thus includes two manipulations with two experimental conditions in each manipulation, resulting in a 2×2 experimental design. The manipulations are assigned using a between-subjects design, in which participants are randomly assigned to one of the four experimental conditions. This ascertains that the manipulation remains hidden from the participants, since each

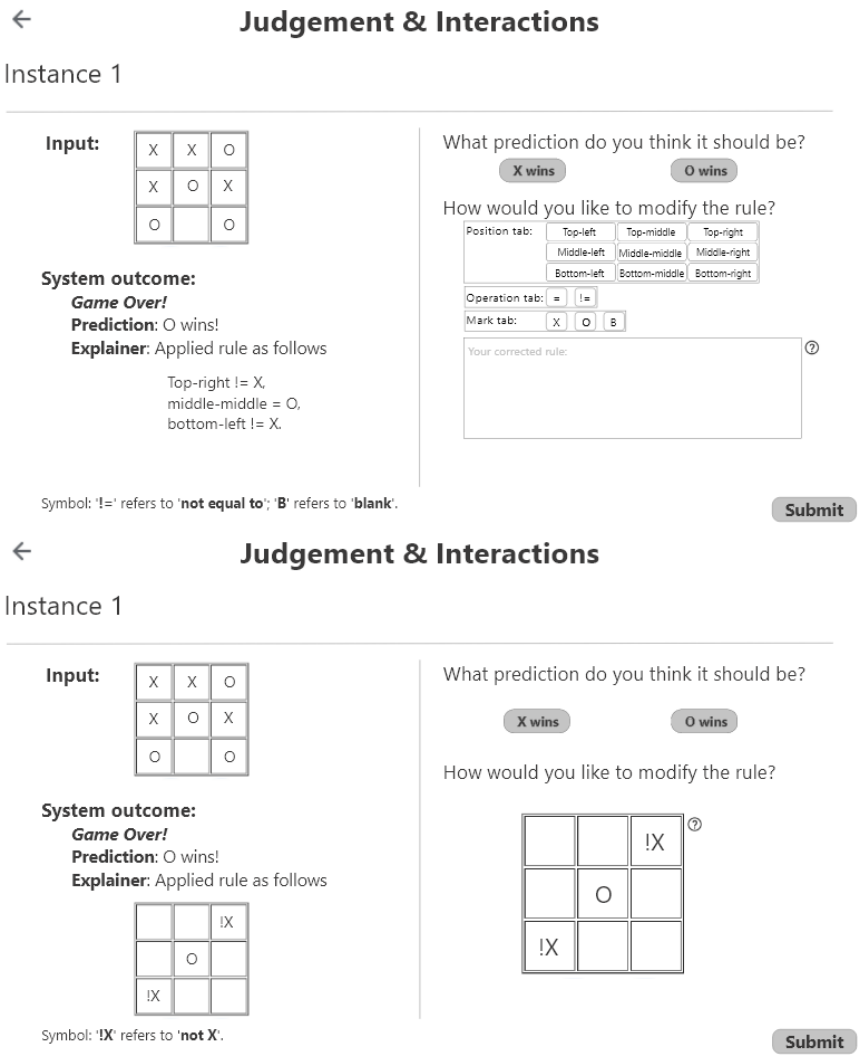

Figure 3.4: The experimental conditions allowing "interaction", with textural explained rule (top) and grid-visualized explained rule (bottom) provided.

participant can see only one condition. Since users of real systems usually only see a single version of a system, such a between-subjects experiment makes the study more realistic [128].

3.3.3.4 Measurements

I measure participants' perceptions of the subjective system aspects (SSA) and user experience with the system (EXP) with six measurement scales, adopted from [128]:

- `Understandability`: participants' perception of the understandability of the XIML system.
- `Perceived control`: participants' perception of their control over the XIML system.

- `Perceived feedback quality`: participants' perception of the quality of the feedback given by the XIML system.
- `Perceived feedback difficulty`: participants' perception of the difficulty of giving feedback to the XIML system.
- `Perceived system trust`: participants' trust in the XIML system.
- `System satisfaction`: participants' satisfaction with the XIML system.

Each scale consists of multiple statements—34 in total (see Table 3.2—that participants are asked to rate on a 5-point agreement scale (strongly disagree, disagree, neutral, agree, strongly agree). An analysis of the validity of these constructs is presented in Section 3.4.

I integrate the experimental manipulations and the measured constructs in a hypothesized path model (Figure 3.5). I hypothesize that the grid-based visualization increases the understandability of the machine learning rules and that it reduces the difficulty of providing feedback to the system (perhaps mediated by understandability). I further hypothesize that participants' ability to edit the rules of the machine learner increases the understandability and makes them feel more in control of the system (again, perhaps mediated by understandability). Understandability is further hypothesized to increase perceived feedback quality, and both feedback quality and control are hypothesized to increase participants' perception of trust in the system. Finally, I hypothesize that perceived control, feedback quality, and trust ultimately increase participants' satisfaction with the system.

I do not formulate specific hypotheses regarding the personal characteristics of the participants; in the results section, these effects are added to the model where significant in an ad-hoc manner.

3.4 Results

I first validated the measurement model regarding the SSA and EXP constructs using a Confirmatory Factor Analysis (CFA) and then fitted a Structural Equation Model (SEM) that describes the hypothesized and ad-hoc causal relationships between the two manipulations, the subjective constructs, and the measured personal characteristics. An SEM can be conceptualized as a series of linear regressions between latent (SSA, EXP) and observed (OSA, PC) variables.

Table 3.2: Items presented in the final survey. Items without a factor loading were excluded from the analysis.

Considered aspects	Items	Factor loadings
Understandability (SSA) AVE: NA	I liked the explanations provided by the system. I found the explanations appealing. The explanations unravel my confusion with the system. The explanations were necessary. The system provided too many unnecessary explanations. I did not like any of the provided explanations.	
Perceived control (SSA) AVE: 0.657	I had limited control over the way the machine learner made predictions. The system restricted me in my interactions with the machine learner. Compared to how I normally get predictions, this system was very limited. I would like to have more control over the interactions. I decided which information was used for predictions.	0.829 0.814 0.789
Perceived feedback quality (SSA) AVE: 0.614	I liked the feedback options provided by the system. I found the option of giving feedback appealing. The ability to provide feedback unravels my confusion with the system. The ability to provide feedback was necessary. The system provided too many unnecessary means to provide feedback. I did not like any of the provided feedback processes.	0.868 0.831 0.630
Perceived feedback difficulty (SSA) AVE: 0.657	I was in doubt between several feedback options the system provided. I changed my mind several times before providing feedback. The task of providing feedback was overwhelming. It was easy to decide how to provide feedback. Providing feedback to the machine learner took a lot of effort.	0.776 0.801 0.819 0.844
Perceived system trust[88] (EXP) AVE: NA	I believe that there could be negative consequences when using the system. I feel I must be cautious when using the system. I believe that the system will act in my best interest. I think that the system is competent and effective in its interaction.	
System satisfaction (EXP) AVE: 0.781	I would recommend the system to others. I like using the system. Using the system is a pleasant experience. Overall, I am satisfied with the system. I would quickly abandon using the system. I could quickly abandon using this system. Using the system is annoying. The system is useful.	0.869 0.934 0.888 0.885 0.841

Table 3.3: Factor-fit metrics. Off-diagonal values are correlations, diagonal values are the square roots of the average variance extracted $\left(\sqrt{AVE}\right)$ per factor.

	Perceived control	Perc. feedback quality	Perc. feedback difficulty	System satisfaction
Perceived control	**0.811**	-0.323	0.288	-0.238
Perceived feedback quality	-0.323	**0.783**	0.122	0.792
Perceived feedback difficulty	0.288	0.122	**0.811**	0.042
System satisfaction	-0.238	0.792	0.042	**0.884**

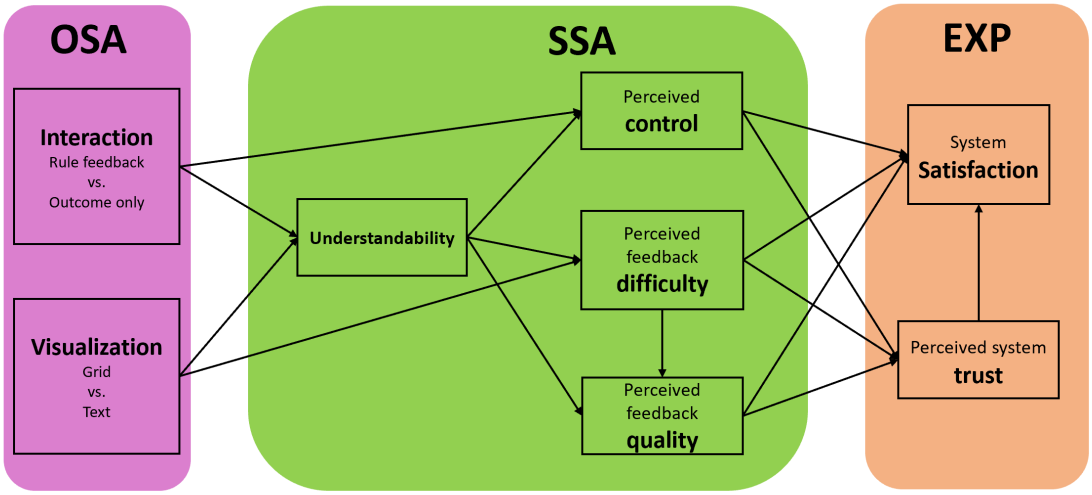

Figure 3.5: The hypothesized research path model.

3.4.1 Measurement model (CFA)

In fitting the initial CFA, I found that the understandability scale was too highly correlated with both the perceived feedback quality scale ($r = 0.838$) and the satisfaction scale ($r = 0.873$), thereby violating the principle of discriminant validity (rs larger than the $\sqrt{AVE}$ of understandability, which is 0.781). Similarly, the trust scale is too highly correlated with the satisfaction scale ($r = 0.901$, while the $\sqrt{AVE}$ of trust is only 0.809). I therefore remove understandability and trust from the initial CFA model to guarantee that the resulting CFA model meets the requirement of discriminant validity. I also removed 2 items from the perceived control scale, 1 item from the perceived feedback difficulty scale, and three items from the perceived satisfaction scale, due to low commonality (< 0.3) or high modification indices (both of which indicate misfit). Table 3.2 displays the item fit metrics and Table 3.3 displays the factor fit metrics of the resulting CFA model.

3.4.2 Structural model (SEM)

The remaining subjective constructs (i.e., perceived control, perceived feedback difficulty, perceived feedback quality, and system satisfaction) are then structurally related to each other and to the experimental manipulations (i.e. the "interaction" and "visualization" manipulations) in a structural model based on the hypothesized path model (Figure 3.5). In line with suggestions by Knijnenburg and Willemsen [127], I created a *saturated* model (with as many estimated parameters linking OSAs to EXP variables via SSAs as there are constructs included in the model) and

then trimmed the non-significant effects from the saturated model iteratively. Significant effects of personal characteristics are subsequently added on an ad-hoc basis.

The resulting structural model is displayed in Figure 3.6. This model has a good overall model fit with $\chi^2(161) = 232.852$, $p < 0.001$, CFI = 0.980, TLI = 0.989, RMSEA = 0.047 with a 90% confidence interval of [0.033, 0.060]. According to Bentler et al., theoretically, a good model is not statistically different from the fully specified model (i.e., the p-value of the χ^2 should be > 0.05), but this statistic is commonly regarded as too sensitive [23]. As such, Hu and Bentler proposed cut-off values for the alternative fit indices to be: CFI > 0.96, TLI > 0.95, and RMSEA < 0.05, with the upper bound of its 90% CI falling below 0.10 based on extensive simulations [106].

In the model in Figure 3.6, the path coefficients in the final SEM model are standardized. This means that coefficients on all arrows (A $\rightarrow$ B) denote the standardized increase or decrease in B, given a one standard deviation (1 SD) increase or decrease in A (except for the effects of OSAs, where the coefficients represent the standardized difference between the two experimental conditions). The number in the parentheses denotes the standard error of this estimate, and the asterisk marks beside the parentheses denote the statistical significance of the effect. The R^2 of each subjective construct denotes the proportion of variance of that construct that is explained by the model. The significant paths in the model are explained in more detail below.

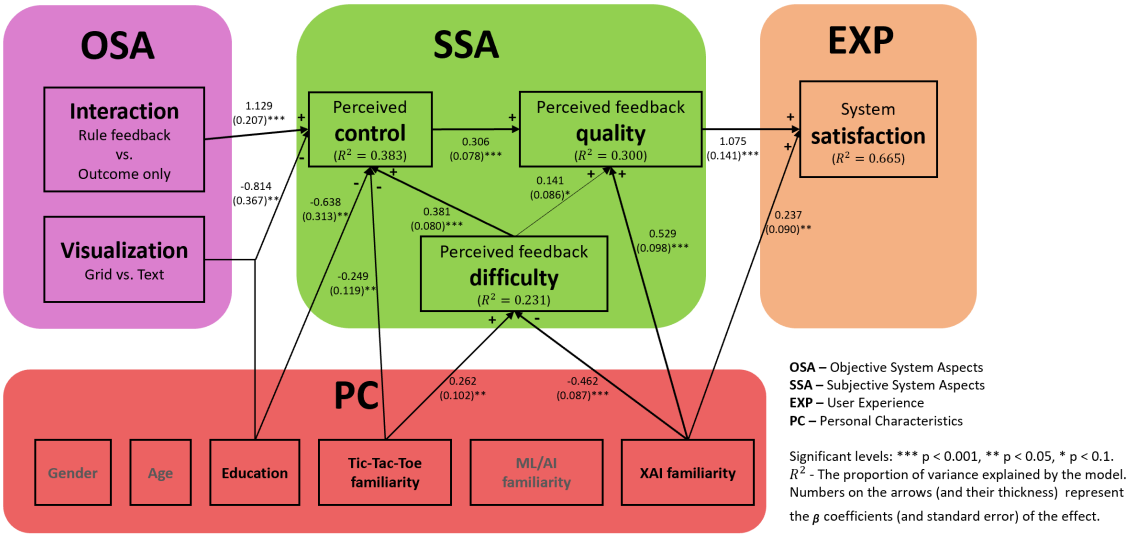

Figure 3.6: The structural equation model for the data of the experiment.

3.4.3 Subjective Experience

As shown in Figure 3.6, the manipulation "interaction" has a significant effect on the perceived control: Participants who were able to edit the rules scored 1.129 standard deviations higher on perceived control than participants who were only able to give feedback on the outcomes—a large effect. The manipulation "visualization" does not have a significant main effect on the subjective system aspects. However, there is a significant interaction effect of visualization and education level (a personal characteristic variable) on the perceived control: participants whose education was limited to high school perceived less control during their interaction with the system if they were shown the grid-based explanation style (see Figure 3.7).

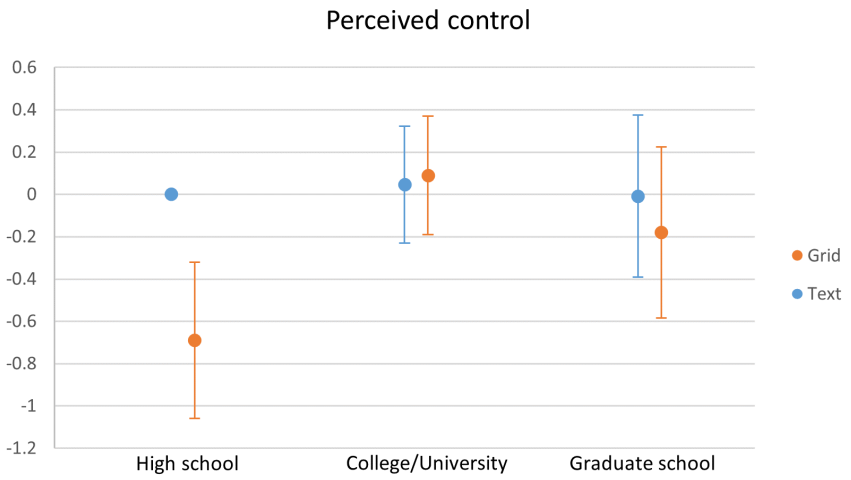

Figure 3.7: Marginal effects of visualization and education on the perceived control, the effect of the "text" condition at "high school" education level is set to zero, and the y-axis is scaled by the sample standard error.

Participants' perception of control is significantly related to their perception of the quality of the feedback given by the system. Moreover, participants' perception of the difficulty of giving feedback to the system also has a significant positive effect on both the perceived control and the perceived feedback quality. Finally, participants' perception of the feedback quality is significantly related to their satisfaction with the system. Feedback quality fully mediates all other system-related effects on satisfaction.

3.4.4 Personal Characteristics

In terms of personal characteristics, participants' age, gender, and their familiarity with ML/AI do not have any significant influences on the other variables in the model, nor do they interact with the experimental manipulations to mediate any of their effects.

Aside from the aforementioned interaction effect, participants' education level does have a significant main effect on their perceived control: participants with a high school education feel less in control compared to participants at other education levels.

Participants' familiarity with the Tic-Tac-Toe game has a positive effect on participants' perception of feedback difficulty and a negative effect on their perception of control over the system, suggesting that participants who are more familiar with the Tic-Tac-Toe game find it more difficult to give feedback and feel less in control during their interaction with the system. Finally, participants' familiarity with XAI has a negative effect on perceived feedback difficulty but a positive effect on both perceived feedback quality and system satisfaction, showing that experienced XAI participants have an easier time giving and receiving feedback, and that they have a higher overall satisfaction.

3.5 Discussion

Based on the results of this online user study, I can describe in detail how the benefits of interaction and visualization in explanatory interactive machine learning come about. I can also describe these results in the light of participants' personal characteristics.

3.5.1 Interaction and Visualization

Both the "interaction" and "visualization" manipulations influence the user's experience with the system, primarily due to their effects on users' perception of control over the system. Allowing users to give feedback on the rules increases their perception of control, which in turn causes them to perceive higher quality feedback from the system, thereby increasing their satisfaction with the system. Perceived control increases feedback quality arguably because the feedback of the system is *in response to participants' input*: the more effective this input is judged to be (perceived control), the more effective the feedback is judged to be as well (perceived feedback quality). The quality of this feedback loop is essential for the successful of systems that employ "human-ML teaming" [165], and my results thus suggest that such systems should allow users to give feedback not only on the

output of the ML system, but on the actual rules that govern its behaviors. Arguably, the ability to engage with the system's rules can support the formation of "team cognition" in human-ML teams [51].

The "visualization" manipulation works differently depending on users' personal characteristics. Specifically, high school educated participants who were shown the grid-based explanation of the machine learning rules perceived less control over the system. This is surprising, as the grid-based visualization was meant to improve the ease of understanding the system, especially for users with no training in formal logic. Arguably, though, compared to the more formal textual explanations, the grid-based visualization obscures the workings of the system to these users, to the extent that it might make it more difficult for them to understand how the edits influence the system (users with a higher level of education may, on the other hand, have an intuitive idea of how their feedback to the grid-based rule is translated into machine rules). Again, further research in the area of "team cognition" could investigate how visualizations can more effectively create a shared understanding of the task and the workings of the system at hand [51].

3.5.2 Personal Characteristics

The results of this experiment demonstrate that users' personal characteristics have a significant influence on their experience with this system. Participants with a high school education may feel less in control because they understand that predictions are the outcome of complex interdependencies between multiple rules—this system only allows them to interact with the rule that currently applies, not with the full body of learned rules.

Likewise, people who are familiar with the Tic-Tac-Toe game rate their perceived control lower and the feedback difficulty higher because they understand that the intricacies of the game are impossible to capture by tweaking a single rule. Additionally, they may be more likely to attempt to interpret the game instances according to their own set of rules, rather than trying to engage with (and optimize) the system-provided rules. This implies that it would be beneficial to users if we could distinguish targeted users by their familiarity with the ML model or AI application and serve them the different specialized interfaces accordingly (cf. Knijnenburg and Willemsen [126, 123]), so as to avoid a situation where a unified interactive interface is too complicated for novice users but too redundant to ML experts, which helps on the acceptance of the ML application among different user groups.

Conversely, experienced XAI users may have had an easier time giving and receiving feedback due to their ability to reflect on their past experiences using such systems. Moreover, this system moves beyond existing XAI systems in several ways, which may be the reason for their increased satisfaction with the system. Arguably, then, familiarity with XAI may be an important precondition for a fluent user experience when using XIML systems, and this may raise questions regarding the equitable distribution of the benefits of XIML systems: if only users who are well-versed in the concept of XAI can benefit from such systems, XIML could become the context for a new digital divide between expert and novice users (cf. [82]).

3.5.3 Systems with Direct Control

Norman's foundational theory of Human-Computer Interaction suggests that systems should aim to provide a *direct mapping* between the user interface and the underlying system mechanisms it controls [183]. In ML systems, the underlying mechanisms may be too complex for end-users to understand, hence real-world user-facing ML systems tend to provide a user interface that offers an indirect mapping to the underlying system. In this study, I added two types of control interfaces to a rule-based ML system: one version allowed users to interact with text-based rules, while the other version mapped the rules onto what we thought would be a more intuitive visualization. The visualization turned out not to be as effective as we expected—perhaps because it constitutes a less direct mapping to the underlying system than the text-based version.

The text based version offered more direct control over the ML mechanisms, and once end-users figure out how these mechanisms work, interacting with them ultimately resulted in a higher satisfaction with the system. This outcome suggests that there is a benefit in providing control interfaces with direct mappings to the underlying system, even if these may initially feel less intuitive to the user. Future work could explore optimal means of explaining such direct control interfaces to novice end-users.

3.6 Conclusion, Limitations, and Future Work

The results of this empirical online user experiment show that explainable machine learning (XML) systems (and arguably XAI systems in general) indeed benefit from mechanisms that allow users to interact with the system's internal decision rules. While this example system serves a

relatively simple scenario (i.e., determining the outcome of a Tic-Tac-Toe game), I find that even in this simple scenario explanation-driven interactive machine learning (XIML) systems have a better user experience, partially because they encourage users to engage in a mutual feedback loop that helps improve the system's performance. Specifically, XIML systems that allow users to edit the decision rules (as compared to only give feedback on the decision itself) make users feel more in control over the system, which increases the perceived quality of the system's feedback and, in turn, the overall system satisfaction.

The simplicity and the relatively objective nature of this scenario is a limitation that I carefully considered in setting up the study. The field of interactive ML/AI has repeatedly shown that tackling simple, objective tasks is a good way to study co-interaction mechanisms before implementing more complicated interactive AI systems that optimize more subjective decisions. As a recent example, Colella et al. designed a simplified interactive optimization task—a 1-dimensional function optimization setting—to study how humans interact with an interactive intelligent system. Due to the simplicity of the task, the study actually works well in making the intelligent system's behavior transparent to the user [43]. Future work could investigate how interaction and visualization as presented by Colella et al. and the current study would translate to more complex tasks, and to optimization for more subjective scenarios (e.g., loan application).

Despite the simplicity of the task domain, this study shows that the effects of XIML systems are not universal, but depend on participants' personal characteristics. This is true regardless of this somewhat limited sample (US-based Mechanical Turk participants), and would likely be exacerbated if I had studied a more diverse global audience. I found that the intuitive grid-based visualization did not improve the user experience, and actually made things worse for users with a high school education level. This finding suggests that visualizations developers may deem "intuitive" are perceived as the opposite by end-users without any training in formal logic, and emphasizes the needs for user-centric research (e.g., participatory design, usability studies) in the field of Machine Learning. This evaluation framework can support such studies.

Likewise, I found that participants who were more familiar with the Tic-Tac-Toe game felt less in control over this XIML system and found it more difficult to give feedback. Familiarity with XAI, however, had positive effects, suggesting that XAI may be a useful—if not necessary—stepping stone towards XIML.

The lack of a positive effect of the grid-based visualization may be an artifact of the partic-

ular domain (Tic-Tac-Toe) or of the quality of the visualization. Intuitive visualizations may indeed make a positive difference in the future—the current study is an initial cautionary step in the investigation of the effect of such visualizations in the context of a Tic-Tac-Toe XIML system. I propose that future work should iterate on these ideas, and/or implement them in a more complex XIML domain that may require a more complex visual design. I do caution future researchers not to take the benefit of their visualization solutions for granted, but to instead carefully evaluate how participants perceive the difference such visualizations make from the perspective of user experience.

3.7 Summary

While the Tic-Tac-Toe game example system serves a relatively simple scenario (i.e., determining the outcome of a Tic-Tac-Toe game), I find that even in this simple scenario, explanation-driven interactive machine learning (XIML) systems create a better user experience, partially because they encourage users to engage in a mutual feedback loop that helps improve the system's performance (Overall RQ1).

The findings of this study motivate me to take a step further to apply the combination of interactivity and explanation in recommender systems that serve a more complex and subjective scenario (Overall RQ3) compared to the rule-based Tic-Tac-Toe game.

Chapter 4

Study II: The Effect of Recommendation Source and Justification on Professional Development Recommendations for High School Teachers

(Note: This work has been published in the 33rd ACM Conference on Hypertext and Social Media (HT '22) [93].)

In this work I conducted a scenario-based study to understand the effect of justification method (needs-based vs. interest-based justification) and recommendation source (a human expert vs. an AI algorithm) on professional development recommendations. In Chapter 2 I have clarified the distinction between *explanations* and *justifications* of recommendations. The effect of explanations together with interactivity on Machining Learning systems has been studied in Chapter 3; in this chapter, I will dig into the effect of justification method (Overall RQ2) together with recommendation source (Overall RQ3) in recommending personalized professional pathways in a scenario-based online user study. This study implies how we should consider the recommendation source design when

designing a recommender systems for different purpose.

4.1 Introduction

The continued improvement of teaching professionals throughout their careers is an essential consideration for modern school systems. Past efforts have often targeted generic professional development methods that only aim to reach required standards, rather than try to fill the specific professional gaps of each individual teacher [46]. Targeting those specific gaps ultimately provides teachers with more effective professional development, which in turn increases the quality of education for the students they teach [32]. While *personalized professional development* has been implemented on a smaller scale, it has been found difficult to expand to wider audiences due to its reliance on experts collecting and evaluating user data to create personalized education plans [153]. Recommender systems can provide a means of expanding these expert decisions to wider audiences in an efficient manner by rapidly considering user data and recommending those professional development resources that are most beneficial and relevant to each individual teacher [200].

While the fields of entertainment and e-commerce were the first to adopt recommender systems in a commercial setting [212, 18, 249, 211], recommenders have more recently found their way to professional settings as well. Our project considers a system that recommends personalized professional development pathways to high school teachers seeking to increase their disciplinary knowledge and/or their teaching skills. The recommendations provided serve as promoted suggestions when signing up for professional development activities with the goal of guiding teachers towards development opportunities they will both enjoy and highly benefit from.

The first iteration of our recommender system, while simplistic compared to the state-of-the-art in this area, uses real-world teacher data to provide recommendations that benefit their professional development: teachers indicate their interests and needs by filling out a "needs assessment" questionnaire, which gets processed by a rule-based system that assigns weights to various professional development options (ranging from single-course microcredentials, to multi-course endorsements and comprehensive master programs) based on teachers' answers to the needs assessment questions, subsequently listing the Top 3 options as recommendations. While these recommendations originate from our system, a lot of work by human experts has gone into the development of the "algorithm"—which is essentially a formalization of a vast body of expert knowledge about how

teachers' needs and interests would translate into professional development options that meet these needs and interests.

In addition to the content of the recommendations provided, our system must also consider the *presentation* of the recommendations: the characteristics of the interface that presents the recommendations to the teachers are critical for ensuring that they carefully consider the recommendations as valid and useful professional development options [26]. In this paper, we investigate two important design considerations regarding the way recommendations are presented to the end-users. Firstly, we acknowledge that teachers' professional development decisions are driven by their interests (What courses would I like to take?) and their needs (What courses would be most beneficial for me to take?), and these may not always be perfectly aligned. The trade-off between these two possible considerations, especially when justifying the recommendations to the user, provides an important avenue for improving the perception of the recommendations provided. Hence, we posit the following research question:

> **Study II - RQ1:** Do teachers have more favorable perceptions of professional development recommendations that are presented as items they would **like**, or as items that would be **most beneficial** to them?

Secondly, given that the recommendations originate from a system that embeds a vast amount of human expert knowledge, we ask ourselves whether it would be better to present the recommendations as originating from an AI algorithm or from a human expert. Past research has found conflicting results on this topic—some have carefully documented cases of "algorithm aversion", where users tend to prefer to receive recommendations from a human rather than an algorithm [34, 114], while others have found situations where algorithmic suggestions are preferred to human suggestions [89, 146, 215]. Given that in our case, one could argue either way about the source of the recommendations, we posit the following research question:

> **Study II - RQ2:** Do teachers have more favorable perceptions of professional development recommendations that are presented as originating from an **AI algorithm** or from a **(human) expert**?

Finally, recent research has suggested that the "algorithm aversion" phenomenon is task-dependent: it is stronger for subjective tasks and/or hedonic decisions than for objective tasks

and/or utilitarian decisions [35]. In this light, one could argue that presenting the recommendations as something the user would *like* frames the decision as a subjective task and focuses them on the hedonic aspects of the decision, while presenting the recommendations as something that would *benefit* the user frames the decision as a more objective task and focuses them on the utilitarian aspects of the decision. This could result in an interaction effect between the justification and the source of the recommendations:

> **Study II - RQ3:** Does the effect of recommendation source **(AI vs. human)** on teachers' perceptions differ depending on the justification for the recommendations **(interest vs. needs)**?

Our work is the first to reconcile the "algorithm aversion" phenomenon (and its inverse) with research on explainable AI (xAI). In particular, whereas the existing research suggests that "algorithm aversion" depends on the recommendation *domain*, our study investigates the interaction between the recommendation source and the type of *justification* within a single domain. A significant effect in our study has considerable practical implications: for one, it would give researchers the opportunity to overcome algorithm aversion—or its inverse—by justifying the recommendations in a needs- or interest-oriented manner. Conversely, our results would provide guidance for recommender system developers to present the recommendations as stemming from either an AI system or a human, depending on how the recommendations are justified.

We answer the stated research questions in the context of our professional development recommender for high school teachers but argue that our core findings are likely applicable to a broader spectrum personalized professional development scenario and perhaps even to recommender systems in general.

4.2 Study Design

We tested the effect of recommendation source (AI vs. human) and justification type (interest vs. needs) on teachers' perceptions of the system in a scenario-based controlled experiment using a prototype of the real system. Testing users' reactions to AI-based systems with prototypes is a common practice in HCI research [43]. While we plan to eventually test the effects of recommendation source and justification in our field trial where teachers receive real personalized professional

development pathway recommendations, we decided to first run a more tightly controlled experiment to study these effects outside the inherently noisy environment of a field trial. In our controlled experiment, all participants receive the same recommendations after reading a scenario describing the abilities, needs and preferences that ostensibly served as input for these recommendations (more details in Section 4.2.3). By using this scenario-based study setup, we can benefit from not having individual recommendation quality interfere with the effect of the presentation of the recommendations (i.e., recommendation source and justification). This setup also makes it possible to manipulate the justification type, as it allowed us to create a scenario where both interests and needs may conceivably underlie the recommendations. Conversely, in the real world some teachers' recommendations follow only their interests or only their needs, making it impossible to claim otherwise (and thus difficult to manipulate the justification type).

Furthermore, the sample of our study is recruited via Prolific, and is thus different from the teachers in our real-world field trial. We took special care, though, that the participants in this study also identified as teachers, and that the recommendations were accompanied by a realistic scenario for which the presented recommendations would be appropriate. This ascertained that participants would be able interpret the quality of the provided recommendations and (more importantly) the justifications, as their background as teachers would allow them to personally relate to the presented scenario. The remainder of this section outlines the setup of the controlled experiment. We note that our study procedures were approved by our Institutional Review Board (IRB).

4.2.1 Participants

We recruited participants for our experiment on Prolific, an online recruitment platform that has a set of detailed filters that can be used to target a particular sample of participants. Using these filters, we limited participation to adult teachers with a completed undergraduate degree or higher, so that the participants would closely match the users of the actual system under development. 207 participants took part in the study, yielding 190 usable data points after filtering out 17 participants who did not carefully read the presented information. Of the 190 participants, 120 identified as women, 66 as men, 3 as non-binary, and 1 participant preferred no to answer our gender question. The sample includes 13 participants between the ages of 18 and 24, 82 between 25 and 34, 55 between 35 and 44, 24 between 45 and 54, and 16 older than 54. Most participants completed the study in 6 minutes. They each received USD 1.20 for their participation.

4.2.2 Procedure

Participants were shown a welcome page with an introduction to our study and a consent form containing statements of possible risks, discomforts, and incentives. We then presented them with a scenario asking them to imagine being a teacher with certain professional development needs/interests (see below), followed by a reading comprehension check question ("Which of the following is not true based on the scenario?").

Next, participants were shown two screenshots of the proposed system: an information screen and a recommendation screen. The information screen (Figure 4.1) welcomed the participant to the system and explained that the recommendations were provided by either a human expert or AI-based-algorithm. The recommendation screen (Figure 4.2) displayed the recommended professional development pathways, including justifications for the recommendation process, as well as each of the individual recommendations, based on either the interests or the needs of the imagined teacher. This screen was followed by another reading comprehension check question[1] ("Which of the following is not one the recommendations?").

Finally, participants were asked to answer a survey containing 38 questions (see below) measuring their opinions about and user experience with the presented system. The user study procedure is shown in 4.3.

4.2.3 Scenario

To provide enough context for participants to understand the recommendations, we created a scenario asking participants to imagine that they are a teacher with a carefully selected set of professional development needs and interests (Figure 4.4). To make the scenario match the source of the recommendations, it emphasized the teacher's interests for participants in the "interests" conditions, while emphasizing the teacher's needs for participants in the "needs" conditions. Note that this difference is merely one of presentation—the content of the two versions of the scenario remained the same. Also note that the participants all identified as teachers themselves, making the scenario (which was rooted in real-world teacher data) easily relatable.

[1]The two reading comprehension check questions each had 4 options. Participants were allowed to fail each question twice before answering correctly. If they failed the question a third time, they would be redirected to the end of the survey, and we would discard their data.

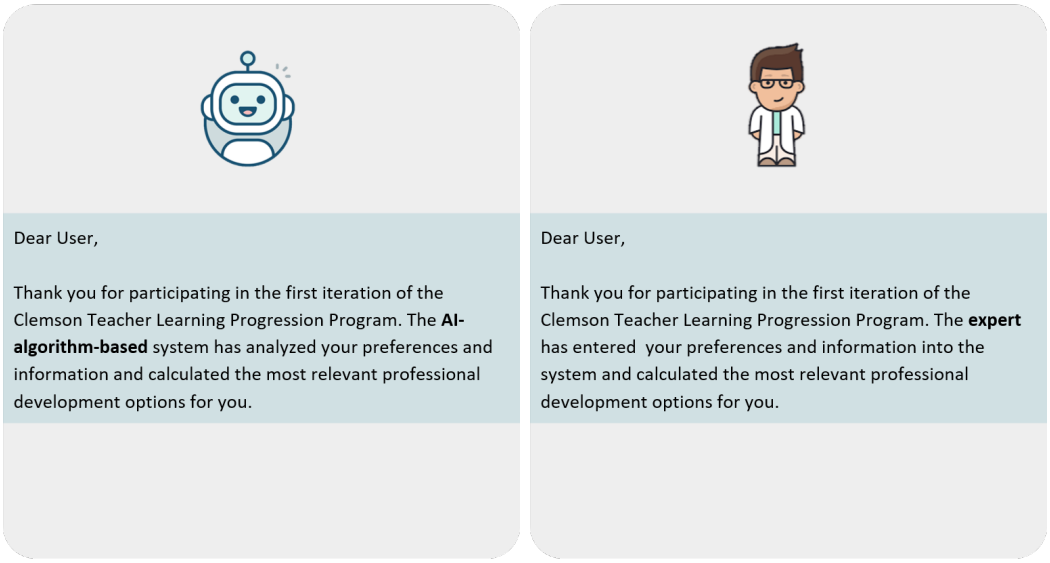

Figure 4.1: The information screen in the two source conditions: AI algorithm (left) and human expert (right).

4.2.4 Experimental Manipulations

The experiment involved two between-subjects manipulations: recommendation source and justification. The recommendation source was presented as either an **AI algorithm** or a **human expert**. The source was printed in bold and accompanied by a robot or human avatar on both the information screen (Figure 4.1) and the recommendation screen (Figure 4.2). The recommendation source manipulation *only* manipulated whether the recommendations were provided by an AI algorithm or a human expert—all other information on the screens was kept as similar as possible, so as to be able to particularly test the effect of the source of the recommendations.

The justifications for the recommendations were presented on the recommendation screen (Figure 4.2) as either the teacher's **interests** ("The [source] thinks you would **like** the following recommendations based on the information you provided.") or their **needs** ("The [source] thinks the following recommendations would be **most beneficial** to you based on the information you provided."). Furthermore, the "reason" listed for each recommendation also reflected the teacher's interests or their needs, depending on the experimental condition. Finally, as mentioned above, the scenario was framed in such a way that these reasons would match the reasons presented in the scenario.

We randomly assigned participants to one of the four experimental conditions in a 2 × 2

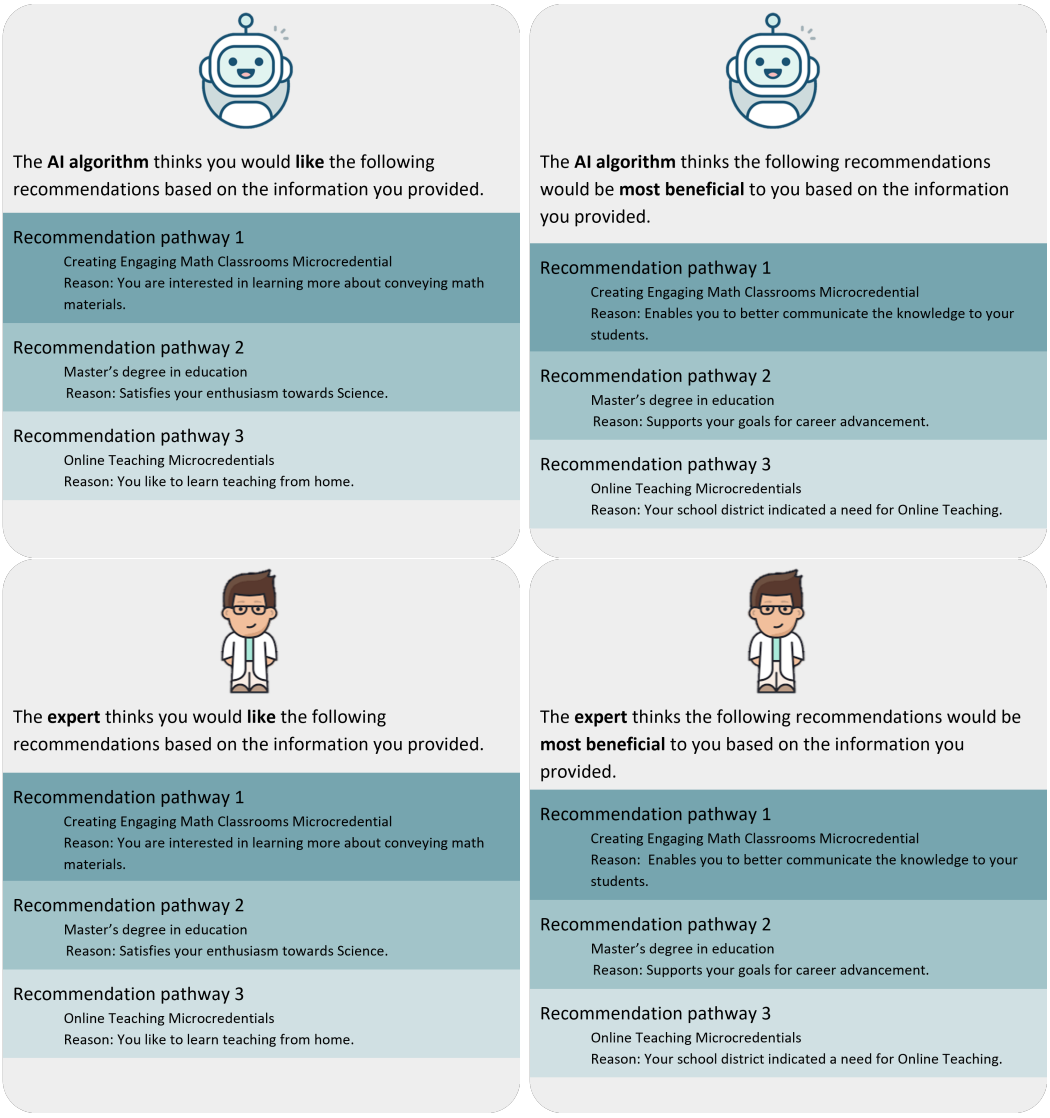

Figure 4.2: The recommendation screen in the four conditions. Top left: recommendations based on interests presented by an AI algorithm; top right: recommendations based on needs presented by an AI algorithm; bottom left: recommendations based on interests presented by a human expert; bottom right: recommendations based on needs presented by a human expert.

between-subjects design—a between-subjects manipulation was used to increase ecological validity and to prevent "demand characteristics" from influencing the study [127].

4.2.5 Dependent Variables

We used the following eight scales (adopted from related work) to measure participants' perceptions of the system attributes and user experience with the system presented in the scenarios

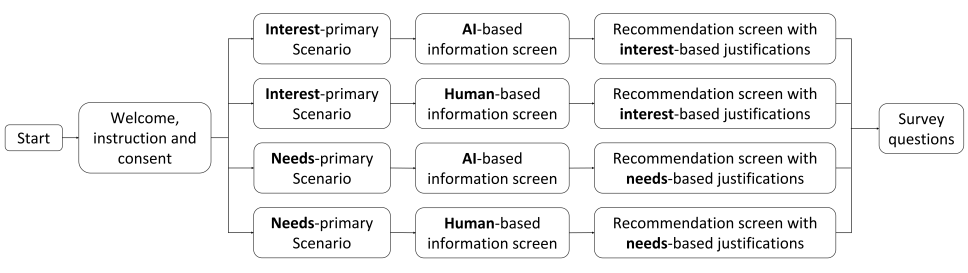

Figure 4.3: The structural equation model for the data of the experiment.

Imagining you are in the following scenario:

You have a bachelor's degree in Math and are currently teaching at the middle school in your neighborhood. However, your district decided to change textbooks this year, and you find the new book somewhat confusing. To keep up with the recent changes, you think that you need more training in Math. Being more knowledgeable about Math makes you feel relaxed in the classes and less worried about students asking questions which you are unable to answer. Furthermore, more training can help you communicate the knowledge to the students better. In addition, you are wondering about moving your background from Math to Science by getting a master's of education degree in teaching over the next few years. This may not only address your enthusiasm for science but also positively impact your salary. Finally, you think teaching online from home and avoiding the commuting hassle is a pretty neat option. You heard from an official that soon, there will be a high need for online teachers.

Now you want to plan for your future.

Imagining you are in the following scenario:

You have a bachelor's degree in Math and are currently teaching at the middle school in your neighborhood. However, your district decided to change textbooks this year, and you find the new book somewhat confusing. To keep up with the recent changes, you think that you need more training in Math. More training can help you communicate the knowledge to the students better. Furthermore, being more knowledgeable about Math makes you feel relaxed in the classes and less worried about students asking questions which you are unable to answer. In addition, you are wondering about moving your background from Math to Science by getting a master's of education degree in teaching over the next few years. This may not only positively impact your salary but also address your enthusiasm for science. Finally, you heard from an official that soon, there will be a high need for online teachers. You think teaching online from home and avoiding the commuting hassle is a pretty neat option.

Now you want to plan for your future.

Figure 4.4: The scenario presented to participants was manipulated alongside the justification manipulation: on the left, interests are presented as the primary source preferences and needs are presented as secondary; on the right, needs are presented as the primary source of preferences and interests are presented as secondary.

with eight subjective measurement constructs:

- `Understandability`: participants' self-reported understanding of the recommendation process, as derived from the justifications (adopted from [122, 120]).
- `Perceived recommendation quality`: participants' perception of the recommendation quality (adopted from [128]).
- `Perceived system effectiveness`: participants' perception of the effectiveness of the system (adopted from [128]).
- `Explainability`: participants' perception how well the provided justifications explained the recommendations.
- `Fit with preference`: the perceived fit of the recommendation with the participant's preferences (adopted from [84]).

- `Competence belief`: participants' perception of the ability, skills, and expertise of the system to perform effectively in its specific domain (a sub-scale of trust, adopted from [248]).
- `Benevolence belief`: participants' perception that the system cares about the consumer and acts in the consumer's interest (a sub-scale of trust, adopted from [248]).
- `Integrity belief`: participants' perception that the system adheres to a set of principles (e.g., honesty and keeping promises) that are generally accepted by consumers (a sub-scale of trust, adopted from [248]).

Each scale consists of multiple items, and a total of 38 items were administered in the questionnaire. Participants were asked to rate each item on a 5-point agreement scale (from strongly disagree to strongly agree).

4.3 Results

Confirmatory Factor Analysis (CFA) was performed to validate the subjective scales that serve as dependent variables in our experiment. We subsequently fitted a Structural Equation Model (SEM) that demonstrates the causal relationships between the manipulations and the validated subjective constructs, as well as mediation effects.

4.3.1 Measurement Model (CFA)

Our CFA indicated 4 questionnaire items with either low loadings (< 0.70) or high modification indices (both of which indicate misfit). These items were removed from subsequent analyses. While all factors had an adequate convergent validity (AVE > 0.50)[2], we found that several factors showed a lack of discriminant validity[3]. Particularly, we found that *explainability* was too highly correlated with *perceived recommendation quality*, *perceived system effectiveness*, *competence belief*, and *benevolence belief*; *Perceived recommendation quality* was too highly correlated with *fit with preference*; and *competence belief* was too highly correlated with *perceived system effectiveness* and *benevolence belief*. To avoid multicollinearity in our subsequent SEM model, we removed *explain-*

[2]Average Variance Extracted (AVE) is an indicator of the convergent validity of the measurement scales, with the recommended lower bound threshold of 0.5.

[3]Discriminant validity is established when the $\sqrt{AVE}$ of a factor is larger than its correlations with each of the other factors

Table 4.1: Items of the 4-factor model. Items without a factor loading were excluded from the analysis.

Considered aspects	Item	Factor loading
Understandability AVE: 0.748 Cronbach's α: 0.91	I understand how the system came up with the recommendations.	0.871
	The recommender explained the reasoning behind the recommendations.	0.812
	I am unsure how the recommendations were generated.	-0.888
	The recommendation process is clear to me.	0.878
	The recommendation process is not transparent.	-0.874
Effectiveness AVE: 0.745 Cronbach's α: 0.92	I would recommend the system to others.	0.954
	The system is useless.	-0.904
	The system makes me more aware of my choice options.	0.768
	I make better choices with the system.	0.884
	I can find better pathways without the help of the system.	-0.789
	I can find better pathways using the recommender system.	
	The system showed useful pathways.	0.904
Fit with preference AVE: 0.885 Cronbach's α: 0.86	The recommended pathways reflect what I want.	0.938
	The recommended pathways suit my needs.	0.964
	The recommended pathways are exactly what I want.	0.919
Integrity AVE: 0.757 Cronbach's α: 0.83	This system provides unbiased pathway recommendations.	0.770
	This system is honest.	0.879
	I consider this system to be of integrity.	0.952

ability, *perceived recommendation quality*, *competence belief*, and *benevolence belief* from subsequent analyses[4].

We again performed a CFA with the remaining four factors (i.e., *understandability*, *fit with preference*, *perceived system effectiveness*, and *integrity belief*). In this analysis, one additional item was dropped from the model due to a low loading (< 0.70). The consistency coefficients (Cronbach's α) of the final four factors showed high to excellent scale reliabilities[5] and the AVE of the four factors ranged from 0.745 to 0.885, indicating that the 4 constructs meet convergent validity requirements (see Table 4.1). The final 4-factor model also meets the discriminant validity requirements (see Table 4.2).

4.3.2 Structural Equation Model (SEM)

A Structural Equation Model (SEM) was fitted to the four constructs and the experimental manipulations (i.e., recommendation source and justification). An SEM enables one to specify

[4]Note that the concept of *explainability* is still represented in the model by its close analog *understandability*, and the concept of *perceived recommendation quality* is represented by *fit with preference*. In addition, while we removed two trust factors (competence and benevolence), one trust factor (integrity) remains in the model.

[5]A commonly used rule of thumb is that an α of 0.7 indicates acceptable reliability, 0.8 or higher indicates good reliability, and 0.9 or higher indicates excellent reliability [209].

Table 4.2: Factor-fit metrics. Off-diagonal values are correlations, diagonal values are the square roots of the average variance extracted $\left(\sqrt{AVE}\right)$ per factor.

	U	E	F	I
Understandability	**0.865**	0.651	0.582	0.609
Effectiveness	0.651	**0.863**	0.780	0.779
Fit with preference	0.582	0.780	**0.941**	0.720
Integrity	0.609	0.779	0.720	**0.870**

the relationships between exogenous variables (the manipulations) and latent constructs (the CFA factors) as a structured model of regressions [127]. An important benefit of SEM is that fit statistics are provided for the model as a whole, as well as for the individual regression coefficients. We built our model following two principles:

1. Justification (RQ1), recommendation source (RQ2) and their interaction (RQ3) were hypothesized to influence *understandability*, *fit with preference*, *integrity*, and *system effectiveness*.

2. Each effect outlined in (1) is allowed to mediate the subsequent effects (e.g., *understandability* is allowed to mediate the effect of justification and/or recommendation source on *fit with preference*).

We first specified a saturated model with all hypothesized effects and mediations. We then iteratively trimmed non-significant effects. The resulting model is displayed in Figure 4.5. This model has a good overall fit with $\chi^2(158) = 246.918$, $p < 0.001$, CFI = 0.990, TLI = 0.991, RMSEA = 0.055 with a 90% confidence interval of [0.041, 0.067][6].

The model shows that the recommendation source and justification manipulations have significant interaction effects on the dependent variables. Particularly, the manipulations have a significant interaction effect on the understandability of the system ($p = .036$) and a marginally significant effect on participants' perceived system effectiveness ($p = .098$); understandability mediates the interaction effects on the perceived fit of the recommendations with the teacher's presented preferences and on participants' perception of the integrity of the system.

Figure 4.6 displays the total (mediated + direct) effects of the two manipulations on the dependent variables. The total interaction effect between source and justification is significant for

[6]Theoretically, a good model is not statistically different from the fully specified model (i.e., the p-value of the χ^2 should be > 0.05), but this statistic is commonly regarded as too sensitive [23]. As such, Hu and Bentler proposed cut-off values for the alternative fit indices to be: CFI > 0.96, TLI > 0.95, and RMSEA < 0.05, with the upper bound of its 90% CI falling below 0.10 based on extensive simulations [106].

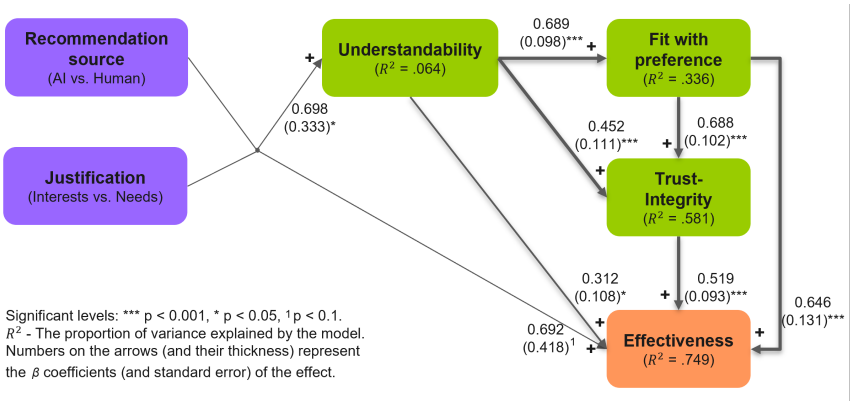

Figure 4.5: The structural equation model for the data of the experiment.

all dependent variables—understandability ($p = .036$), fit with preference ($p = .036$), integrity ($p = .040$), and effectiveness ($p = .011$). In particular, Figure 4.6 shows that when participants were told that the source of the recommendation is a human expert, there was no significant difference in *understandability*, *fit with preference*, *integrity*, or *effectiveness* between participants who were told that the recommendations were based on the teacher's interests vs. the teacher's needs. However, among participants who were told that the source of the recommendation is an AI algorithm, those who were told that the recommendations are based on the teacher's interests perceived a significantly higher level of *understandability* (a large, significant effect; Cohen's $d = 0.78$, $p < .001$), *fit with preference* (a medium-sized, significant effect; Cohen's $d = 0.56$, $p < .001$), *integrity* (a large, significant effect; Cohen's $d = 0.74$, $p < .001$), and *effectiveness* (a large, significant effect; Cohen's $d = 1.12$, $p < .01$) than participants who were told that the recommendations are based on the teacher's needs.

4.4 Discussion

4.4.1 Revisiting the Research Questions, and Comparison to Related Work

We set out to test the effects of justification (interests vs. needs, Study II - RQ1), recommendation source (human expert vs. AI algorithm, Study II - RQ2), and their interaction (Study II - RQ3) on teachers' perceptions of and experience with a personalized professional development pathway recommender in a 2×2 between-subjects controlled experiment. In light of our research

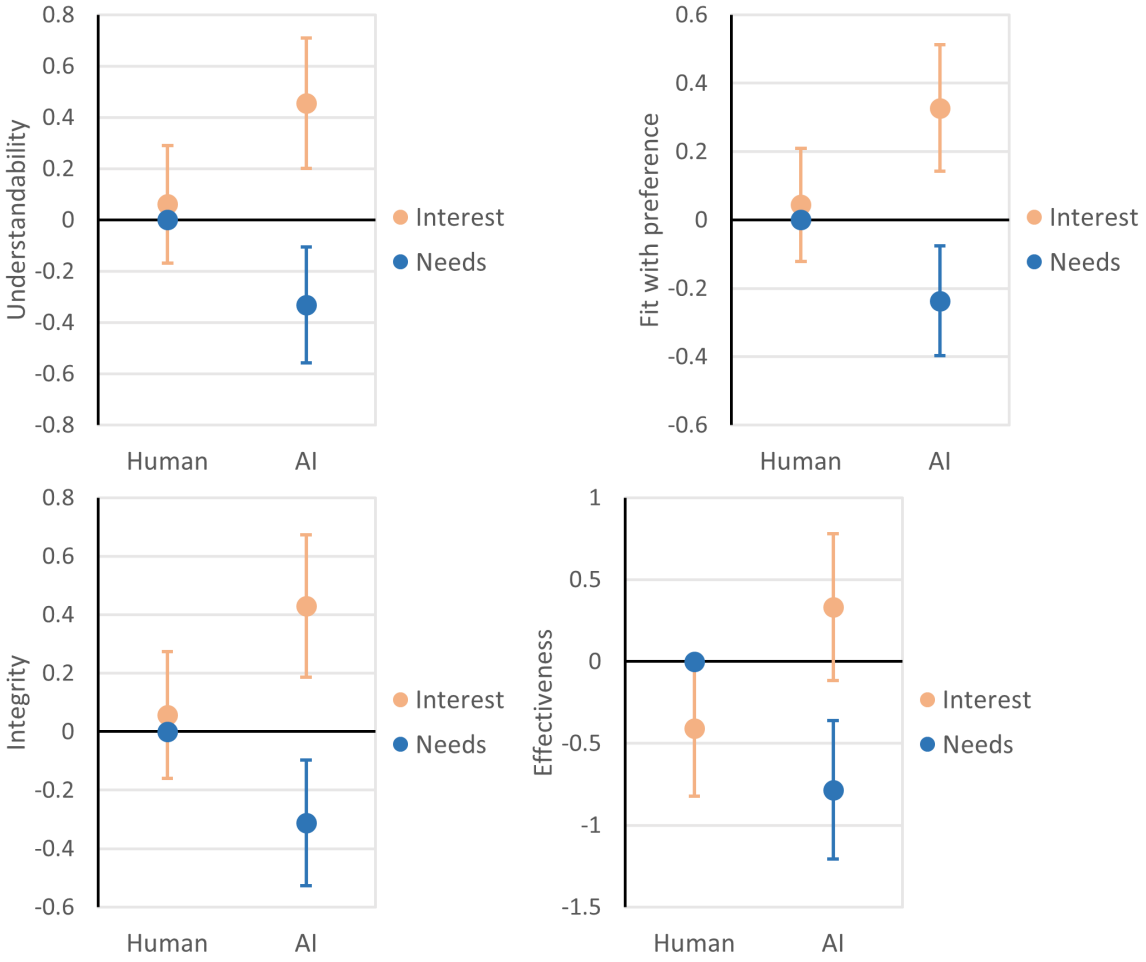

Figure 4.6: Total effects of recommendation source and justification on the perceived understandability (left) and the system effectiveness (right). The effect of the "Human" source with the "Needs" based justification condition is set to zero, and the y-axis is scaled by the sample standard error.

questions, the results show that an interest-based justification outperforms a needs-based justification (Study II - RQ1), but only for users who are told that the recommendations originate from an AI algorithm rather than a human expert (Study II - RQ3). Conversely, the effect of presenting a human expert vs. an AI algorithm as the source of the recommendations (Study II - RQ2) completely depends on the presented justification (Study II - RQ3): users who are told that the recommendations are based on their interests have a better experience when the recommendations are presented as originating from an AI algorithm, while users who are told that the recommendations are based on their needs have a better experience when the recommendations are presented as originating from a human expert.

Notably, the uncovered interaction effect runs counter to existing research, which shows that the "algorithm aversion" phenomenon is stronger for subjective tasks and/or hedonic decisions than for objective tasks and/or utilitarian decisions [35]—assuming that interest-based justifications align with a perception of the recommendations as subjective/hedonic while needs-based justifications align with a perception of the recommendations as objective/utilitarian, our results show that subjective/hedonic recommendations actually perform *better* when presented by an AI algorithm, while objective/utilitarian recommendations perform *worse*. Perhaps, then, there is no clear connection between interest vs. needs-based justifications and subjective vs. objective tasks. Research has shown that lay people perceive a task that can be approached by measuring and analyzing relevant quantitative variables as objective, while perceiving a task that can be approached using intuition or gut feelings as subjective [109]. From this perspective, both types of justifications can be considered objective, as each version portrays the system as taking a decidedly calculative approach to the recommendation process.

A possible alternative explanation for our findings is that while both interest and needs-based justifications are considered objective (and hence principally better suited for an AI system), users find a needs-based justifications condescending when coming from an AI system—if so, needs-based recommendations would indeed best be presented as originating from a human expert. Alternatively, one could argue that needs-based recommendations for teachers' professional development (i.e., recommendations that may have a serious impact on their career) are more consequential than interest-based recommendations (i.e., recommendations that simply align with what they enjoy)—if so, teachers may be less willing to trust in algorithms for (high-risk) needs-based recommendations [193] than for (low-risk) interest-based recommendations [146].

4.4.2 Design Implications

Overall, users are most satisfied with interest-based recommendations presented by an AI algorithm—according to the results in Figure 4.6, users find this system the most understandable, they find that the recommendations better fit their preferences, they find that the system has a higher level of integrity, and they find the system more effective. Arguably, users are most excited about interest-based recommendations, but they may believe that only an AI algorithm would be able to handle the complexity of translating their nuanced interests into a series of recommended professional development activities.

Given these results, we suggest that, when both are possible, the presentation of recommendations should emphasize their algorithmic nature, and the justification of recommendations should relate back to users' interests over their needs. In our study, this presentation was implemented both visually and textually, with the system displaying a robot-like icon and the text explicitly stating that "the **AI algorithm** thinks you would **like** the following recommendations", and the justification was implemented with each individual recommendation having a reference to the teacher's interests (see Figure 4.2). Furthermore, while the design elements chosen were not designed to be overly distracting, the medium to large effect sizes of our study would indicate that more subtle designs implementations could still be effective.

Conversely, an AI algorithm based presentation of recommendations justified by users' needs performed the worst. As mentioned above, one possible explanation for this could be that users find a needs-based explanation condescending when presented by an AI system; another possible explanation is that needs-based explanations are too consequential to trust to an AI system. Regardless, needs-based recommendations are best presented as originating from a human expert.

This finding has implications for situations where recommendations are exclusively based on needs—e.g., in situations where interests are not elicited, or where the recommender system is built to prioritize needs in cases where a system prioritizes targeting a user's deficiencies or needs over their interests. In these cases, it would be disingenuous to justify the recommendations by referring to the user's interests. Instead, when justifying recommendations with a user's needs, the system's presentation should *downplay* the algorithmic nature of the recommendation selection process. In our study, we did this by displaying a human-like icon on the recommendation page, and by explicitly mentioning that "the **expert** thinks the following recommendations would be **most beneficial** to you" (see Figure 4.2). Note that one does not have to completely hide the involvement of a recommender system—on the introduction page of the human source condition, we did explicitly mention that "the expert has entered your preferences and information into the system and calculated the most relevant professional development options for you." (see Figure 4.1). Rather the presentation of these recommendations (i.e., the delivery of them to the user) should be perceived as coming from a human expert rather than an algorithmic system.

4.4.3 Limitations and Future Work

An obvious limitation of this work is that in an effort to control the quality of the recommendations and the justification between subjects, the manipulations were introduced in a scenario-based experiment rather than a real recommender system. To some extent, this limits participants' deeper understanding of the personal relevance of the presented recommendations, and the justifications alike. To mitigate this limitation, we carefully outlined a scenario explaining the fictitious teacher's needs and interests. Moreover, we made sure to recruit participant among actual teachers, who are arguably more qualified to understand the presented scenario and recommendations than the general population.

In our future work, we will confirm these effects in the real personalized professional development pathway recommender and attempt to verify these potential reasons for the uncovered effects. The deployment of this study in the "live" system will also give us the opportunity to test the effects on users' choice behavior: do the recommendation source and justification type have an effect on which and how many professional development items they agree to enroll in? And are there perhaps differences between conditions in terms of users' attrition rates (e.g. dropping classes or abandoning them mid-semester—something that tends to happen frequently, as teachers have to balance their professional development commitments with the demands of their teaching job and their personal lives)? While our current study focused on opinions, the upcoming study with the real system will provide a unique opportunity to carefully study these behavioral effects as well. That said, experimental control will be more difficult in the "live" system, since the teachers will approach the system with different goals, constraints, and ambitions. Hence, the current, more carefully controlled study provides valuable insights into the attitudinal effects of recommendation source and justification type—the "live" system study will complement these results.

Another limitation of our study is that in order to carefully single out the effect of the recommendation source (i.e., to not ascribe differing capabilities to either source), the recommendations are presented as the outcome of a system supporting the recommendation calculation process, even in the "human expert" condition. This may have given our scenario a more calculative emphasis, regardless of the recommendation source or the justification type. To further emphasize the difference between human and AI recommendations, future work could present the recommendations in the "human expert" condition as manually curated rather than calculated. One must however be

careful about the ethical ramifications of misrepresenting the true source of the recommendations.

4.5 Conclusion

In this study, we presented a study to investigate the best way to present recommendations to teachers seeking to advance their professional development. In a carefully controlled, scenario-driven experiment, we tested the effect of the justification behind the recommendations (i.e., the teachers' personal interests vs. their needs) and the source of the recommendations (i.e., a human expert vs. an AI algorithm). The results show that this recommender system benefits teachers most if they are told that the recommendations originate from an AI algorithm and are based on their interests. In our future work, we will confirm the uncovered interaction effect in the real recommender system and attempt to verify its underlying cause.

4.6 Summary

The findings of this study suggest that the presentation of recommendations should emphasise their algorithmic nature, and the justification of recommendations should relate back to users' interests over their needs (Overall RQ2). This implies that while building the movie recommender systems, it would provide a better user experience if the movie recommendations are presented as originating from *algorithms* rather than from a *human expert* considering that movie preferences (which will be visualized by the movies' emotion feature) are usually based on users' interest (Overall RQ3).

Chapter 5

Study III: Preference Exploration and Development: The Role of Individual Differences

(Note: Part of his work (the algorithms) has been published in the Doctoral Symposium of the 12th ACM conference on recommender systems (RecSys '22).)

In this work, I developed four alternative algorithms that go beyond the traditional top-N recommendations and build them into a recommender system that recommends items for self-actualization. Instead of only focusing on the traditional top-N recommendations which suggest items the system thinks users will like, we also provide transparency and control with the system. Specifically, transparency was implemented by developing algorithms to dig into users' expressed preference and discover new recommendations from the following four different perspective:

1. Items users might hate. This shows recommendations with a low predicted rating, allowing users to confirm or correct the potential "false negatives";

2. Items they might be among the first to try. This suggests items that haven't been rated by many users yet, recommending new items to try out for users who are open to novelty;

3. Items the system has no clue about. This composes of items for which the system has the lowest confidences in the personalized predicted rating for the current user, helping users to

discover their unexplored potential preferences;

4. Items that are controversial. This shows recommendations which are polarizing among like-minded users, allowing the user to explore tastes that go beyond the mainstream.

With regard to the control element, I designed two rounds of recommendations in the experiment, users are allowed to rate the initial recommendations to refine the estimates from the recommender. This rating process enables users to interact with the system by indirectly correct their preference or confirm the systems' estimates of the alternative recommendations.

I ran an online user study to test if this system creates a better user experience through transparency (Overall RQ1) and interactivity (Overall RQ2) specifically by improving self-actualization (motivates the alternative recommendation lists), which help overcome the "filter bubble" problem.

5.1 Introduction

Recommender Systems that help users handle the abundance of information and choices are available on today's websites (e.g., e-commerce, streaming). These systems filter the catalog of products, suggesting possible relevant items to a user based on e.g., previous behavior or explicitly stated preferences [201].

Past research has focused on making these recommendations as accurate as possible, thereby inevitably ignoring the items that the system thinks the user will not like. Although this approach has been shown to improve the user experience [129], some scholars have argued that this creates a "Filter Bubble" that traps users in their comfort zone [187, 176, 198]. This filter bubble prevents them from discovering new and unknown areas of their own taste, and limits the diversity of presented content.

In previous work Knijnenburg et al. argued that the "filter bubble" problem can be overcome by building "Recommender Systems for Self-Actualization" [125, 253]. These systems concentrate on the more complex situation of helping users in developing their preferences rather than only suggesting accurate items [125, 253]. They focus on exploring previously unknown taste areas instead of enforcing already known preferences. Recommender Systems for Self-Actualization (RSSA) keep the user "in-the-loop" by providing alternative recommendation lists that go beyond the traditional Top-N list which purely concentrates on the algorithm accuracy.

More and more scholars have noticed that accuracy is not the only criterion for evaluating the effectiveness of recommender systems [223, 96, 132], the shift of the focus of recommendation systems from algorithm accuracy to user experience has been the interest of many researchers in recent years. Many scholars' studies have shown that user experience accounts more for a larger proportion of system satisfaction [155, 96, 132], high accuracy metrics may even hurt recommender systems [163]. Thus the evolution from research concentrated purely on accuracy to solutions that improve the user interaction experience with the recommender becomes necessary, as McNee et al. have suggested, recommender systems need a deeper understanding of users and their preference seeking task to improve the recommendation quality through investigating the interactions between human and the recommenders [164].

Although prior work have acknowledged the importance of going beyond accuracy in order to achieve the effectiveness of recommender systems, these work did not develop a real recommender system that really focuses on improve the effectiveness of the system through user experience, for instance, McNee et al. proposed a Human-Recommender Interaction (HRI) framework to understand users, their tasks, and recommender algorithms using a common language [164]; He et al. argued for the importance of other user-centered factors beyond accuracy through presenting an interactive visualization framework that combines recommendation with visualization techniques to support human-recommender interaction [96]. In this work, we fulfilled the concept of Recommender Systems for Self-Actualization (RSSA) in a real recommender system through providing four alternative recommendation lists that go beyond accuracy to help users examine and understand their own tastes and preferences.

With our four RSSA features (i.e., the four novel alternative recommendation lists in Recommender Systems for Self-Actualization), we aim to achieve this goals: a) supporting rather than replace decision making so as to help users develop and express their preferences; b) focusing on exploration rather than consumption (in other words, the four RSSA features do not focus on optimizing the probability that the user will like recommendations, but instead focus on exploring underdeveloped recommendations); c) attempting to cover users' tastes. Because users' preferences are not singular, but rather multifaceted and loosely connected, and an ideal recommender should be able to fit any part of a users' preferences.

The effectiveness of a recommender system is likely to be affected by users individual differences [112]. Psychologically, personality accounts for the individual differences in users' preferences

and behaviour [234]. Past research have indicated that personality quizzes can be a viable and promising way to build user profiles to recommend entertainment products [107]. For instance, domain knowledge has an influence on users' perception of the recommender system [108]. The personality factors for an observed user can be explicitly acquired through questionnaires [234]. Hence, other than testing the effect of the four novel RSSA features on users experience, we were also trying to explore the role of individual difference acquired from the personality questionnaire while users interacting with the system.

In this work, we designed and developed a novel recommender system providing the four RSSA features and conducted an online user study to subjectively evaluate of the effect of the novel RSSA features and the role of individual difference in this system.

5.2 Algorithms

In this section, we first describe the data that was used in our experiment. Then, we explain our approach employed to develop the algorithms used in the system.

5.2.1 Data

The MovieLens 25M dataset[1] was used for this study. This dataset contains around 25 million 1 to 5 star ratings on 62000 movies by 162000 users. Users were selected at random for inclusion, all selected users had rated at least 20 movies. We enrich the dataset with additional movie information (e.g. synopsis, cast, genre, poster, etc) that were extracted from IMDB database site[2] with the open-source code from GitHub[3,4].

In our recommender system, the observed data will be managed as a user-item matrix (Figure 5.1, i.e., matrix with user IDs as row indices, item IDs as column indices and the user's item rating as the value for each cell. In Figure 5.1, the question mark represents the fact that an item has not yet been rated by the user.

[1]The dataset was accessible at https://grouplens.org/datasets/movielens/
[2]https://datasets.imdbws.com/
[3]https://github.com/alberanid/imdbpy
[4]https://github.com/babu-thomas/movielens-posters

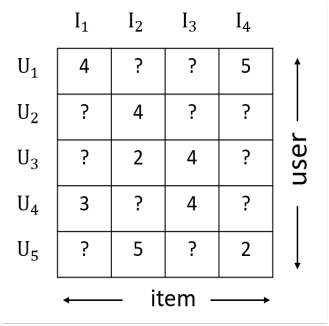

Figure 5.1: User-item matrix.

5.2.2 Recommender System Approaches

Our algorithms are based on collaborative filtering, an approach that identifies user-item associations by detecting the interdependencies between items and relations between users. Collaborative filtering has been widely used in both commercial applications [143] and academic studies [199, 221]. Among collaborative filtering approaches, the neighborhood-based approach and latent factor models are the two primary methods.

The neighborhood-based approach can be further classified into item-based and user-based approaches. Item-based collaborative filtering exploits the relationship between items to identify "neighboring items" to those items that are known to match the user's preferences, neighboring items are items that have similar ratings when rated by the same user. User-based collaborative filtering works similarly, but exploits the relationship between users.

Latent factor modeling, on the other hand, is a technique that characterizes both items and users by vectors of "feature factors" inferred from item rating patterns through dimensional reduction. Recommended items have a vector that shows a high correspondence with the current user's vector [133]. Matrix factorization is the most common realization of latent factor models, which is widely used in recommender systems as well.

5.2.3 Recommendation Lists Algorithms

The first two of our proposed alternative recommendation lists, i.e., "Things we think you will hate" and "Things you will be among the first to try", use the matrix factorization method to get the predictions for unrated movies, but use these ratings in alternative ways rather than simply reporting the Top-N predicted ratings.

We use the SVD (Singular Value Decomposition) technique to get the latent feature factors, reducing the high-dimensional user-item rating matrix R into two lower-dimensional matrices P and Q that describe users and items as vectors in a latent feature space [133, 254]. Then the user-item rating matrix can be approximated by the product of these two lower-dimensional matrices P and Q, which produces the predicted ratings:

$$\hat{R} = P \times Q^t$$

Since we use exactly the same algorithm as specified in [133] to get the predicted ratings, we do not address the realization of the basic matrix factorization algorithm in detail in the current chapter. Instead, we focus on specifying why and how to use the predicted ratings for the unrated movies in the user-item matrix to construct alternative lists of 'Things we think you will hate" and "Things you will be among the first to try".

5.2.3.1 Things we think you will hate

Since recommender systems usually recommend items with high predicted ratings, items with low personalized predicted ratings are rarely ever shown to users. However, it is possible that the recommender systems mislabeled certain items as something the user dislikes, and it would be useful for the user to be able to correct such mistakes. Moreover, exposure to items that the system predicted they will dislike may give the user interesting insights into their tastes and preferences.

Therefore, we propose, as our first alternative list of recommendations, a list of "things we think you will hate". Rather than simply selecting the items that have the lowest predict ratings based on our matrix factorization approximations, we select movies that have a predicted rating that is much lower than the average rating. The reasoning is that these items are generally appreciated (i.e., not universally bad items) but simply do not match the user's tastes. This algorithm is addressed in Table 5.1.

5.2.3.2 Things you will be among the first to try

Most recommender systems have to deal with an influx of new items. As these items initially have no or only a few ratings, they have a relatively low probability to be in any user's Top-N recommendations. For example, if a new movie on Netflix happens to get rated by a few

Table 5.1: Things We Think You Will Hate

Algorithm 1: Last 10 items

Input: Personalized predicted ratings $\{\hat{R}_{ui}\}$ set
Get mean predicted rating $\bar{R}_i$ over all observed ratings
For $u \epsilon U$:
  Sort $\bar{R}_i - \hat{R}_{ui}$ for unrated items in ascending order
  Get the last 10 items

users that do not like its genre, then the initial lower ratings will make it less likely to show up among any Top-N recommendations, even for users who really like that genre. Most recommender systems solve this *cold start problem* through content-based filtering: in the absence of ratings, they resort to item attributes (e.g. genre) to make initial recommendations. However, these cold start solutions ignore the fact that users may actually like to try new items.

Taking this situation into account, we propose a list of "Things you will be among the first to try", which lists items that have personalized predicted ratings in the Top-200 (or some other number larger than N, allowing for some good-but-not-great items) that were rated by the fewest number of other users. We get this list with the algorithm specified in Table 5.2.

Table 5.2: Things You Will Be Among the First to Try

Algorithm 2: Novel items

Input: Personalized predicted ratings $\{\hat{R}_{ui}\}$ set
For $u \epsilon U$:
  Sort $\hat{R}_{ui}$ for unrated items in descending order
  Get the top 200 items
  Get the numbers of observed ratings for those top 200 items
  Get 10 items with lowest number of observed ratings
  Sort them in descending order based on predicted ratings

5.2.3.3 Things we have no clue about

It is difficult for recommender systems to predict the rating of an item when there is not sufficient information about whether the user will like the item or not. For example, if a Netflix user happens to rate only movies in one particular genre, then the system will have a hard time predicting the rating of movies of a different genre, and such movies will rarely get a high enough predicted rating to be featured in the user's Top-N. As such, recommender systems often end up targeting a specific subset of the user's preferences, and struggle to get a more holistic representation

of the user's tastes.

To remedy this situation, we propose to show a list of hard-to-predict items, i.e. "Things we have no clue about", that may be used to identify unexpressed preferences. This list is composed of movies for which the system has the lowest confidence in the personalized predicted rating for the current user. The algorithm producing this list thus has to go beyond a single predicted rating and consider the *confidence* of the prediction as well. Whereas the predicted rating represents the algorithm's best guess on what the user's actual rating will be, the confidence of the corresponding predicted rating indicates how certain the algorithm is that the prediction is in close vicinity to the actual rating. Mazurowski [158] has created an algorithm that can estimate the confidence of individual rating predictions in item-based collaborative filtering recommender systems. We use this algorithm on top of a standard item-based collaborative filtering algorithm (Table 5.3) to generate a list of "Things we have no clue about" (Table 5.4).

Table 5.3: Item-based Collaborative Filtering Algorithm

Algorithm 3: Item-based CF
Input: Observed ratings $\{R_{ui}\}$ set Get general mean of observed ratings μ For $u \epsilon U$: Get mean ratings $\bar{R}_u$ over all rated items Get user bias by $b_u = \bar{R}_u - \mu$ For $i \epsilon I$: Get mean ratings $\bar{R}_i$ over all observed ratings Get item bias by $b_i = \bar{R}_i - \mu$ For $u \epsilon U$ and $i \epsilon I$: Get $b_{ui} = \mu + b_u + b_i$ For $i, j \epsilon I$: $S(i,j) = \frac{\sum_{u \epsilon U}((R_{ui} - b_{ui})(R_{uj} - b_{uj}))}{\sqrt{\sum_{u \epsilon U}(R_{ui} - b_{ui})^2}\sqrt{\sum_{u \epsilon U}(R_{uj} - b_{uj})^2}}$ For $i \epsilon I$: Get nearest neighbors $N_{s(i)}$ set For $u \epsilon U$: $\hat{R}_{ui} = \frac{\sum_{n \epsilon N_{s(i)}}(S(i,n)(R_{un} - b_{un}))}{\sum_{n \epsilon N_{s(i)}}\lvert S(i,n)\rvert}, n = \left\lvert N_{s(i)} \right\rvert$

5.2.3.4 Things that are controversial

As mentioned in Section 5.2.2, recommender systems based on neighborhood-based collaborative filtering provide recommendations based on the preferences of the user's nearest neighbors. These recommendations usually consist of items that the neighbors unanimously like. However, it

Table 5.4: Confidence of Predicted Ratings

Algorithm 4: RESAMPLE
Input: Observed ratings $\{R_{ui}\}$ set
m - number of repetitions
α - portion of R_{ui} to be re-sampled each time
Get N - the number of observed ratings in $\{R_{ui}\}$
Get $n = \alpha \times N$ - sample size
For $j = 1 : m$:
Randomly select n observed ratings $\{R_{ui}^{(j)}\} \subset \{R_{ui}\}$
Execute Item-based CF on $\{R_{ui}^{(j)}\}$ to get $\{\hat{R}_{ui}^{(j)}\}$
Get $C_{ui} = \frac{1}{sd(\hat{R}_{ui}^{(j)})}$, confidence of $\hat{R}_{ui}$
For $u \epsilon U$:
Sort C_{ui} in descending order
Get the last 10 items

is possible that for certain items, these neighbors are divided into two groups with different preferences: some of them love the disputed item, while others hate it. These items usually do not get recommended though, because their predicted rating is an average over all neighbors, resulting in a relatively low score. However, these disputed items may be of special interest to users who want to explore items that go beyond the "mainstream", allowing them to further develop their unique tastes. Moreover, a recommender system can learn much more about a user's preferences by learning their rating on such a contested item rather than an item that is already universally liked among the user's nearest neighbors.

In our final list, we therefore propose to show these disputed items. The algorithm to generate this list of "Things that are controversial" is a variant of the traditional item-based collaborative filtering algorithm, but focuses on finding items that show the largest variance among the user's nearest neighbors (Table 5.5).

5.3 Experimental Setup

In this section, we first offer details into the design and development of the movie recommender. Then, we lay out the procedure that guided the experiment.

Table 5.5: User-based Similarity and Controversial Items

Algorithm 5: Controversial items
Input: $\{R_{ui}\}$ set and $\{\hat{R}_{ui}\}$ set
Get the same μ, b_u, b_i, b_{ui} with Item-based CF
For $u, v \epsilon U$:
$S(u,v) = \frac{\sum_{i \epsilon I}((R_{ui}-b_{ui})(R_{vi}-b_{vi}))}{\sqrt{\sum_{i \epsilon I}(R_{ui}-b_{ui})^2}\sqrt{\sum_{i \epsilon I}(R_{vi}-b_{vi})^2}}$
For $u \epsilon U$:
Get nearest neighbors $N_{s(u)}$ set
For i in unrated items by the user:
Sort $var(\hat{R}_{vi}), v \epsilon N_{s(u)}$ in ascending order
Get the last 10 items

5.3.1 Design Rationale

We developed a new movie recommender system to answer our research questions. Prior studies within the RS community have used the movie domain to explore preference alignment [162, 117, 228, 40, 10] and by extension taste development [31, 17, 208, 16, 141]. Thus, we implement and test our RSSA features alongside traditional Top-N recommendations. As displayed in Figure 5.2, our system had the ability to seamlessly display items from both a traditional Top-N recommender as well as one of our RSSA features (i.e., *Things we think you will hate* (the "hate" items), *Things you will be among the first to try* (the "hipster" items), *Things we have no clue about* (the "no-clue" items), and *Things that are controversial* (the "controversial" items)). As a comparison, we have two baselines: a) we only show a single list of the top-N recommendations (the single list condition); b) we display *More things you may like* (the "next N" items) which presents the additional N movies ranked righted after the top-N recommendations (the *next-N* recommendations). This ended up with one manipulation with six conditions (two baselines plus four RSSA lists). We randomly assigned participants to one of the six experimental conditions in a between-subjects design—a between-subjects manipulation was used to increase ecological validity and to prevent "demand characteristics" from influencing the study [127].

On the recommendation lists, each recommended movie featured the corresponding movie poster and synopsis. Figure 5.2 shows traditional Top-N items on the left while simultaneously providing RSSA items or the baseline items in the list (for example, "Things that are controversial") on the right.

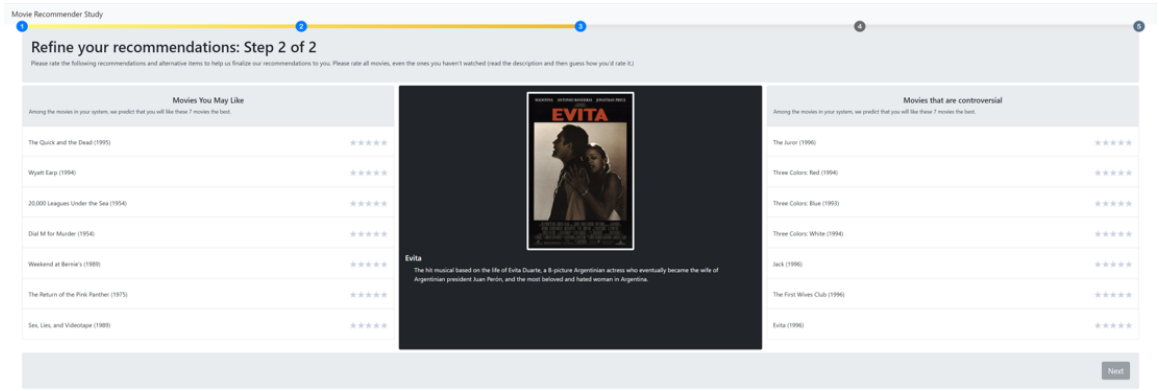

Figure 5.2: The recommendation page of the "Things that are controversial" condition: top-7 list on the left and 7 items of "Things that are controversial" on the right.

5.3.2 Participants

This study was submitted to our Institutional Review Board and considered exempt. We conducted the study on the Prolific platform [5], limiting our recruitment to adult users living in the US. One attention check question ("Regardless of your answer, choose "disagree" to the following") was randomly placed among the 37 user experience questions to track if participants were actually paying attention to what they are doing during the completion of the study. We used these questions plus the time taken to complete the study to filter out participants who clearly rushed through the study.

488 participants took part in the study, yielding 483 usable data points after filtering out 5 participants who did not carefully read the presented information. Of the 483 participants, 223 identified as women, 245 as men, 9 as non-binary, and 6 participant preferred not to answer our gender question. The sample includes 111 participants between the ages of 18 and 24, 143 between 25 and 34, 122 between 35 and 44, 64 between 45 and 54, 42 older than 54, and 1 participant prefers not to disclose. Most participants completed the study in 13 minutes. They each received USD 2.75 for their participation.

5.3.3 Procedure

The procedure, summarized in figure 5.3, contained the following steps:

1. *Introduction and consent:* Upon joining the study, participants were shown a welcome message

[5] https://www.prolific.co/

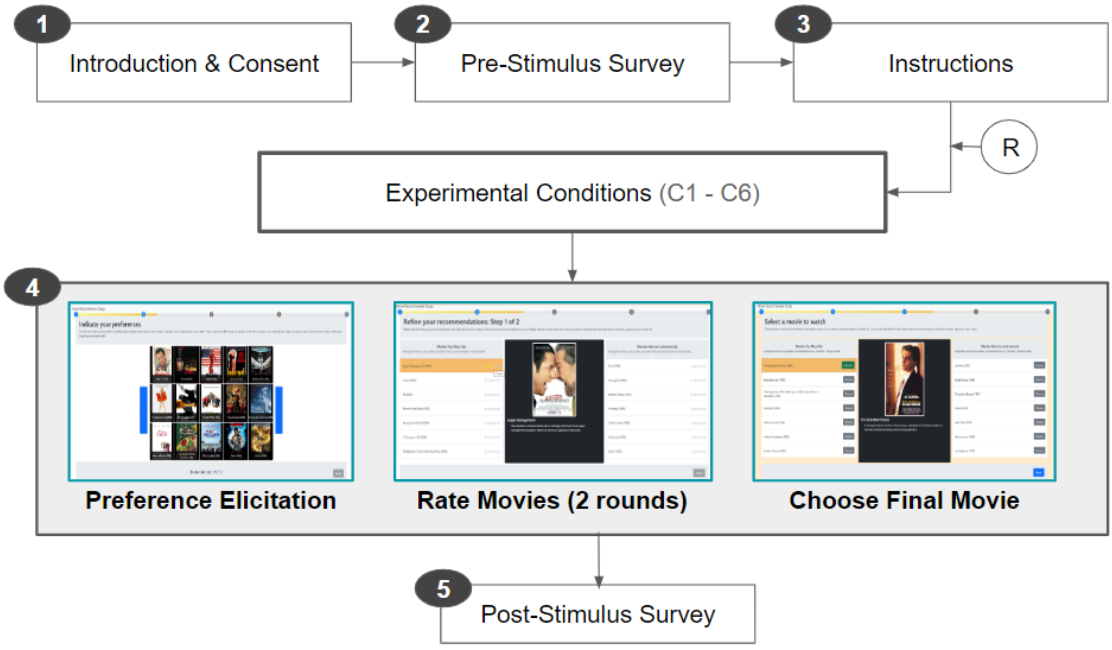

Figure 5.3: The figure above summarizes the key procedural steps of the experiment. "R" denotes random assignment to the experimental conditions.

that provided introductory information about the study to provide insights into expectations. They were then shown the consent information related to the study.

2. *Pre-study questionnaire:* In this section, participants were asked about personal characteristics that could influence users' decision-making when interacting with an online movie recommender such as movie expertise, need for novelty, the extent to which people fear of missing out, and people's perception of their maximizing tendencies. These questions were presented at this stage to avoid confounding effects after interacting with the system. These factors of personal characteristics are described in section 5.3.4.1.

3. *Instructions:* Once the pre-study questionnaire was completed, participants were then offered an overview of all of the steps in the study. Our system also clearly labeled an updated progress bar prominently at the top of the interface to allow users to keep track of the stage in the experiment.

4. *Preference Elicitation:* To provide a personalized experience, we need to first collect a subset of ratings to better understand user preferences. Participants were asked to rate *at least*

ten movies that they have seen before on a scale of five. 24 movies distributed in a gallery of three rows and five columns were displayed for users to rate; users can scroll through the gallery for more movies rate with a next-functioned button. At this step, it is possible that users (especially who are not movie lovers) may not find movies that they have seen on the first few pages; to avoid users clicking through the movies to rate the required ten movies they have seen in this situation, I developed a randomization algorithm to display the movies from our database by popularity, as shown in Table 5.6; for example, on second page of the movie gallery, two random movies were drawn from the top-200 movies (ranked by popularity), seven movies ranking between 201 and 1000, seven movies ranking between 1001 to 2000, three movies ranking between 2001 and 5000, three movies ranking between 5001 and 10000, and 2 movies from the remaining ranking tail.

After rating at least the minimum, they were allowed to proceed to the next page. We added a buffering page before the next page which asked participants to "please hang on while we find the recommendations for you". This was done enhance perceptions of a personalized experience as well as to normalize the wait time between conditions.

5. *Interaction with recommender:* Based on the selections from the preference elicitation stage, the system generated two lists. On the left, the recommendations were from traditional Top-N recommender, whereas, the list on the right included items from one of our six conditions. We iterated the recommendations on three progressive recommendation pages. On the first page, the two lists of recommendations were generated based on the preference elicitation. Participants were asked to rate all the recommended movies, even the ones they have not watched (read the description and guess how they would rate them). On the second page, two lists of recommendations were generated based on the preference elicitation and the ratings they submitted on the previous recommendation page. And then they were again asked to rate all the recommended movies on this page. The system then took the new ratings and fine-tune the final recommendations, and generated the third round of recommendation in the same two lists on the third recommendation page, on this third recommendation page, participants were asked to select one movie that they would like to watch right away if they could.

6. *Post-study questionnaire:* Lastly, participants were asked to evaluate their perception of their interaction with the system. All of the factors that were evaluated are described in section

Table 5.6: The randomization algorithm for the preference elicitation step.

	Page 1	Page 2	Page 3	Page 4	Page 5
Top 200	0	2	4	6	6
200 – 1k	5	7	9	9	9
1k – 2k	5	7	9	9	9
2k – 5k	5	3	1	0	0
5k – 10k	5	3	1	0	0
10k+	4	2	0	0	0

5.3.4.1. Demographic questions were asked towards the end of the study.

5.3.4 Measurements and Hypotheses

5.3.4.1 Measurement scales

As mentioned above, we measure participants' personal characteristics on the following four aspects:

- `Movie expertise`: participants' perception of their movie expertise (adopted from [121]).
- `Need for novelty`: participants' perception of their need for novelty (adopted from [38]).
- `Fear of Missing Out (FOMO)`: the extent to which participants fear of missing out about movies (adopted from [194]).
- `Maximization tendency`: participants' perception of their maximizing tendencies (adopted from [70]).

We used the following six scales (mostly adopted from [128]) to measure participants' perceptions of the subjective system aspects (SSA) and user experience with the system (EXP) in the post-study questionnaire:

- `Diversity`: participants' perception of the diversity of the recommendation (adopted from [254]).
- `Recommendation quality`: participants' perception of the quality of the recommendation (adopted from [121]).
- `Taste coverage`: participants' perception of the extent that the recommender system reflects their tastes.

- `Conformity`: participants' perception of the ability that the recommender system can distinguish between different preferences.
- `Choice satisfaction`: participants' perception of the satisfaction of the recommendation (adopted from [30, 254]).
- `System satisfaction`: participants' perception of the satisfaction of the recommender system (adopted from [121]).

Each scale consists of multiple items, and a total of 61 items were administered in the questionnaire (i.e., 24 items measuring personal characteristics in the pre-study questionnaire and 37 items measuring their subjective opinion and experience with the system in the post-study questionnaire). Within the six scales presented in the post-study survey, *Diversity*, *recommendation quality*, and *Conformity* were measured for the items on the left only. Participants were asked to rate each item on a 5-point agreement scale (from strongly disagree to strongly agree). We also measure participants' interactions (INT) with the system by logging their clicks when they interacted with the system. In this study, we logged their ratings on the first two recommendation pages, the time spent on selecting a movie to watch on the third recommendation page.

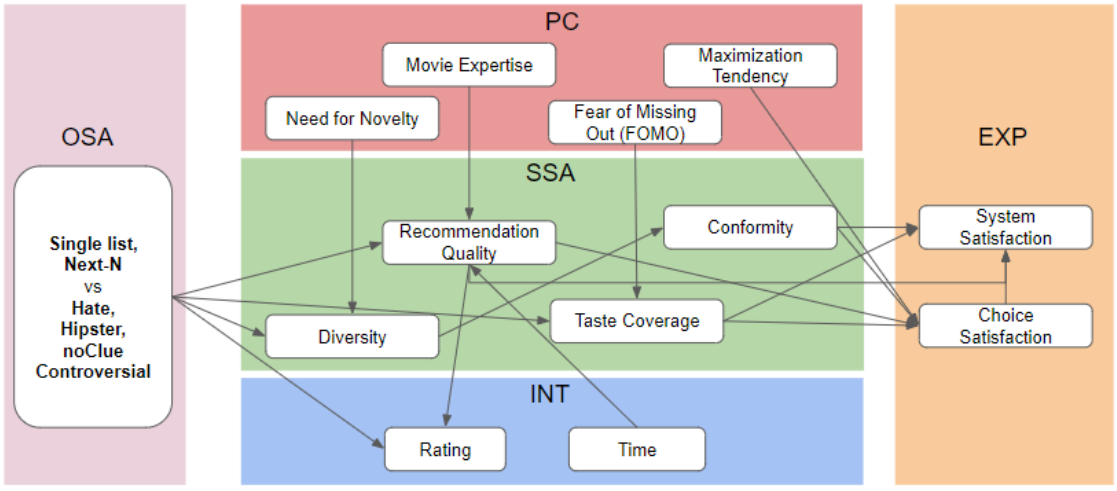

Figure 5.4: The conceptual model of the study based on the user experience framework by Knijnenburg et al. The following are explanations for abbreviated terms above: OSA means objective system aspects, SSA means subjective system aspects, INT means interactive components, PC stands for personal characteristics, and EXP relates to the user experience.

5.3.4.2 Hypotheses

We integrate our experimental manipulation (the RSSA features), personal characteristics, interaction variables, and the subjective constructs in the hypothesized path model (Figure 5.4). We hypothesize that the RSSA features increase the perceived recommendation quality, the perceived diversity, and the perceived taste coverage. Diversity is further hypothesized to increase the conformity of the system; recommendation quality, taste coverage, and conformity are hypothesized to increase users' satisfaction with the provided RSSA recommendations and the recommender system itself (the effect perhaps mediated by the choice satisfaction).

Finally, we hypothesize that there are correlations between need for novelty and perceived diversity, between movie expertise and perceived recommendation quality, between fear of missing out (FOMO) and perceived taste coverage, and between maximization tendency and choice satisfaction. We further hypothesize that there are differences on the recommendation ratings among different RSSA recommendation lists, and this differences may be mediated by the recommendation quality. We also hypothesize that the time spent on selecting one movie on the last recommendation will reflect the recommendation quality of the RSSA features to some extent.

In the results section 5.4, these effects are added to the model where significant in an ad-hoc manner.

5.4 Results

We first validated our measurement model regarding the PC, SSA and EXP constructs using a Confirmatory Factor Analysis (CFA) and then fitted a Structural Equation Model (SEM) that describes the hypothesized and ad-hoc causal relationships between our RSSA features, the subjective constructs(SSA, EXP), the measured personal characteristics(PC), and the interactive components (INT). An SEM can be conceptualized as a series of linear regressions between latent (SSA, EXP) and observed (OSA, PC, INT) variables.

5.4.1 Measurement Model (CFA)

The validated CFA model indicated that 21 questionnaire items with either low factor loadings (< 0.6) or high modification indices, which indicate misfit of the CFA model, those items were removed from the subsequent analyses. Surprisingly, all the six items of the diversity factor

Table 5.7: The factors of personal characteristics with the Average Variance Extracted (AVE) and the consistency coefficients (Cronbach's α), and the items per construct with item factor loadings. Removed items are colored in grey

Considered aspects	Items	Factor loadings
Movie expertise AVE: 0.723 Cronbach's: 0.87	I am a movie lover. Compared to my peers I watch a lot of movies. Compared to my peers I am an expert on movie. I only know a few movies .	0.886 0.926 0.857 -0.717
Need for novelty AVE: 0.628 Cronbach's α: 0.8	When I see a new or different brand on the shelf, I often pick it up just to see what it is like. I like introducing new brands and products to my friends. I enjoy taking chances in buying unfamiliar brands just to get some variety in my purchase. I often read the information on the packages of products just out of curiosity. I get bored with buying the same brands even if they are good. I shop around a lot for my clothes just to find out more about the latest styles.	0.815 0.853 0.701
Fear of missing out (FOMO) AVE: 0.658 Cronbach's α: 0.79	I fear others may find more entertaining movies than me. I get worried when I find out others are finding better movies than me. I get anxious when I think about all the possible movies that are out there. Sometimes, I wonder if I spend too much time trying to make sure I have checked out every interesting movie. It bothers me when I miss an opportunity to learn about new available movies. When I miss out on an opportunity to watch a good movie, it bothers me. Once I decide to go watch a certain movie, I still check on other movies that are playing to see if there is anything better available.	0.635 0.943 0.826
Maximization tendency AVE: 0.596 Cronbach's α: 0.88	No matter what I do, I have the highest standards for myself. I never settle for second best. No matter what it takes, I always try to choose the best thing. I don't like having to settle for "good enough." I am a maximizer. I will wait for the best option, no matter how long it takes. I never settle.	0.781 0.895 0.758 0.746 0.712 0.695 0.801

are among the 21 removed items, three of them show low loadings on diversity, while the other two show high modification indices due to the high correlations between these two items themselves.

All the other nine factors with the remaining 40 questionnaire items had an adequate convergent validity (AVE > 0.50)[6]; the consistency coefficients (Cronbach's α) of the final nine factors (four PC factors and five SSA, EXP factors) showed acceptable to excellent scale reliabilities[7] Table 5.7 and Table 5.8 reflect all the factors and corresponding items along side the Average Variance Extracted (AVE), the consistency coefficients (Cronbach's α), and the respective factor loadings from the CFA. The validated CFA model also meets the discriminant validity requirements[8] (see Table 5.9).

[6]Average Variance Extracted (AVE) is an indicator of the convergent validity of the measurement scales, with the recommended lower bound threshold of 0.5.

[7]A commonly used rule of thumb is that an α of 0.7 indicates acceptable reliability, 0.8 or higher indicates good reliability, and 0.9 or higher indicates excellent reliability [209].

[8]Discriminant validity is established when the $\sqrt{AVE}$ of a factor is larger than its correlations with each of the other factors

Table 5.8: The factors of subjective system aspects (SSA) and the user experience (EXP) with the Average Variance Extracted (AVE) and the consistency coefficients (Cronbach's α), and the items per construct with item factor loadings. Removed items are colored in grey.

Considered aspects	Items	Factor loadings
Recommendation quality AVE: 0.821 Cronbach's α: 0.94	I liked the movies recommended by the movie recommender.	0.964
	I found the recommended movies appealing.	0.97
	The recommended movies fit my preference.	0.95
	The recommended movies were relevant.	0.892
	The system recommended too many bad movies.	-0.851
	I did not like any of the recommended movies.	-0.794
Taste coverage AVE: 0.695 Cronbach's α: 0.89	The movie recommender catered to all of my potential interests.	0.861
	The movies that were recommended did not reflect my diverse taste in movies.	-0.871
	The movie recommender seemed to target only a small subset of my interests.	-0.599
	The movie recommender treated me as a one-dimensional person.	0.883
	The lists of recommendations matched a diversity of my preferences.	
	The recommended movies were a perfect fit for me on many different levels.	0.914
	The movie recommender seemed to stereotype me in a particular category of viewers.	
Conformity AVE: 0.671 Cronbach's α: 0.86	I feel like I was recommended the same movies as everyone else.	
	I think the recommendations are unique to me.	0.927
	I believe that the system is giving me a one of a kind experience.	0.916
	I believe that the movies recommended to me are rather different from the movies recommended to others.	0.761
	I would not be surprised if the system recommended the same movies to many other users.	-0.638
Choice satisfaction AVE: 0.643 Cronbach's α: 0.78	I like the movie I've chosen from the final recommendation list.	
	The chosen movie fits my preference.	
	I would recommend my chosen movie to others/friends.	0.767
	I was excited about my chosen movie.	0.926
	I think I chose the best movie from the options.	
	I know several items that are better than the one I selected.	
	I would rather watch a different movie from the one I selected.	-0.695
System satisfaction AVE: 0.737 Cronbach's α: 0.91	I like using the system.	0.86
	Using the system is a pleasant experience.	
	I would recommend the system to others.	0.908
	I can find better movies using the system.	0.757
	I would quickly abandon using the system.	-0.854
	I would use the system more often if possible.	0.903

Table 5.9: Factor-fit metrics. Off-diagonal values are correlations, diagonal values are the square roots of the average variance extracted $\left(\sqrt{AVE}\right)$ per factor.

	MOVE	NOV	FOMO	MAXT	QUAL	COVE	CONF	ChoSAT	SysSAT
Movie expertise (MOVE)	**0.85**								
Need for novelty (NOV)	0.346	**0.792**							
Fear of missing out (FOMO)	0.422	0.393	**0.811**						
Maximization tendency (MAXT)	0.171	0.326	0.168	**0.772**					
Recommendation quality (QUAL)	0.317	0.185	0.169	0.065	**0.906**				
Taste coverage (COVE)	0.204	0.141	0.098	0.067	0.766	**0.834**			
Conformity (CONF)	0.211	0.231	0.109	0.108	0.679	0.755	**0.819**		
choice satisfaction (ChoSAT)	0.309	0.183	0.096	0.125	0.608	0.597	0.512	**0.802**	
System satisfaction (SysSAT)	0.238	0.344	0.237	0.115	0.654	0.72	0.705	0.553	**0.858**

5.4.2 Structural Equation Model (SEM)

A Structural Equation Model (SEM) was fitted to the nine constructs(SSA, EXP), the experimental manipulation (i.e., the RSSA recommendation lists), and the interactive components (INT). We built our model following the hypotheses formulated in section 5.3.4.2.

We first specified a saturated model with all hypothesized effects and mediations. We then

iteratively trimmed non-significant effects. The resulting model is displayed in Figure 5.5. This model has a good overall fit with $\chi^2(627) = 1411.430$, $p < 0.001$, CFI $= 0.982$, TLI $= 0.985$, RMSEA $= 0.45$ with a 90% confidence interval of $[0.42, 0.48]$[9].

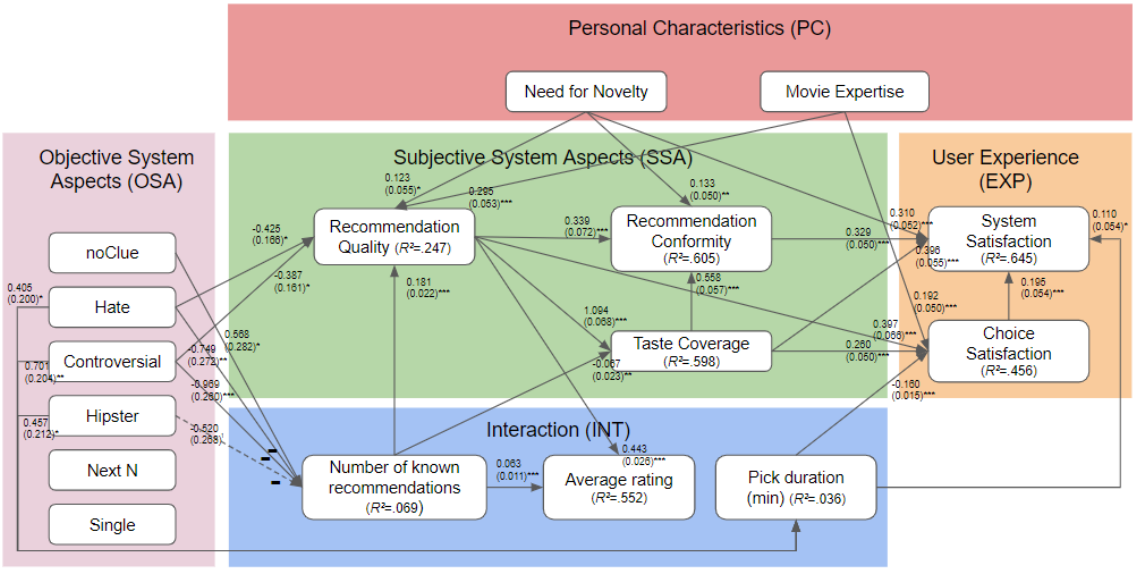

Figure 5.5: The resulting SEM model for the data of the experiment. Significance levels: solid arrows $p < .05$, dashed arrows $p < 0.1$. R^2 is the proportion of variance explained by the model. Minus symbol beside the arrows represent negative effects.

5.4.2.1 Subjective Experience

As shown in Figure 5.5, the RSSA features have a significant effect on users' perceived quality of the recommendations: Participants who received movies of *Things we think you will hate* and *Things that are controversial* on the right scored significantly lower on perceived recommendation quality than participants who only received the single traditional top-N movies (see Figure 5.10 as well). The perceived recommendation quality has a significant effect on participants' satisfaction with the recommendations, this effect is significantly dependent on which recommendation lists participants received (see Figure 5.6): the effect of recommendation quality on choice satisfaction is significantly stronger for participants in the "controversial" condition ($p = 0.092$) but weaker for participants in the "next N" condition ($p = 0.037$) compared to participants who received only the

[9]Theoretically, a good model is not statistically different from the fully specified model (i.e., the p-value of the χ^2 should be > 0.05), but this statistic is commonly regarded as too sensitive [23]. As such, Hu and Bentler proposed cut-off values for the alternative fit indices to be: CFI > 0.96, TLI > 0.95, and RMSEA < 0.05, with the upper bound of its 90% CI falling below 0.10 based on extensive simulations [106].

single top-N recommendations.

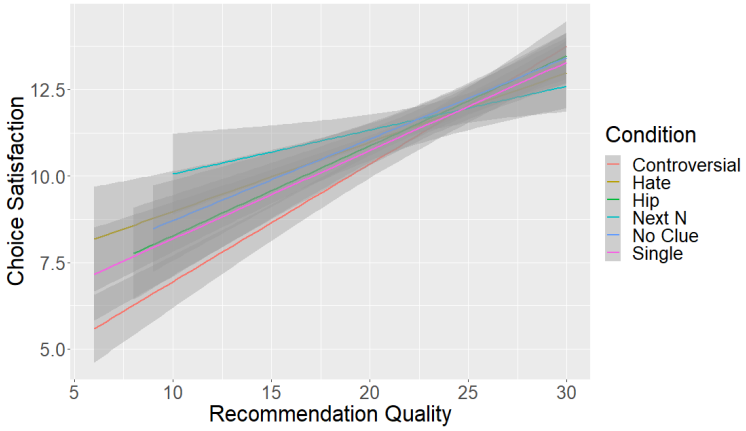

Figure 5.6: The interaction effect of recommendation quality and RSSA features on choice satisfaction.

Participants' decreased perception of the recommendation quality decreases the recommendation conformity, this effect is partially mediated by the perceived taste coverage of the recommendations. The perceived taste coverage is in turn related to participants' satisfaction with the recommendation choices. The perceived conformity and taste coverage of the recommendations finally determine participants' satisfaction with the system.

Figure 5.15 to 5.20 display the total (mediated + direct) effects of the RSSA features on the dependent variables, which presents the influence of RSSA features on users' experience more intuitively. The total effects of the"controversial" condition and the "hate" condition are significantly lower for all dependent subjective variables compared to the single list condition.

5.4.2.2 Personal Characteristics

Participants with higher movie expertise perceived higher recommendation quality and choice satisfaction. Participants who are in higher need for novelty perceived higher in recommendation quality, recommendation conformity, and system satisfaction. Particularly, the effect of need for novelty on perceived recommendation quality is significantly dependent on which recommendation lists participants received (see Figure 5.7): the effect of need for novelty on participants' perceived recommendation quality is significantly less stronger for participants in the "next N" condition ($p = 0.024$) and in the "no clue" condition ($p = 0.018$) compared to participants who received only the single top-N list. However, the extent to which people fear of missing out (FOMO) and

people's perception of their maximizing tendencies (maximization tendency) do not significantly impact participants subjective perceptions of their experience with this study.

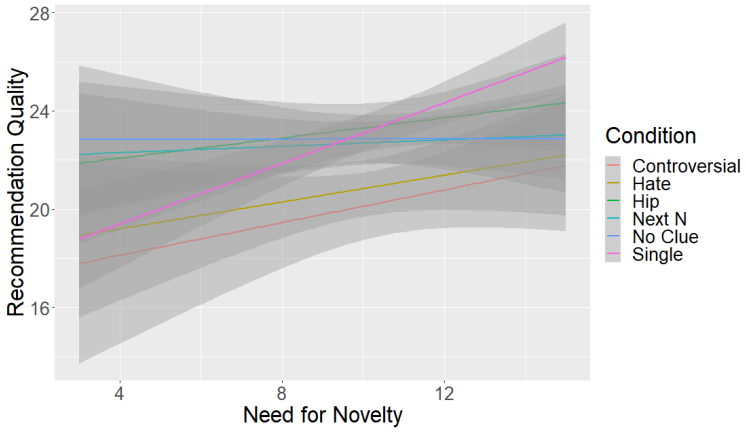

Figure 5.7: The interaction effect of need for novelty and RSSA features on recommendation quality .

5.4.2.3 User Behaviors

Participants who received the recommendations of *Things we think you will hate*, *Things you will be among the first to try*, and *Things that are controversial* indicated that they knew less of the recommendations (on the left list) than those who only received the single traditional top-N recommendations. In turn, the less recommendations the participant already knows, the lower is the perceived recommendation quality, but the higher is the perceived taste coverage.

The perceived recommendation quality and the number of known recommendations determine the average rating participants give to the final top-N recommendations. As shown in Figure 5.9), the marginal effects of the RSSA features on the average ratings of the final top-N recommendations (the left list only) indicate that the average ratings of the final top-N recommendations in the controversial condition (mean: 3.160, $p = 0.005$) and the hate condition (mean: 3.091, $p < 0.001$) are significantly lower than the single condition (mean: 3.446).

The pick duration (the amount of time participants took to select one movie to watch right away) mediate the effects of the RSSA features on participants' perceived choice satisfaction as well as the system satisfaction. In the controversial condition, participants take more time to choose one movie to watch right away ((about 24 seconds more, see Figure 5.8) compared to the next-N condition (I did not compare the pick duration against the single list condition since the total amount

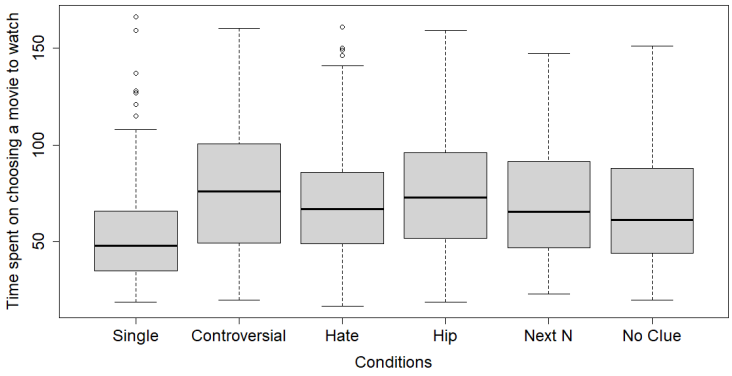

Figure 5.8: The main effect of RSSA features on the amount of time participants took to select one recommendation to consume right away.

of movies to inspect is different).

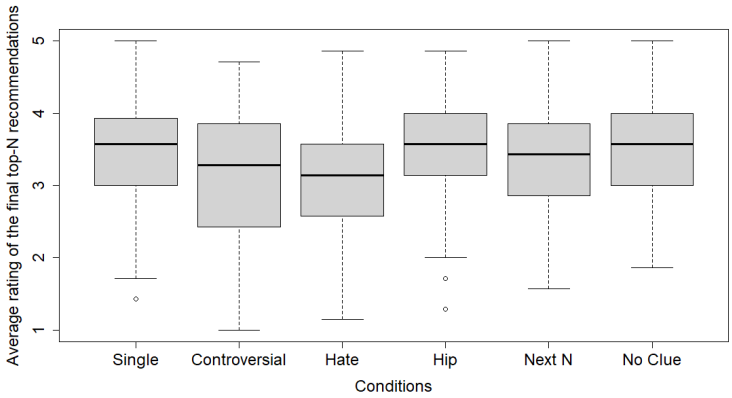

Figure 5.9: The main effect of RSSA features on the average ratings of the final top-N recommendations.

5.5 Discussion

Based on the results of our experiment, we can describe in detail how the benefits of the RSSA features in movie recommenders come about. We can also describe these results in the light of users' personal characteristics. Finally, we can provide some preliminary suggestions on the performance of the RSSA features.

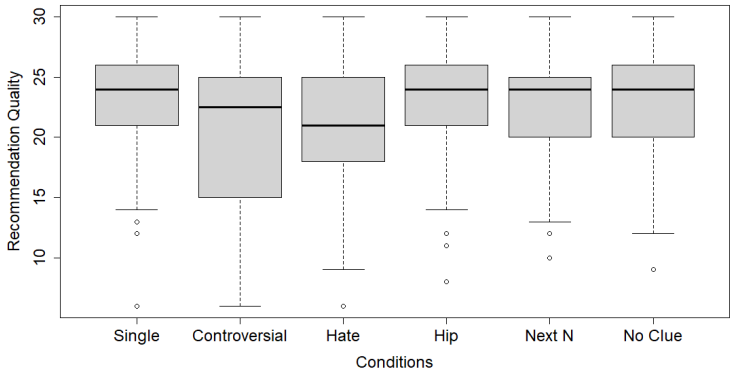

Figure 5.10: The main effect of RSSA features on recommendation quality.

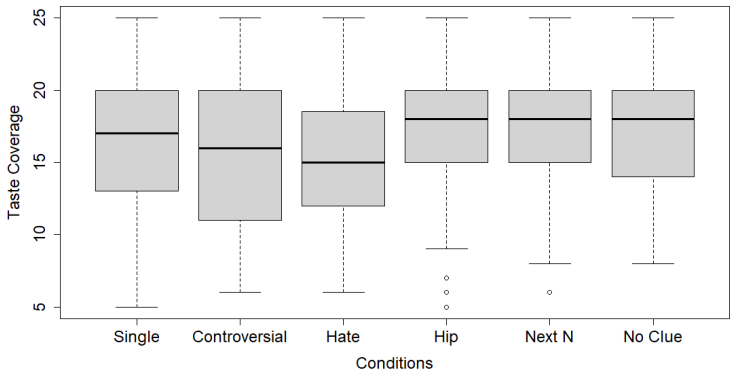

Figure 5.11: The main effect of RSSA features on taste coverage.

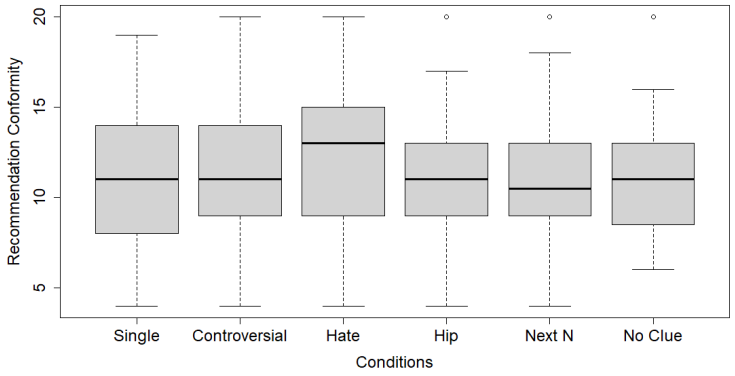

Figure 5.12: The main effect of RSSA features on recommendation conformity.

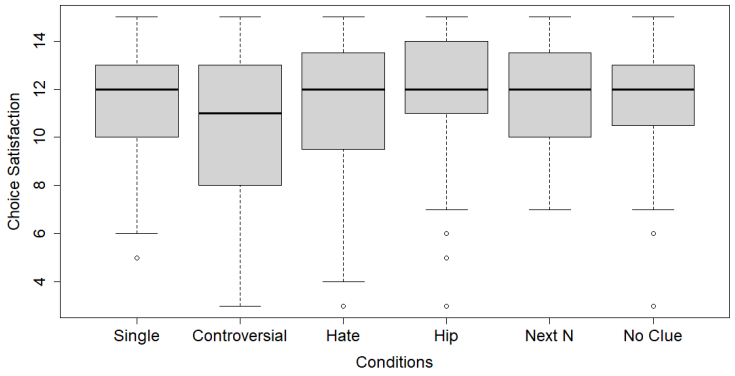

Figure 5.13: The main effect of RSSA features on choice satisfaction.

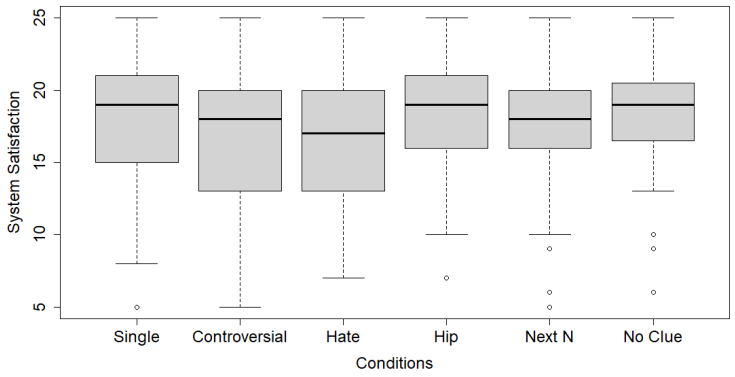

Figure 5.14: The main effect of RSSA features on system satisfaction.

Figure 5.15: The total effects of RSSA features on recommendation quality.

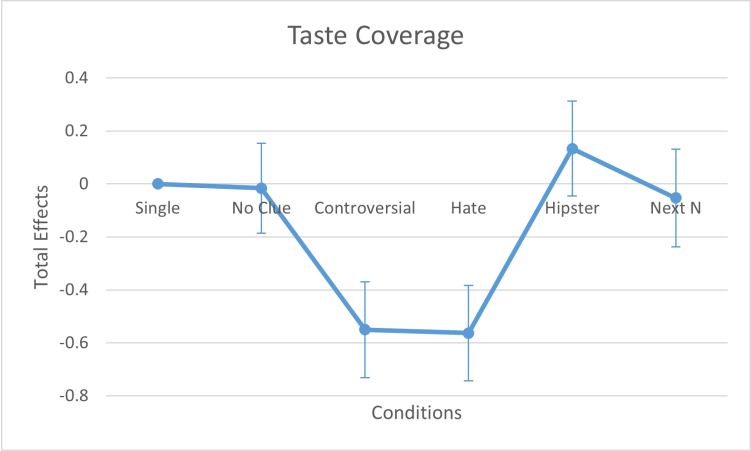

Figure 5.16: The total effects of RSSA features on taste coverage.

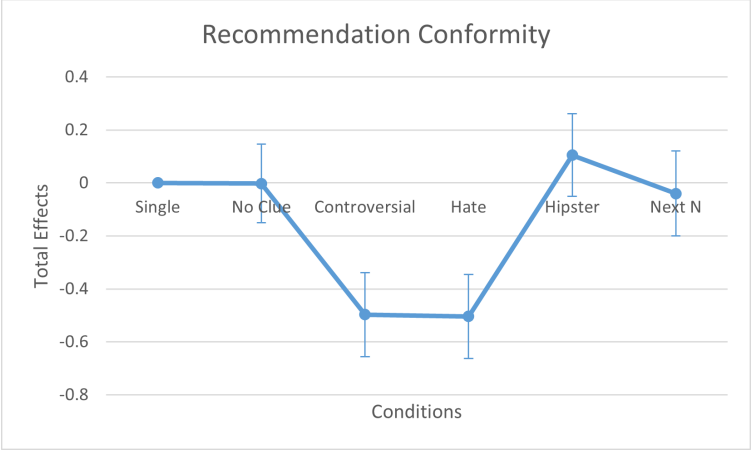

Figure 5.17: The total effects of RSSA features on recommendation conformity.

5.5.1 RSSA Features

The RSSA features have a significant effect on the user experience, primarily because the "controversial" items (*Things that are controversial*) and "hate" items (*Things we think you will hate*) decrease the perceived quality of the recommendations. The decreased recommendation quality in turn causes users to feel less in taste coverage and recommendation conformity, also indicated by the decreased average ratings of the final top-N recommendations. Finally, the lower taste coverage and recommendation conformity cause users to be less satisfied with both the recommendations and the system.

RSSA features work partially due to a direct effect on recommendation quality, and partially due to its influence on user behavior. Specifically, users in the "hate" (*Things we think you*

Figure 5.18: The total effects of RSSA features on choice satisfaction.

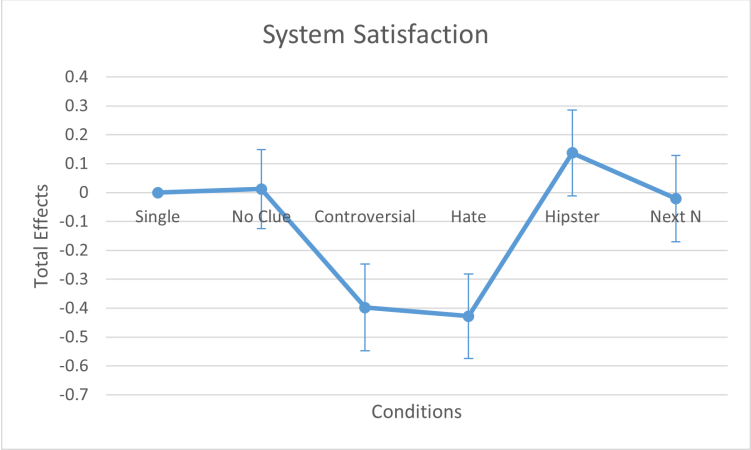

Figure 5.19: The total effects of RSSA features on system satisfaction.

will hate), "controversial" (*Things that are controversial*), and "hipster" (*Things you will be among the first to try"*) conditions already know less of the recommendations while users in the "no clue" (*Things we have no clue about*) condition already know more of the recommendations (the less the recommendations were already known, the lower of the perceived quality of the recommendations, but the higher of the taste coverage). These effects of RSSA features on the number of recommendations that the participant already know is intuitive, because the idea of "hate" , "controversial", "hipster", and "no clue" algorithms seeks to recommend items that are unrelated to the Top-N from different angles: the "hate" recommendations present a list of items the system predicts the user will hate; the "controversial" recommendations present a list of items that are polarized with the Top-N; the "hipster" items present a list of yet-to-be-rated items to users that are identified (using

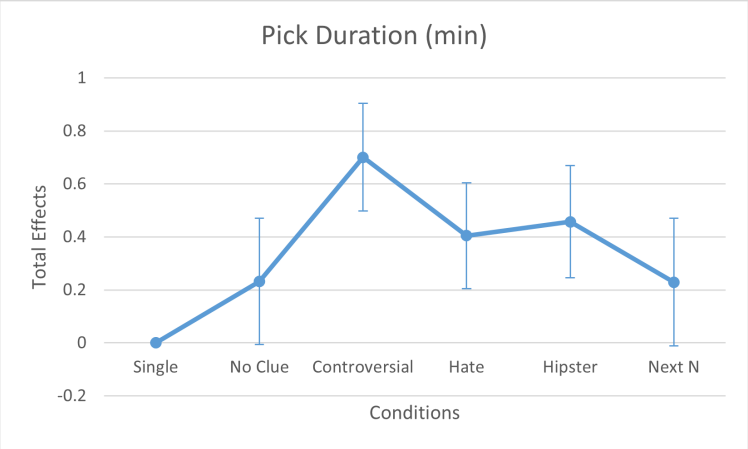

Figure 5.20: The total effects of RSSA features on pick duration (the amount of time participants took to select one movie to watch right away).

a "hipster measure") as having a high willingness to try out new items (thus users knew less of the final top-N recommendations generated after the three algorithms had taken users' confirmation or correction of the recommended items in these three conditions); while the "no clue" recommendations present a list of hard-to-predict items that may be used to identify unexpressed preferences (which shows efficiency of the "no clue" algorithm that the users know more items in the final top-N recommendations after the algorithm took their ratings on the prior two iterations of recommendations). These effects in line with our goal of attempting to support users in developing, exploring, and understanding their unique tastes and preferences [124].

Unsurprisingly and consistently, users take more time to choose a recommendation to consume in the "hate" , "controversial", and "hipster" conditions, since they need more time to inspect the recommendations due to less familiarity with them before picking one to consume. which decreases choice satisfaction, but increases system satisfaction.

From the total effects of the RSSA features on the subjective variables displayed in Figure 5.15, 5.16, 5.17, 5.18, and 5.19, the "hate" and "controversial" items indeed negatively affect the recommendation quality, taste coverage, recommendation conformity, choice satisfaction, and system satisfaction, while the "hipster", the "no clue", and the "next N" items do equally well compared to the single list condition (even though the "hipster" items performs a little bit better on participants' perception of recommendation quality, taste coverage, recommendation conformity, and system satisfaction, but these effects are not statistically significant).

Even though the "hate" and "controversial" items negatively affect user experience with the

system while the "hipster" and the "no clue" items performs equally well compared to the "next N" items and the single list condition, they do support rather than replace decision making (as shown in Figure 5.20). In the follow-up studies, it will be quite interesting to measure three more subjective variables: taste clarification potential (i.e., participants' perception of how well the recommender helps users understand their preferences), taste development potential (i.e., participants' perception of the ability that the recommender helps users develop their unique taste), and self-actualization (i.e., participants' perception of the extent to which the recommendations are attuned to helping them meaningfully improve their own lives [124]) to actually measure the potential of taste clarification and taste development of the RSSA features and participants' perception of self-actualization.

5.5.2 Personal Characteristics

Out of the four personal characteristics we have tested in the prior-study survey — movie expertise, need for novelty, the extent to which people fear of missing out (FOMO), and people's perception of their maximizing tendencies (maximization tendency), two of them have an effect on users' subjective experience when using the systems. Movie expertise has a positive effect on recommendation quality and choice satisfaction, the explanation could be that movie experts are better at judging the quality of the recommendations than non-movie-experts, which are in line with the similar findings in [30, 128, 254, 122].

Moreover, users in higher need for novelty feel higher quality and conformity of the RSSA recommendations, which may be due to the fact that our RSSA features display items outside the Top-N, thus users have more options of items they might hate, yet-to-be-rated items they might be highly willing to try out, and items they have not expressed preferences for [124]. When digging into the perceived recommendation quality together with need for novelty, the "no clue" (*Things we have no clue about*), the "hipster" (*Things you will be among the first to try*), and the "next N" (*More items you may like*) lists work better on the perceived recommendation quality for users with low need for novelty, while the single list works best for users with high need for novelty. Users' need for novelty also has a positive effect on the system satisfaction, which is consistent with the findings in [161].

Surprisingly, maximization tendency and fearing of missing out (FOMO) turn out to have no significant effects on the subjective variables (recommendation quality, taste coverage, recommendation conformity, choice satisfaction, and system satisfaction).

5.5.3 Differences among the RSSA Features

Other than comparing the RSSA features against the "single" condition, we are also interested in comparing the RSSA features against each other. Figure 5.10 and 5.11 show that the perceived recommendation quality and taste coverage are consistently lower for the "controversial" and "hate" conditions than for the "no clue" and "hipster" conditions, these differences are statistically significant. Figure 5.12 shows that the perceived recommendation conformity is significantly higher for the "hate" condition than for the other three conditions (i.e., the "controversial", "hipster", and "no clue" conditions), but only the differences between the "hate" condition and the "controversial" condition is not statistically significant, the other two differences are statistically significant. While the system satisfaction is lower in the "hate" condition than the other three conditions, but only two of differences ("hate" vs. "hipster" and "hate" vs. "no clue) are statistically significant. In Figure 5.13, the choice satisfaction is significantly lower for the "controversial" condition than for the other three conditions, these difference are statistically significant (the difference of the "controversial" condition and the "hate" condition are only marginally significant). The differences between the "hipster" condition and the "no clue" condition is consistently small in all subjective variables and even in the average rating of the final top-N recommendations. The only marginal significant difference between the two control conditions is in duration users took to pick one recommendation to consume right away ($p = 0.06$).

5.6 Conclusion

The results show that the effects of the RSSA features on users' perceptions of recommendation quality, taste coverage, recommendation conformity, and system satisfaction are contrary to our expectation. Items from the "controversial" (*Things that are controversial*) and "hate" (*Things we think you will hate*) perspective are not a good idea according to the results of this empirical experience. Visiting back to the idea of the "hipster" and "no clue" features, however, items from these two perspectives focus more on diversity: the "hipster" items present a list of yet-to-be-rated items to users that are identified (using a "hipster measure") as having a high willingness to try out new items (thus users knew less of the final top-N recommendations generated after the three algorithms had taken users' confirmation or correction of the recommended items in these three conditions); the "no clue" recommendations present a list of hard-to-predict items that may be used

to identify unexpressed preferences (which shows efficiency of the "no clue" algorithm that the users know more items in the final top-N recommendations after the algorithm took their ratings on the prior two iterations of recommendations). Our results show that these two lists perform at least as good as the traditional top-N recommendations, the element of diversity in these two lists is more fruitful than from the perspective of "controversial" and "hate" items. Thus, I shifted away from these RSSA features and focus on diversity of the recommendations instead in the next study (Study IV).

Based on the out-of-expectation results of this study, it could be possible that the features that focus on recommending items diversified by emotions in Study IV may not have an effect on recommendation quality but that could highly potential to help people explore their preferences. To this end, I will measure the taste clarification potential, taste development potential, and self-actualization (which were not measured in this current RSSA study) to actually measure the potential of taste clarification and taste development of the proposed features, as well as participants' perception of self-actualization.

5.7 Summary

The findings of this study suggest that the recommender system with the four alternative recommendation lists (motivated by self-actualization) implemented does not create a better user experience, which implies that providing transparency (Overall RQ2) implicitly (the 4 alternative recommendation lists were embedded in the back-end algorithms in this study) and allowing indirect interactivity (Overall RQ1) might lead to a negative user experience, even though they were designed with a highly promising motivation (i.e., self-actualization, which helps overcome the "filter bubble" problem). Thus in the next study (Study IV), I will design the interactivity and explanation in a more intuitive manner.

Chapter 6

Study IV: Testing a Diverse and Controllable Movie Recommender System

(Note: We plan to submit this chapter to CHI '24 or IUI '24.)

In the previous studies, I described the phases of understanding the effects of explanation (in the forms of visualized explanation, justification, and alternative recommendations revealed from users' preferences) and interactivity on recommender systems (and even on intelligent systems in general).

In Chapter 3, I investigated the effect of interactivity and explanation in the context of a Tic-Tac-Toe XIML system. I found that even in the simple scenario, explanation-driven interactive machine learning (XIML) systems have a better user experience, partially because they encourage users to engage in a mutual feedback loop that helps improve the system's performance. Specifically, XIML systems that allow users to edit the decision rules (as compared to only give feedback on the decision itself) make users feel more in control over the system, which increases the perceived quality of the system's feedback and, in turn, the overall system satisfaction. In Chapter 4, the findings suggest that the presentation of recommendations should emphasise their algorithmic nature, and the justification of recommendations should relate back to users' interests over their needs. This implies that it would provide a better user experience if the movie recommendations are presented

as originating from algorithms rather than from a human expert considering that movie preferences (which will be visualized by the movies' emotion feature) are usually based on users' interest. In Chapter 5, I studied the effects of the four novel RSSA features on recommender systems for self-actualization, however, the results is out of my expectation that the RSSA features do not perform significant better on users' experience with the system. Thus I decided to shift away from these RSSA features and focus on using emotion (from the perspective of item feature) for diversification, interactivity, and visualization to investigate the associated effect on the user experience.

Thus, in this final study, I integrated the explanation-driven (in the form of visualization) interactive mechanisms into a movie recommender system that features using emotions for diversification. By allowing users to explicitly input their emotion preference on movies and immediately get the updated recommendations, accompanied by the implementation of diversification and the information assistance of visualization (which reflects the emotion attributes of the recommendations), the direct interactive process between the user and the system has the potential to improve users' overall experience with the system, and in particular, help users on self-actualizing by supporting them in exploring and understanding their unique personal tastes.

I conducted an online study with this novel system to explore the effects of interactivity (Overall RQ1) and visualization (visualized explanation, Overall RQ2) on user experience and how do they depend on the diversification and personal characteristics (Overall RQ3).

6.1 Overview

Users are faced with endless options to choose from while searching products/services online. One way that would potentially benefit users in making decisions is reading the online reviews, especially for experience products and services – products and services where users do not know how much they like them until they actually experience them. Online reviews for products and services provide a representation of the emotions that the product/service evoked. Dr. Mokryn has leveraged the content of such reviews to develop an eight-dimensional emotional vector describing every product/service on each of the eight emotions of Plutchik's wheel of emotions [191]. This eight-dimensional vector represents the "emotional signature" of the item. This triggered my interest in exploring whether emotional signatures can be used as a novel selection criteria for users to find, evaluate, and select products and services that meet their preferences. I did this by integrating the

emotional signatures into online recommender systems.

Recommender Systems that help users handle the abundance of information and choices are available on today's websites (e.g., e-commerce, streaming). These systems filter the catalog of products, suggesting possible relevant items to a user based on e.g., previous behavior or explicitly stated preferences [201]. However, traditional recommender systems that provide users with personalized recommendations can result in "choice overload" when users are provided with large lists of attractive products or services, especially when users do not have conspicuous preferences. Diversifying the recommendations has been shown to be an effective means to overcome "choice overload" and to provide a balanced set of recommendations. Existing diversification strategies are based on either objective (e.g. author, genre) or latent features [254, 149]. However, diversification based on items' latent features limits the average user's perception of diversity [137, 60] due to the complexity and opaqueness of the diversification algorithm as well as the lack of expert knowledge of the diversity dimensions such as the latent feature [254, 149] that extracted from the matrix factorization [133].

Based on the existing work that Mokryn et al. have done on extracting emotional signatures from the online reviews of movies, I consider using emotion as an item attribute to diversify movie recommendations. Considering the tangible characteristic nature of emotion, diversifying movie recommendations by emotion would potentially help users to perceived the diversity. Over the past two decades, the field of psychology has dedicated significant attention to investigating the influence of emotions on the decision-making process [65, 145, 189, 190]. More recently, researchers have put forth the hypothesis that users' emotions and personalities play a crucial role in understanding the variations in their preferences. This understanding has the potential to contribute to the enhancement of personalized systems [236]. The relevance of the emotions evoked from item reviews and their importance to viewers' experience have long been recognized in the movie domain [85, 227].

In this study, my goal is to develop a novel recommender system that diversifies its recommendations based on the emotional signature of items. I thus built this diversification mechanism in line with our original intention of creating "recommender systems for self-actualization" [124, 91] – systems that follow a more holistic human-centered personalization practice by supporting users in developing, exploring, and understanding their unique tastes and preferences [124]. The diversified recommendations enable the recommender to gain a more holistic view of the user from the perspective of emotions and allow the user to learn more about themselves as well.

Recommender systems have appeared to users as black boxes due to the complex algorithms

and techniques implemented. This opaqueness can lead to feelings of discomfort or even creepiness despite the recommendation matching a user's interest and shows high accuracy [128]. Explanations are an important mechanism to promote trust, as they provide human-understandable interpretations of the inner working of the system [76], and visualization leverages visual representations to facilitate human perception [96]. Thus, I consider making the emotion-based recommender system[1] more intuitive by visualizing the emotions of movies to reveal the novelty of this system. In the meanwhile, for the purpose of self-actualization, I believe that allowing users to take control over the diversification process would benefit users in exploring their unique tastes from the perspective of emotions. This is approachable in this emotion-based recommender system by allowing users to interactively specify which emotional dimensions they want the system to prioritize. Making the recommendation process transparent and controllable has the potential to alleviate the negative effects caused by the opaqueness of the complex algorithms and techniques implemented in recommender systems. This effect aims to make recommendations more aligned with users' long-term goals and ambitions.

For my final study, I designed, developed, and evaluated these means of leveraging emotional signatures for recommender systems by combining three distinct directions (diversification, visualization, and interactivity). My ultimate goal is to support users in exploring and understanding their unique personal tastes from the perspective of emotions extracted from the online reviews of items. Although existed works have investigated individually or in combinations of the three directions in the area of recommender systems [76, 254, 181], to the best of my knowledge, no research has explored the interplay of these directions in emotion-based recommender systems.

With a movie recommender system with the above ideas integrated, I aim to address the following research questions:

Study I - RQ1: Does using the evoked emotions from movie reviews for diversification contribute to users' perceived diversity of the recommendations?

Study II - RQ2: How does the visualization influence the transparency of the recommender system that leverages the emotional signature?

Study III - RQ3: Does users' control over the recommender system improve their perceived interactivity of the system and thus increase their satisfaction with the system?

[1]The *emotion-based recommender* here refers to a recommender that re-ranks and diversifies the recommendations based on their emotional signatures

Table 6.1: Data used in the system.

Data Source		Original Dataset	Used Dataset
IMDB & Movie Lens	Ratings	24,702,320	15,056,975
	Movies	57,533	9,064
	Users	162,541	133,969

Study IV - RQ4: How do users' personal characteristics moderate the effects of the main features of the movie recommender on users' experience with the system?

6.2 Data and Recommendation Algorithms

6.2.1 Data

In this final study, both the movie rating dataset from MovieLens [2] and movie emotional signature dataset were used. The MoviesLens 25M rating dataset — contains around 25 million 1 to 5 star ratings on 62,423 movies by 162,541 users (users were selected randomly for inclusion, and all selected users had rated at least 20 movies) — was used to generate the traditional top-N recommendations, and the associated movie information such as synopsis, cast, genre, and poster were enriched from the IMDB database that can be accessed on the public IMDB database website[3]. The emotional signatures of the movies — eight-dimensional vectors represents the "emotional signature" of the movies, generated by a system built by an IS Capstone Project group which Dr. Mokryn guided [19, 170] — were used for diversification, visualization, and interactivity.

These two data sets consist of two different movie sets with ratings and emotional signatures. I merged the two data sets by excluding movies without emotional signatures extracted (due to some technical reasons such as insufficient review numbers) from the rating dataset in other words, only movies with both valid user ratings and valid emotional signatures, as well as movie information extracted were used in the database for this proposed work (as shown in Table 6.1). The final dataset contains a total of 15,056,975 ratings on 9064 movies by 133, 969 users.

[2]accessible at https://grouplens.org/datasets/movielens/
[3]https://datasets.imdbws.com/

6.2.2 Recommendation Algorithms

6.2.2.1 Matrix Factorization

The algorithms for recommending in this system are based on collaborative filtering, an approach that identifies user-item associations by detecting the interdependencies between items and relations between users. Collaborative filtering has been widely used in both commercial applications [143] and academic studies [199, 221]. Among collaborative filtering approaches, latent factor modeling is a technique that characterizes both items and users by vectors of "feature factors" inferred from item rating patterns through dimensional reduction. Recommended items have a vector that shows a high correspondence with the current user's vector [133]. Matrix factorization is the most common realization of latent factor models, which is widely used in recommender systems as well.

I adopted the LensKit for Python[4] [59] that was developed by Ekstrand as a supporting platform to generate the traditional top-N recommendations with the matrix factorization algorithm applied.

6.2.2.2 Diversification Algorithm

The diversification algorithm from Willemsen et al. [254] was adopted for diversifying in this system. Willemsen et al. used latent features extracted from matrix factorization for diversification while I used the emotional signatures instead for diversification. The pseudo-code of the diversification algorithm using emotional signature is shown in Table6.2, I applied the exact same diversification technique when diversifying the the traditional top-N recommendation by emotional signatures. I implemented this diversification algorithm on the LensKit for Python platform.

6.2.2.3 Taking Users' Emotion Input

As mentioned above, each movie has the emotion feature reflected on an 8-dimension vector termed emotional signature, which describes every movie on each of eight emotions of Plutchik's wheel of emotions [191]. We allow users to specify their emotion taste in movies, more specifically, of all the 8 emotions (anticipation, anger, trust, disgust, fear, joy, sadness, and surprise), users can specify either 'more' or 'less' on each emotion (corresponds to a positive weight '0.125' for 'more'

[4]https://lkpy.readthedocs.io/en/stable/

Table 6.2: Diversification Algorithm

Algorithm: Diversification
Input: Top-N predicted items T, Item emotional signatures $Q_T = \{i \epsilon T, q_i \epsilon Q_T\}$, h (number of diversified item set H) } Set $H = \{\}$, an empty set; Get $centroid(Q_T)$, take the average along each dimension of the emotional signatures over all the candidate top-N items; *first-item* = the $i \epsilon T$ for which $d(Q_i, centroid(Q_T))$ is minimal; Add *first-item* to H and Remove it from T; while $num(H) < h$: *next-item* = the $i \epsilon T$ for which $\sum_{j \epsilon H} d(Q_i, Q_j)$ is maximal Add *next-item* to H and Remove it from T End

and a negative weight '-0.125' for 'less', respectively) to indicate how much extent they would like the recommended movies to have on the emotion if possible. Among the eight emotions, users can choose to specify only some of the 8 emotions and leave others unspecified.

In the diversification version where the diversified recommendations are shown,

- Step 1: the algorithm takes the predicted top 200 movies — generated from the traditional top-N algorithm with Matrix Factorization applied — as candidates;
- Step 2: the algorithm diversifies the 200 candidates by the unspecified emotions to generate a diversified ranking;
- Step 3: the algorithm calculates a new ranking score for each item by integrating the weights of the specified emotions and the diversified ranking;
- Step 4: the algorithm recommends the seven movies with the highest new ranking scores.

the pseudo-code of this algorithm is shown in Table 6.3.

In the non-diversified version where the top-N recommendations are shown, the only difference in taking users emotion input is that in step 2, the algorithm does not diversify the top 200 movies (candidates) by the unspecified emotions, instead, the algorithm generates a top-N ranking sorted by the original predictions from the Matrix Factorization algorithm for the follow-up calculations stated in step 3 and step 4.

Table 6.3: Tuned-Diversification Algorithm

Algorithm: Tuned Diversification
Input: Top-N predicted items T, User-unspecified emotion set E_u, Item emotional signatures $Q_T = \{i \in T, q_i \in Q_T\}$, h (number of diversified item set H). Apply the above diversification algorithm described in Table 6.2 to the item candidate set T on the user **unspecified** emotion set E_u, generate a new diversified ranking for the item candidate set r_d; For for the **specified** emotions e_i ($i \in \{1, 2, 3, ..., 8\}$), the 'high' label corresponds to a positive weight of $w_i = 0.125$ while the 'low' label corresponds to a negative weight of $w_i = -0.125$; Apply the following formula to each candidate item to calculate a new ranking score r_n: $r_n(item) = \sum_{i=1}^{8}(w_i * e_i) + (1 - \sum_{i=1}^{8}\|w_i\|) * r_d(item)$ Sort items by the new ranking score and get the top 7 items as the new recommendations; End

6.3 Study Setup

6.3.1 Procedure

This study was submitted to our Institutional Review Board (IRB) and considered exempt. Participants were recruited from the Prolific platform [5]. Participants were limited to adult users located in the United States and to be fluent in English.

The procedure, summarized in figure 6.1, contained the following steps:

1. *Introduction and consent:* Upon joining the study, participants were shown a welcome message that provided introductory information about the study to provide insights into expectations. They were then shown the consent information related to the study. After the participants agreed to participate in this study, they were then offered an overview of all of the steps in the study: pre-survey, indicating preference, interacting with the system, and post-survey.

2. *Pre-study survey:* In this section, participants were asked about personal characteristics that could influence users' decision-making when interacting with an online movie recommender such as movie expertise, need for novelty, and their familiarity with visualization. These questions were presented at this stage to avoid confounding effects after interacting with the system.

3. *Preference Elicitation:* To provide a personalized experience, we need to first collect a sub-set of recommendations to better understand user preferences. Participants were asked to rate at least 10 movies that they have seen before on a scale of five. After rating at least the minimum (10 movies), they were allowed to proceed to the next page. We added a buffering

[5] https://www.prolific.co/

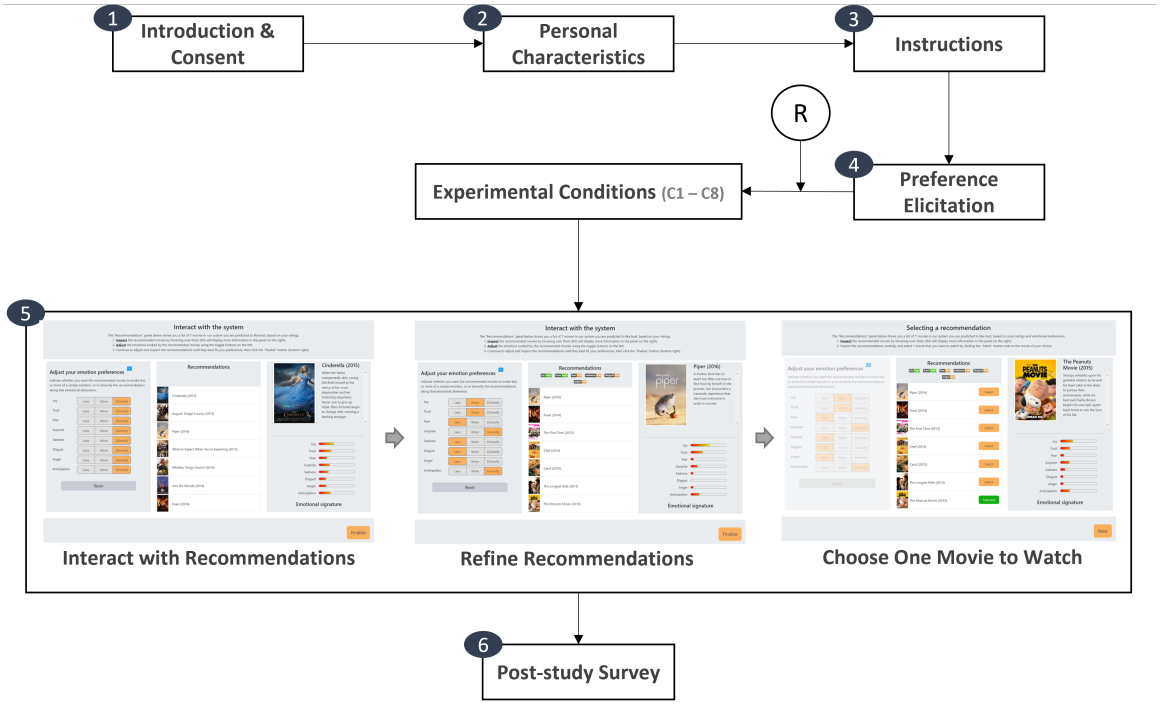

Figure 6.1: The key procedural steps of the experiment. "R" denotes random assignment to the experimental conditions.

page before the next page which asked participants to "please wait while the system prepares your recommendations". This is done to enhance perceptions of a personalized experience.

4. *Interaction with the recommender:* Based on the ratings from the preference elicitation stage, the system generated recommendations for the participants "lively". For the conditions without the control panel which allows participants to specify their tastes on emotions of the movies, each participant was asked to inspect the movies by hovering the recommendation list and select one movie that they would like to watch right away if they could. For the conditions with the control panel, we iterated the recommendations on two progressive recommendation pages: on the first page, after participants finished inspecting the recommended movies, they were asked to specify their tastes in movie emotions and play around with the control panel to interact with the recommender to get recommendations based on their emotion tastes in movies; on the second page, the control panel was locked off and participants were asked to select one movie that they would like to watch right away if they could.

5. *Post-study survey:* Lastly, participants were asked to evaluate their perceptions of their in-

teractions with the system. We firstly presented one feedback page to allow users to inform us issues if any by asking participants "Did anything go wrong while using the system?", which also serves as an attention check. The page were then followed by pages of the post-survey questionnaire. All of the factors that were evaluated are described in section 6.3.4. Demographic questions were asked towards the end of the study.

6.3.2 Participants

A power analysis with a 0.25 effect size and a 0.85 power shows that a total of 146participants are needed for this study. Each participant received 2.75 US dollars as compensation at the end of the study.

Other then the feedback page following the recommendation mentioned above, there are two extra attention checks in the style of "Please select 'agree' for this question regardless" randomly placed among the questionnaires, one was in the pre-study survey and the other in user experience questions, to track if participants were actually paying attention to what they are doing during the completion of the study. We used these questions plus the time taken to complete the study to filter out participants who clearly rushed through the study.

In the actual online study, 282 participants were collected from the Prolific platform, 6 participants failed the attention checks, which ended up with 276 usable data points for data analysis. The sample includes 52 participants between the ages of 18 and 24, 93 between 25 and 34, 62 between 35 and 44, 46 between 45 and 54, and 23 older than 54. Most participants completed the study in 15 minutes. Among these 276 participants, 148 identified as women, 118 as men, 8 as non-binary, and 2 participant preferred not to answer our gender question.

6.3.3 Experimental Design

As aforesaid, this recommender system features emotion-based diversification, visualization, and interactivity, so the experiment involves these three between-subjects manipulations: diversification, visualization, and interactivity. The emotion-based diversification technique were tested against the traditional top-N recommendations as the baseline, as shown in Figure 6.2), the middle panel (panel 1) displayed either the top N recommendations or the recommendations diversified by emotional signatures (i.e., diverse N recommendations). The visualization focused on the emotion

feature of the movies, which explicitly displays the emotions of a movie on eight dimensions (the bar graph of emotional signature, i.e., panel 2 in Figure 6.2) so that the users can intuitively understand the movie's tone on the perspective of emotions. The visualization was tested against not to show any visualization of the emotions.

The interactivity was tested by either allowing or not allowing users to interact with the system by specifying the relative weight of the emotional dimensions underlying their preferences (the default values of the weights are included in Table 6.3). The design of the emotion preference panel is labeled as panel 3 in Figure 6.2. In the conditions where users were allowed to interact with the system (i.e., specify the relative weight of the emotion dimensions by tweaking the toggle buttons on the emotion preference panel as shown in Figure 6.2), the system took their inputs and recommend items based on the conditions they are assigned to (more details described in Section 6.2.2.3).

Our experiment thus includes two experimental conditions in each manipulation, resulting in a $2 \times 2 \times 2$ between-subjects controlled experiment (as shown in Table 6.4).

Table 6.4: Manipulations and conditions.

Manipulations		Diverse Recommendations	Top-N Recommendations
Allow Input	**Visualization**	**C1:** Diverse items With emotion input panel With visualization panel	**C5:** Top-N items With emotion input panel With visualization panel
	No Visualization	**C2:** Diverse items With emotion input panel Without visualization panel	**C6:** Top-N items With emotion input panel Without visualization panel
No Input	**Visualization**	**C3:** Diverse items Without emotion input panel With visualization panel	**C7:** Top-N items Without emotion input panel With visualization panel
	No Visualization	**C4:** Diverse items Without emotion input panel Without visualization panel	**C8:** Top-N items Without emotion input panel Without visualization panel

6.3.4 Measurements and Hypothesized Structural Equation Model (SEM)

We measured both participants' personal characteristics (i.e., movie expertise, need for novelty, visualization familiarity; reflecting participants individual differences) and their perceptions of the subjective system aspects and user experience with the system with the following measurement scales:

- `Movie expertise`(PC): participants' perception of their movie expertise (adopted from [121]).

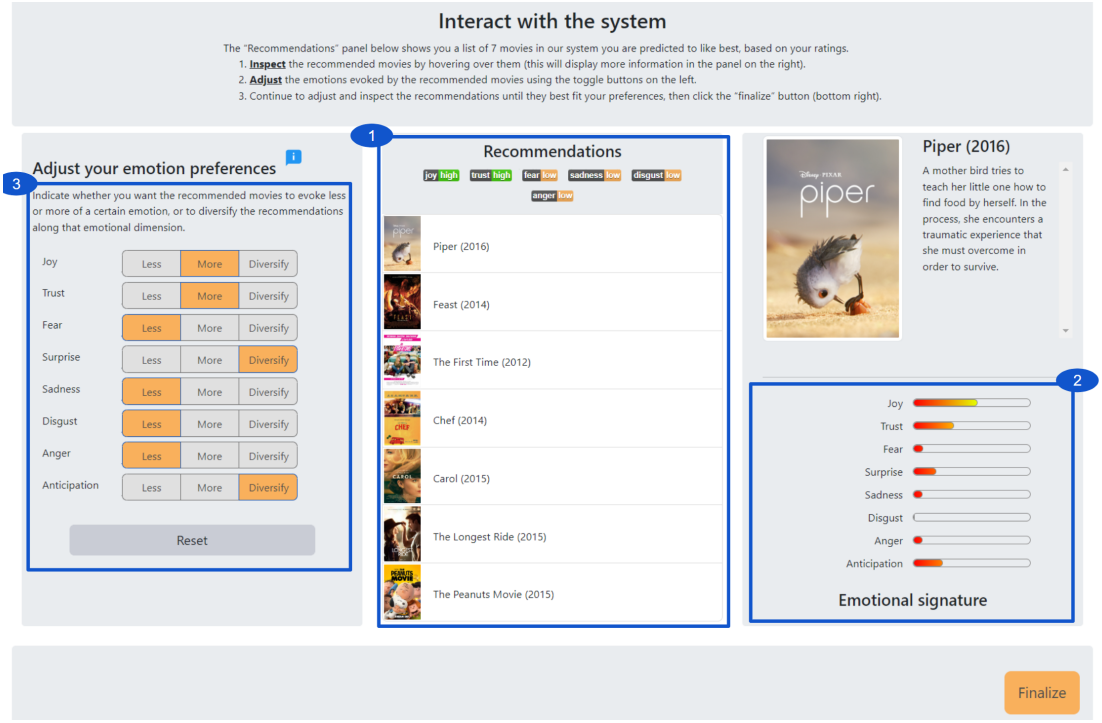

Figure 6.2: The interface with the three experimental manipulations of the recommendation page in the actual system.

- `Need for novelty`(PC): participants' perception of their need for novelty (adopted from [38]).
- `Visualization familiarity`(PC): participants' familiarity with visualization (adopted from [135, 134]).
- `Perceived interactivity`: the extent to which participants perceive that they are able to interact with the recommendation process.
- `Understandability`: participants' perception of the understandability of the emotion-based recommendations (adopted from [121]).
- `Perceived diversity`: participants' perception of the diversity of the recommendations (adopted from [254]).
- `recommendation quality`: participants' perception of the quality of the recommendations (adopted from [121]).
- `Taste coverage`: participants' perception of the extent that the recommender system reflects all of their tastes

- `Taste clarification potential`: participants' perception of how well the recommender helps users understand their preferences.
- `Taste development potential`: participants' perception of the ability that the recommender helps users develop their unique taste.
- `Choice difficulty`: participants' perception of the difficulty of selecting movies from the recommendations (adopted from [30, 254]).
- `System satisfaction`: participants' satisfaction with the recommender system (adopted from [121]).
- `Choice satisfaction`: participants' satisfaction with the selected recommendations (adopted from [30, 254]).
- `Self-actualization`: refers to participants' perception of the extent to which the recommendations are attuned to helping them meaningfully improve their own lives [124].

The experimental manipulations and the measured constructs were integrated into the following path model (as shown in Figure 6.3). I hypothesize that:

Hypothesis 1: *The diverse recommendations based on emotions increase the perceived diversity.*

Hypothesis 2: *The emotion visualization of the recommendations increases the understandability.*

Hypothesis 3: *Users' ability to edit the emotional preference of the movies increases the understandability and perceived interactivity.*

Hypothesis 4: *The perceived diversity reduces the difficulty of the selection of the recommendations and increases the perceived recommendation quality (perhaps mediated by the choice difficulty).*

Hypothesis 5: *Understandability of the recommendations is hypothesized to reduces the choice difficulty and increases the taste development potential.*

Hypothesis 6: *The perceived interactivity is then hypothesized to increase the perceived recommendation quality.*

Hypothesis 7: *The increased perceived diversity and recommendation quality is hypothesized to increase the perceived taste coverage as well as the taste clarification potential.*

Hypothesis 8: *The perceived interactivity, choice difficulty, perceived recommendation quality, and taste coverage ultimately have significant effects on choice satisfaction.*

Hypothesis 9: *The taste clarification potential and taste development potential are hypothesized to increase users' satisfaction with the system.*

Hypothesis 10: *The choice satisfaction and system satisfaction are expected to increase users' perception of self-actualization.*

Hypothesis 11: *Users' need for novelty contributes to their perceived diversity of the diverse recommendations.*

Hypothesis 12: *Users' familiarity of visualization as well as their movie expertise increase the perceived recommendation quality (perhaps mediated by understandability).*

Hypothesis 13: *Users' education level is correlated with the understandability.*

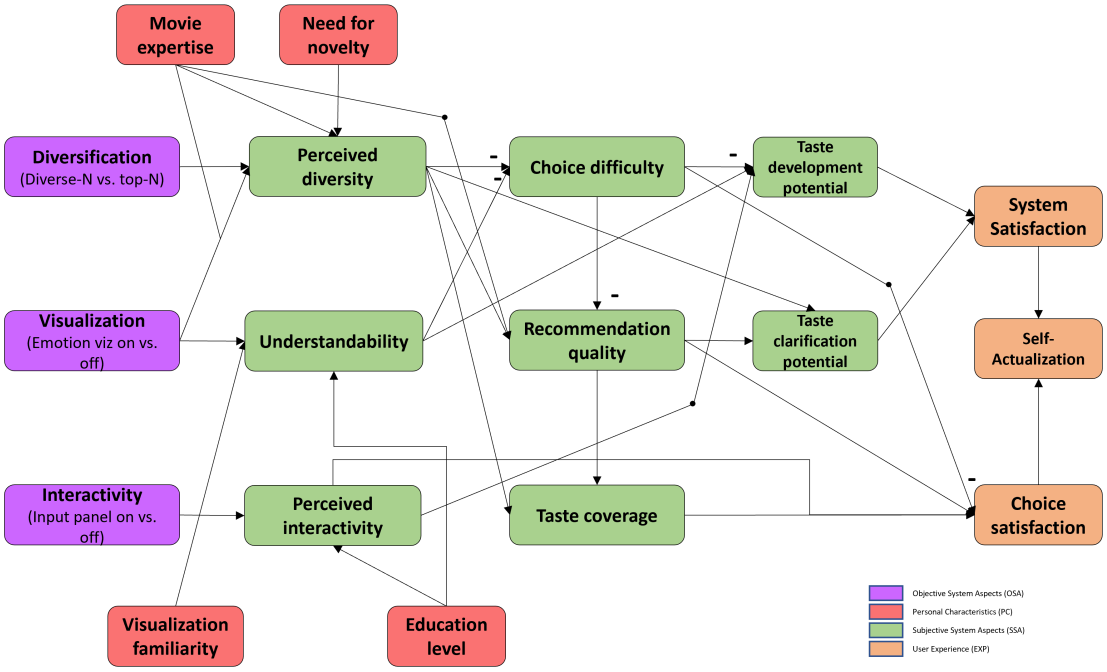

Figure 6.3: The hypothesized path model.

6.4 Results

With the collected survey data, I first verified the the diversity of the diversified recommendations with the defined AFSR metric (more details in Section 6.4.1). I validated our measurement

model regarding the PC, SSA and EXP constructs using a Confirmatory Factor Analysis (CFA) and then fitted a Structural Equation Model (SEM) that describes the hypothesized and ad-hoc causal relationships between the objective system manipulations (OSA), the subjective constructs(SSA, EXP) and the measured personal characteristics(PC).

6.4.1 Objective Diversity of the Initial Recommendations

Among the three manipulations (i.e., diversification, visualization, and interactivity), visualization and interaction were intuitively reflected on the interface of the recommendation page (see Figure 6.2) through the design of the emotion input panel (panel 3) and the bar graph visualization panel of emotional signature (panel 2), while diversification was implicitly reflected in the algorithm integrated in the back end of the system, thus I firstly check on the objective diversity of the initial recommendations (generated before the system took into account participants' emotion preferences) of all the experimental conditions. Referring to the metric AFSR (Average Feature Score Range) used to measure the objective diversity of recommendations diversified by the item latent features [254], I applied this same metric but replaced the latent feature with emotion signature since I diversified the recommendations by the emotion signatures of movies, hereby, I labelled this metric as AESR (Average Emotion Score Range). The statistical result shows that the AESR of the diverse N recommendations is significantly higher than the AESR of the traditional top N recommendations ($M = 0.069$, $\beta = 0.051$, $p < .001$, which verifies that the diverse N recommendations significantly differs from the traditional top N recommendations from the perspective of emotional signature (see Figure 6.4).

6.4.2 Confirmatory Factor Analysis (CFA)

I ran a CFA model on the collected data to validate the reliability and validity of the measured 14 scales (3 PC factors and 11 SSA and EXP factors). Items with low factor loadings (< 0.6) or high modification indices (both indicate misfit of the CFA model) were excluded from the subsequent analyses. The consistency coefficients (Cronbach's α) were also calculated to check the reliability of the constructs, the results show all the constructs (3 PC factors and 11 SSA and EXP factors) have at least acceptable reliability metrics[6]. I checked the convergent validity and the

[6]A commonly used rule of thumb is that an α of 0.7 indicates acceptable reliability, 0.8 or higher indicates good reliability, and 0.9 or higher indicates excellent reliability [209].

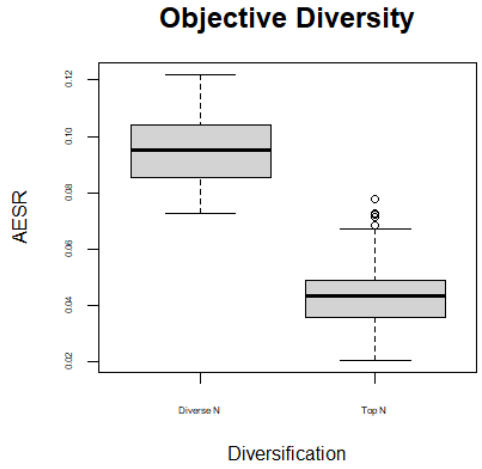

Figure 6.4: The Averaged Emotion Score Range (AESR) of the initial recommendations.

discriminant validity of the CFA model as well — the convergent validity was assessed by checking if the average variance extracted (AVE) are greater than 0.5 [7], all the measured factors have AVEs greater 0.5; while the discriminant validity was assessed by comparing the average variance extracted (AVE) of each factor against its correlation with other factors. I found that the *taste development potential* construct is high correlated with *taste clarification potential*, *system satisfaction*, and *self-actualization*, so I removed *taste development potential* from the CFA model to avoid multicollinearity in the subsequent SEM modeling.

I again ran the CFA model with the remaining 13 factors, no more items were further trimmed (all the remaining items have factor loadings greater than 0.6), this final 13-factor CFA model also meets the convergent validity and discriminant validity (see Table 6.5, 6.6, 6.7).

6.4.3 Structural Equation Model (SEM)

With the build 13-factor CFA model, I then fitted a Structural Equation Model (SEM) to the 13 constructs and the experimental manipulations (i.e., interactivity, diversification, and visualization). An SEM enables one to specify the relationships between exogenous variables (the manipulations) and latent constructs (the CFA factors) as a structured model of regressions [127]. An important benefit of SEM is that fit statistics are provided for the model as a whole, as well as

[7] Average Variance Extracted (AVE) is an indicator of the convergent validity of the measurement scales, with the recommended lower bound threshold of 0.5.

Table 6.5: The factors of Personal Characteristics (PC) with the Average Variance Extracted (AVE) and the consistency coefficients (Cronbach's α), and the items per construct with factor loadings. Items removed from the CFA model are colored in grey

Considered aspects	Items	Factor loadings
Need for novelty (NOV) AVE: 0.516 Cronbach's α: 0.76	When I see a new or different brand on the shelf, I often pick it up just to see what it is like.	0.704
	I like introducing new brands and products to my friends.	0.862
	I enjoy taking chances in buying unfamiliar brands just to get some variety in my purchase.	0.685
	I often read the information on the packages of products just out of curiosity.	
	I get bored with buying the same brands even if they are good.	
	I shop around a lot for my clothes just to find out more about the latest styles.	0.597
Movie expertise (MOVE) AVE: 0.755 Cronbach's α: 0.89	I am a movie lover.	0.938
	Compared to my peers I watch a lot of movies.	0.886
	Compared to my peers I am an expert on movies.	0.887
	I only know a few movies.	-0.752
Visualization familiarity (VizF) AVE: 0.783 Cronbach's α: 0.93	I am competent when it comes to graphing and tabulating data.	0.832
	I frequently tabulate data with computer software.	0.918
	I have graphed a lot of data in the past.	0.934
	I frequently analyze data visualizations.	0.915
	I am familiar with data visualization.	0.859
	I am an expert at data visualization.	0.846

for the individual regression coefficients. We built our model following three principles:

1. The experimental manipulations (ie. diversification (RQ1), visualization (RQ2), and interactivity (RQ3) were hypothesized to influence *understandability*, *Perceived interactivity*, *Perceived diversity*.

2. Each effect outlined in (1) is allowed to mediate the subsequent effects of the manipulations on the subsequent SSA and EXP constructs (i.e. *recommendation quality*, *taste coverage*, *taste clarification potential*, *choice difficulty*, *system satisfaction*, *choice satisfaction*, *self-actualization*).

3. The personal characteristics (RQ4) were hypothesized to moderate users' perceptions of the system.

We first specified a saturated model with all hypothesized effects and mediations. We then iteratively trimmed non-significant effects. The resulting model is displayed in Figure 6.5. This model has a good overall fit with $\chi^2(1780) = 2488.651$, $p < 0.001$, CFI = 0.974, TLI = 0.979, RMSEA = 0.38 with a 90% confidence interval of [0.34, 0.42][8].

[8]Theoretically, a good model is not statistically different from the fully specified model (i.e., the p-value of the χ^2 should be > 0.05), but this statistic is commonly regarded as too sensitive [23]. As such, Hu and Bentler proposed cut-off values for the alternative fit indices to be: CFI > 0.96, TLI > 0.95, and RMSEA < 0.05, with the upper bound of its 90% CI falling below 0.10 based on extensive simulations [106].

Table 6.6: The factors of Subjective System Aspects (SSA) and User Experience (EXP) with the Average Variance Extracted (AVE) and the consistency coefficients (Cronbach's α), and the items per construct with factor loadings. Items removed from the CFA model are colored in grey

Considered aspects	Items	Factor loadings
Perceived interactivity (INT) AVE: 0.648 Cronbach's α: 0.88	I felt in control of the movies shown to me.	0.912
	I felt in control of the movie recommendation process.	0.901
	I feel unable to intervene in the movie recommendation process.	-0.758
	I was able to interact with the movie.	
	I was able to manage the movies shown to me.	0.796
	I was able to supervise the process through which the movies are shown to me.	
	I had limited control over the way the movie recommender made recommendations.	-0.801
	The system restricted me in my choice of movies.	
	Compared to how I normally get recommendations, this movie recommender was very limited.	
	I would like to have more control over the recommendations.	
	I decided which information was used for recommendations.	0.630
Understandability (UND) AVE: 0.771 Cronbach's α: 0.90	The recommendation process is clear to me.	
	I understand how the system came up with the recommendations.	0.911
	The movie recommender explained the reasoning behind the recommendations.	0.769
	I am unsure how the recommendations were generated.	-0.941
	The recommendation process is not transparent.	-0.882
Perceived diversity (DIV) AVE: 0.841 Cronbach's α: 0.88	The recommended list of movies suits a broad set of tastes.	0.944
	The recommended movies were from many different genres.	0.836
	The recommendations contained a lot of variety.	0.965
	None of the movies in the recommended list were alike.	
	All the recommended movies in the final list were similar to each other.	
	Most movies were from the same genre.	
Recommendation quality (QUAL) AVE: 0.900 Cronbach's α: 0.94	I liked the movies recommended by the movie recommender.	0.960
	I found the recommended movies appealing.	0.975
	The recommended movies fit my preference.	0.947
	The recommended movies were relevant.	0.911
	The system recommended too many bad movies.	
	I did not like any of the recommended movies.	
Taste coverage (COVE) AVE: 0.681 Cronbach's α: 0.89	The lists of recommendations matched a diversity of my preferences.	0.865
	The movie recommender catered to all of my potential interests.	
	The recommended movies were a perfect fit for me on many different levels.	
	The movies that were recommended did not reflect my diverse taste in movies.	-0.840
	The movie recommender seemed to target only a small subset of my interests.	-0.868
	The movie recommender treated me as a one-dimensional person.	-0.779
	The movie recommender seemed to stereotype me in a particular category of viewers.	-0.768

6.4.3.1 Subjective Experience

The SEM model (as shown in Figure 6.5) shows significant interaction effects from the interactivity, diversification, and visualization manipulations on the dependent variables (the 10 SSA and EXP factors). Particularly, the three experimental manipulations have a direct significant interaction effect on the participants' perceived interactivity of the system ($p = 0.035$): for participants who were not shown the visualization of emotional signatures, when they received the traditional top N recommendations, there was no significant difference in perceived interactivity of the system between participants who were able to specify their emotion preferences on movies vs. not being able to specify their emotion preferences. However, among participants who received the diverse N recommendations, those who were being able to specify their emotion preferences perceived significantly higher interactivity (a large, significant effect; Cohen's $d = 1.13$, $p < 0.001$) than participants

Table 6.7: Continuing table 6.6.

Considered aspects	Items	Factor loadings
Taste clarification potential (CLAR) AVE: 0.908 Cronbach's α: 0.95	Thanks to the movie recommender, I now know what kinds of movies I like.	0.912
	After using the movie recommender, I have no better idea about what different types of movies I like.	
	The movie recommender made me more uncertain about my own preferences.	
	The movie recommender helped me understand what kind of movies I like.	0.958
	Thanks to the movie recommender, I can now better express my preferences in terms of movies.	0.972
	The movie recommender helped me figure out what kind of movies I like.	0.969
Choice difficulty (DIFF) AVE: 0.584 Cronbach's α: 0.76	I was in doubt between several movies on the list.	
	I changed my mind several times before making a decision.	
	The task of making a decision was overwhelming.	0.793
	It was easy to select a movie.	-0.822
	Comparing the movies took a lot of effort.	0.669
System satisfaction (SysSAT) AVE: 0.738 Cronbach's α: 0.89	I would recommend the system to others.	
	The system is useless.	
	The system makes me more aware of my choice options.	0.840
	I make better choices with the system.	0.888
	I can find better items using the recommender system.	0.819
	Using the system is a pleasant experience.	
	The system has no real benefit for me.	-0.888
Choice satisfaction (ChoSAT) AVE: 0.692 Cronbach's α: 0.89	I like the movie I've chosen from the final recommendation list.	0.929
	The chosen movie fits my preference.	0.903
	I would recommend my chosen movie to others/friends.	0.899
	I was excited about my chosen movie.	0.877
	I think I chose the best movie from the options.	0.619
	I know several items that are better than the one I selected.	
	I would rather watch a different movie from the one I selected.	-0.715
Self-actualization (ACTL) AVE: 0.886 Cronbach's α: 0.93	The movie recommender taught me something about myself.	
	The movie recommender helped me get a new perspective on life.	0.928
	The movie recommender helped me reflect on who I want to be.	0.941
	The movie recommender helped me reflect on who I am as a person.	0.954
	The movie recommender would improve my quality of life.	

Table 6.8: Factor-fit metrics. Off-diagonal values are correlations, diagonal values are the square roots of the Average Variance Extracted $\left(\sqrt{AVE}\right)$ per factor.

	NOV	MOVE	VizF	INT	UND	DIV	QUAL	COVE	CLAR	DIFF	SysSAT	ChoSAT	ACTL
NOV	**0.718**												
MOVE	0.513	**0.869**											
VizF	0.147	0.157	**0.885**										
INT	0.330	0.247	-0.015	**0.805**									
UND	0.289	0.204	-0.043	0.647	**0.878**								
DIV	0.235	0.140	0.173	0.375	0.278	**0.917**							
QUAL	0.337	0.291	0.022	0.659	0.371	0.388	**0.949**						
COVE	0.247	0.156	0.009	0.577	0.355	0.688	0.549	**0.825**					
CLAR	0.306	0.097	0.146	0.591	0.319	0.488	0.532	0.570	**0.953**				
DIFF	-0.007	-0.145	0.081	-0.224	-0.245	-0.167	-0.297	-0.263	-0.015	**0.764**			
SysSAT	0.431	0.189	0.058	0.601	0.337	0.410	0.607	0.588	0.790	0	**0.859**		
ChoSAT	0.326	0.255	0.020	0.485	0.301	0.371	0.776	0.468	0.395	-0.41	0.435	**0.832**	
ACTL	0.289	0.097	0.223	0.351	0.167	0.328	0.281	0.301	0.677	0.223	0.639	0.201	**0.941**

who were not being able to specify their emotion preferences. For participants who were shown the visualization of the emotional signatures of movies, when they received the diversified recommendations, there was no significant difference in perceived interactivity of the system between participants being able to vs. not being able to specify their emotion preferences; while among participants who received the traditional top N recommendations, those who were able to specify their emotion prefer-

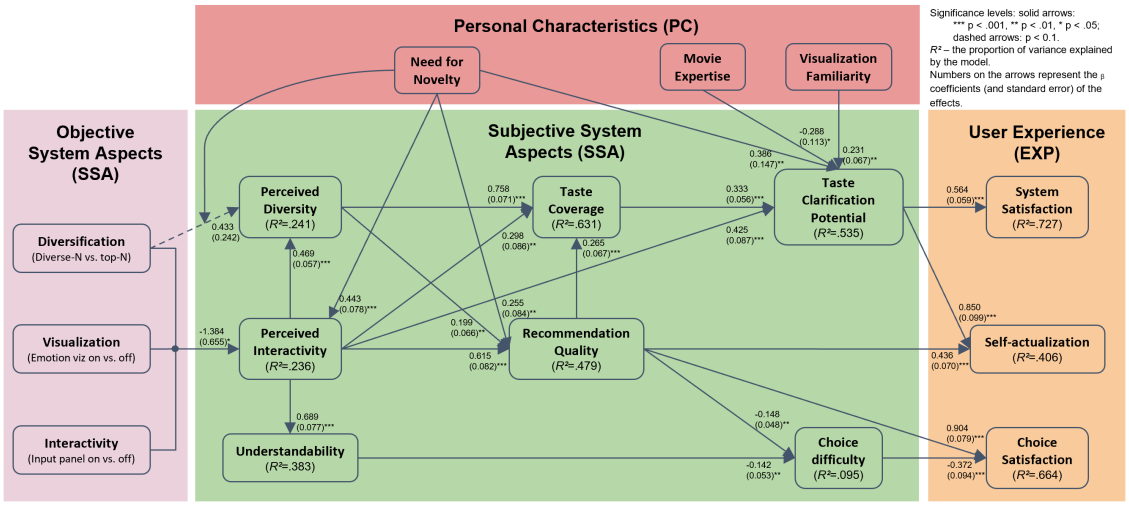

Figure 6.5: The resulting structural equation model (SEM) model of study IV.

ences on movies perceived a higher level of interactivity (a large, significant effect; Cohen's $d = 0.77$, $p < 0.001$) than participants who were not being able to specify their emotion preferences.

This interaction effect among the three experimental manipulations is displayed in Figure 6.6.

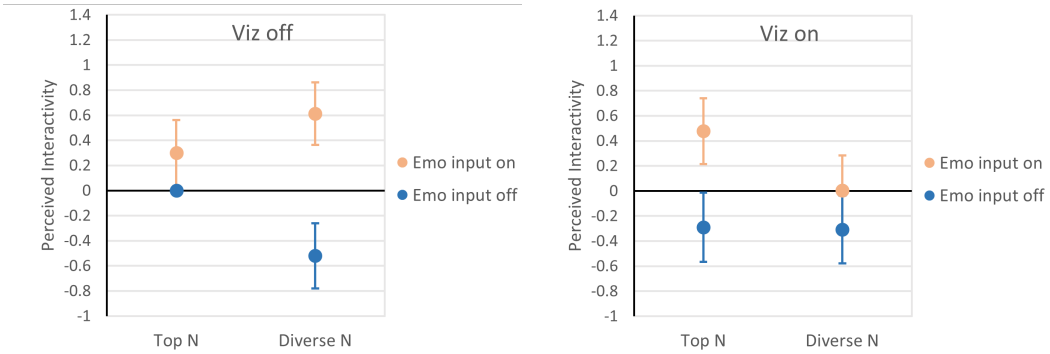

Figure 6.6: The marginal effects of interactivity (emotion input on/off), diversification (top-N/diverse-N), and visualization (Viz on/off) on participants' perceived interactivity of the system, the effect of the "Diverse N" condition with neither the emotion input ("Emo input off") nor the visualization of emotion signature ("Viz off") is set to zero, and the y-axis is scaled by the sample standard error.

Other than the interaction effects of the manipulations on perceived interactivity, the diversification manipulation also has a marginally significant effect on participants perceived diversity of the recommendations ($p = 0.074$): participants who received the diverse N recommendations scored 0.433 standard deviation higher on perceived diversity than participants who received the traditional

top N recommendations.

Perceived interactivity mediates the interaction effects on participants' understandability of the system, on participants' perceived diversity of the recommendations, on participants' perceived quality of the recommendations, on participants' perception of taste coverage, and on participants' perception of taste clarification potential. Understandability is in turn related to participants' perception of choice difficulty. Participants' perceived diversity of the recommendations has a significant positive effect on their perceptions of recommendation quality and taste coverage. The perceived recommendation quality is further significantly related to participants' perception of choice difficulty; participants' increased taste coverage is significantly related to their perception of taste clarification potential. Finally, participants' perception of taste clarification potential is significantly related to participants' satisfaction with the system; the perceived recommendation quality and the perceived choice difficulty are significantly related to participants' satisfaction with the recommendations; the perceived recommendation quality and the perceived taste clarification potential have significant effects on participants' perception of self-actualization.

6.4.3.2 Personal Characteristics

With the resulting SEM model presented in 6.5, I found that participants' personal characteristics play a role in participants subjective perceptions of the interaction with the system (Study IV - RQ4).

Participants with a 1 standard deviation higher need for novelty have an 0.443 standard deviation higher perceived interactivity and an 0.255 standard deviation higher on the recommendation quality, and 0.386 standard deviation higher on taste coverage. Participants' need for novelty has a moderated effect on the effect of the diversification manipulation on their perceived diversity of the recommendations (see Figure 6.7(a): for participants with higher need for novelty, there is no significant difference in perceived diversity between participants who received the traditional top N recommendations vs. the diverse N recommendations. However, for participants with lower need for novelty, those who were presented the diverse N recommendations perceived higher diversity than participants who received the traditional top N recommendations. Digging into this interesting interaction effect on participants perceived diversity, one possible reason is that the perceived diversity might depends on the objective diversity of the initial recommendations and/or final recommendations, some participants in the top N conditions might get more diverse recommendations

if they had already rated more diverse movies on the rating pages. To verify this possible reason, I calculated the objective diversity (i.e., the AESR metric, more details in Section 6.4.1) of both the initial recommendations and final recommendations (these two list of recommendations are the same for participants in the conditions without the emotion preference control panel). Surprisingly, it is not the case that participants with higher level of need for novelty received recommendations with higher objective diversity in the top N conditions, as shown in Figure 6.7. The other possible explanation for this interesting interaction effect is participants perception of diversity. It could be just that people with higher need for novelty perceived higher diversity in general; it could also be that people judge the diversity of the recommendations — especially if the recommendations that are not artificially diversified (i.e., the top N recommendations) — differently depending on how they inspected the details of the recommendations (i.e., the poster, synopsis, and/or the visualized emotion signature of the recommended movies).

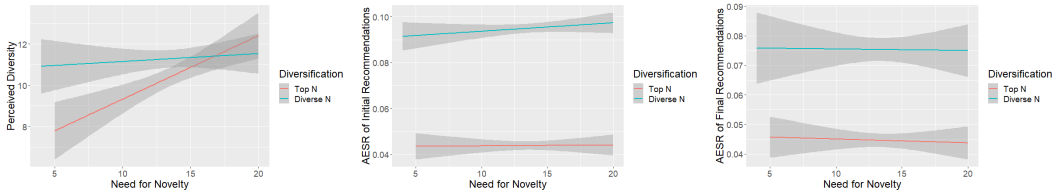

Figure 6.7: The interaction effect of diversification algorithm and need for novelty on participants' perceived diversity (left), objective diversity of the initial recommendations(middle), and objective diversity of the final recommendations (right).

Participants with higher movie expertise perceived lower taste clarification potential (unsurprisingly, because movie experts already know their tastes on movies), while participants' visualization familiarity contributes to their perception of taste clarification potential: participants with 1 standard higher deviation higher visualization familiarity scored 0.231 standard deviation higher on taste clarification potential. These effects are displayed in Figure 6.5

6.5 Discussion

In this study, introducing emotional signature for diversification(Study IV- RQ1), visualization (Study IV- RQ2), and interactivity (Study IV- RQ3) has a significant effect on the user experience, primarily because the significant three-way interaction effect of the three manipulations on participants' perceived interactivity and the marginally significant effect of diversification on

participants' perceived diversity. Perceived interactivity mediates the three-way interaction effects of the manipulations on participants' perceptions of diversity, understandability, recommendation quality, taste coverage, taste clarification potential, and choice difficulty; these outcomes in turn leads to a significant higher satisfaction with both the recommendations and the system, as well as contributing to participants' perception of self-actualization.

6.5.1 Incorporating Emotion as an Item Feature

The improvement in user experience in this study can be attributed to to the introduction of emotions as item feature (Study IV- RQ1). Emotions play a significant role in shaping users' preferences and reactions to content. By introducing emotion as an item feature for diversification, the recommender can recommend items that respond to users' diverse emotional needs. This emotional resonance can lead to a deeper connection between users and recommended items, resulting in higher engagement and satisfaction. In the meanwhile, visualizing emotions of items provides users with a contextual understanding of the emotional tone with each item. By visually representing emotions, users can quickly grasp the emotional attributes of the recommendations, allowing for better comprehension and interpretation of the recommendations.

Emotions are highly subjective and can vary greatly from person to person. By considering emotion as an item feature and thus allowing users to specify their unique tastes on movies from the perspective of emotion, the recommender can tailor recommendations to the emotion preferences of individual users in this study. This personalization allows users to discover content that aligns with their unique emotional needs and enhances their overall viewing or consumption experience, which supports the claim that taking emotions as an user-centric information would be a promising research for personalization [235]. On the other hand, emotions evoked from item reviews are often closely tied to uesrs' mood states, and users' preferences can vary depending on their current mood. By using emotions for interactivity, the recommender can take into account users' mood-based preferences and suggest content that matches their desired emotional state at a given time. This capability adds a new dimension to recommendation personalization, which in turn enhances users' satisfaction with the system.

6.5.2 Benefits of Interactivity

One of the innovations of this study is the introduction of the interactivity element: allowing users to actively interact with the recommendations by specifying their emotion preferences to the system and receive responses (updated recommendations) from the system immediately. Even though the overall effects of the three manipulations on user experience, primarily on perceived interactivity, also depend on the visualization and diversification as reported in section 6.4.3.1, the interactivity manipulation plays a leading role in this interaction effects, because among the three manipulations, allowing users to interact with the recommendations significantly increases ($t(274) = 5.248$, $p < .001$) participants' perceived interactivity with the system, while no significant main effects on perceived interactivity from diversification and visualization were found.

I hereby argue that introducing interactivity into recommender systems can provide benefits in personalization, engagement, and adaptability. Firstly, the interactivity feature allows users to provide explicit preferences as input, the system takes the input and updates the personalized recommendations accordingly, which increases the relevance and accuracy of the recommendations, thus leading to higher user satisfaction. In the meanwhile, the interactivity feature encourages users to actively participate in the recommendation process by tweaking the emotions, making them feel more involved and in control over the system. Users can both express their preferences and refine their recommendations. This engagement improves the overall user experience and strengthens the user-system interaction. Thirdly, the interactivity feature enables the recommender system to adapt and learn from user input over time (taking into account both the rating inputs and the emotion preference inputs of this study). By taking these inputs, the system can continuously refine its recommendations, adapting to users' evolving preferences and ensuring that the recommendations remain relevant and up to date.

6.5.3 Self-actualization

In the resulting SEM model of this study (see Figure 6.5), we note that the effects of using emotions for diversification, visualization, and interactivity on participants' perception of recommendation quality and taste clarification potential, which are mediated by perceived interactivity, perceived diversity, and/or taste coverage, finally determine their perception of self-actualization. This finding is in line with our original intention of creating "recommender systems

for self-actualization" [124, 91] – systems that follow a more holistic human-centered personalization practice by supporting users in developing, exploring, and understanding their unique tastes and preferences [124].

Introducing emotions for diversification help users explore a broader range of items beyond their usual preferences, expose them to a variety of items that they might not have been recommended otherwise, which enables users to discover items that aligns with their emotional interests and supports their self-discovery process, which in turn prompts them to clarify their preferences and refine their taste by discovering new and unexpected content. In addition, visualizing emotional signatures presents the recommendations in a intuitive manner, helps users to understand, evaluate, and compare the recommendations effectively. With enhanced comprehensibility via visibility, users can make more informed decisions based on their individual emotion preferences, resulting in improved recommendation quality; in the meanwhile, it enables users to articulate their preferences regarding the emotional aspects of items, allowing them to fine-tune their taste according to their own criteria and interests related to item emotions. On the other hand, by allowing users to actively engage with the recommender system and explicitly express their unique tastes on item emotions, the system can better understand users' preferences and refine the recommendations accordingly. This iterative interacting loop leads to higher recommendation quality since the system learns from the interactions and hereby adapts to their evolving tastes. Additively, this process encourages users to reflect on their tastes and preferences, helping them gain a deeper understanding of their own preferences and further develop and shape their self-identity.

The combination of interactivity, diversification, and visualization in this study does not show effects in an additive manner: visualizing emotional signatures help users perceive higher interactivity (Study IV- RQ2), but only for the traditional top N recommendations; allowing users to actively indicate their emotion preferences on items also helps on the perceived interactivity (Study IV- RQ3) without the visualization of the emotional signatures. The possible reason is that the visualization of the emotional signature was designed in an item-by-item manner, lacking the ability to show the spread of emotion values (on each dimension) across the whole recommendation lists; this design limitation in the visibility of the overall emotions spread in turn negatively impact users' subjective perceptions of the system.

By combining interactivity, diversification, and more proper visualization (which requires a more careful design), recommender systems would empower users to actively engage with the

system, explore diverse content, and gain insights into their preferences. This process supports self-actualization by enabling users to polish their taste, clarify their preferences, and ultimately curate a personalized content experience that aligns with their individuality and self-expression.

6.6 Conclusion, Limitations, and Future Work

This final study was motivated by my existing works of eXplanation-driven Interpretable Machine Learning (XIML) (presented in Chapter 3), the effect of recommendation source and justification method on professional development recommendations (presented in Chapter 4), and alternative recommendations beyond the traditional top-N recommendations (presented in Chapter 5). This study highlights the following aspects: a) combining visualization and interactivity on a more complex machine learning task, which takes a step further based on the findings from the existing work of chapter 3; b) diversifying the traditional top-N recommendation by the emotion feature of movies, which provides a potential way that supports users in exploring and understanding their unique personal tastes from the perspective of the evoked emotions.

The results show that introducing emotion as an item feature into recommender systems does help in personalization and individual taste exploration; this benefits are greatly optimized through the mechanisms that diverse recommendations by emotional signature, visualize recommendations on the emotional signature, and allow users to directly interact with the system by tweaking their tastes, which further contributes to both user experience and self-actualization.

Based on the potential findings explored in this work, designers, developers, and practitioners could consider how to integrate the explanation and interactivity features into recommender systems for better user experience (Overall RQ1 and Overall RQ2) and supporting taste exploration. Additionally, this work offers additional contributions to the limited literature on how the idea of interactive explainable machine learning systems can be applied in the context of recommender systems with the diversification algorithm implemented (Overall RQ3) to improve users subjective experience as well as supporting self-actualization.

One downside of this work is that I didn't record users interaction behaviors when they were interacting with the emotion preference panel, future work could investigate more into the details of users' interaction process, such as which emotions users are more interested in tweaking, which emotions users are prefer to have high values on; there could be some interesting findings of how users

react on the item emotions, and probably there is a possibility to come up with *emotional preference signatures* from the investigation as a new user feature that can be applied in recommender systems.

Chapter 7

Discussion over All the Four Studies

7.1 Revisiting the Overall Research Questions

In this dissertation, I conducted four studies to investigate the effects of explanation (Overall RQ2) and interactivity (Overall RQ1) on user experiences in adaptive experiences as well as how these effects depend on personal and situational context (Overall RQ3).

The Effect of Interactivity (Overall RQ1). In light of the Overall RQ1, allowing users to edit the rules in Study I (Chapter 3) and implicitly indicate their emotion preferences to get updated recommendations in Study IV (Chapter 6) have significant effects on user experiences, however, allowing users to interact with the alternative recommendation lists in Study III (Chapter 5) doesn't contribute to better user experience.

The Effect of Explanation (Overall RQ2). In light of the Overall RQ2, the grid-visualized explanation in Study I (Chapter 3) does not have a significant main effect on the subjective system aspects compared to the textual explanation. The proposed alternative recommendation lists in Study III (Chapter 5) which reflect how much else does the system learn about users beyond the traditional top-N recommendations doesn't create a positive user experience. The provided justification of recommendations in Study II (Chapter 4) has a significant effect on user experience, but this effect differs in the sources that the recommendations were originated from. The effect of

the visualization in Study IV (Chapter 6) also depends on the other two experimental manipulations (i.e., interactivity and diversification).

How does the above effects of interactivity and explanation depend on personal and situational context(Overall RQ3)? Study I (Chapter 3) shows that participants whose education was limited to high school perceived less control during their interaction with the system if they were shown the grid-based explanation style. In Study II (Chapter 4), the results show that an interest-based justification outperforms a needs-based justification, but only for users who are told that the recommendations originate from an AI algorithm rather than a human expert. In Study III (Chapter 5), participants with higher movie expertise perceived higher recommendation quality and choice satisfaction. Participants with higher need for novelty perceived higher in recommendation quality, recommendation conformity, and system satisfaction. Particularly, the effect of need for novelty on participants' perceived recommendation quality is significantly less stronger for participants received "More things you will like" and "Things we have no clue about" compared to participants who received only the single top-N list (see Figure 5.7). In Study IV (Chapter 6), the effect of the "interactivity" manipulation on participants perceived interactivity depends on the nature of diversification: participants who were allowed to specify their emotion preferences perceived significantly higher interactivity when they were shown the diversified items. Participants with lower need for novelty perceived higher diversity on the diversified recommendations vs. on the traditional top N recommendations, but participants with higher need for novelty perceived similarity on both the traditional top N recommendations and the diversified recommendations; participants with higher movie expertise perceived lower taste clarification potential; participants' visualization familiarity contributes to their perception of taste clarification potential.

7.2 Interactivity

I note that allowing direct control over the system results in improved user experiences. In Study I (Chapter 3), the interactivity manipulation influence users' experience with the system by encourage users to engage in a mutual feedback loop that helps improve the system's performance. Participants perceived higher level of control primarily because they were allowed to provide feedback on the system provided rules, which in turn increases their perceptions of the feedback quality and thus results in a higher system satisfaction. In Study IV (Chapter 6), allowing users to explicitly

indicate their emotion preferences significantly increases participants' perceived interactivity of the design, and thereby contributes to a positive user experience with the system. However, allowing for the indirect control through interaction with the alternative recommendation lists (against the traditional top-N recommendations) leads to a negative user experience (see study III in Chapter 5).

I argue interacting directly with adaptive systems creates a more engaging and interactive user experience. This aligns with Moggridge's [168] claim on the importance of direct interaction in creating engaging user experiences. Introducing control by allowing direct interaction (cf. [168]), such as providing feedback and specifying preferences, users become active participants in the predicting/recommendation process, stimulating their curiosity and encouraging them to be more engaged in the process. Additionally, through this direct interaction, users can learn more about the rationality of the system, their own preferences, discover new items, and gain insights into their interests, fostering a sense of discovery and self-improvement.

7.3 Explanation

The effect of explanation on the user experience depends on how the explanation was designed and implemented. The intuitive visualization (visualized explanation) element in Study I (Chapter 3) turned out not to be as effective as I expected, it works differently depending on users' personal characteristics, probably due to the simplicity and the relatively objective nature of the application(i.e., the Tic-Tac-Toe game). In Study II, the justification of recommendations plays a role in users' experiences, but it also depends on the recommendation source, overall, users are most satisfied with interest-based recommendations presented by an AI algorithm (demonstrated in Chapter 4). The visualization of the emotional signature overall helps on users understandability of the system, even though this main effect is marginally significant ($t(274) = 1.66$, $p = 0.097$), it performs best when participants received the top N recommendations and were allowed to interact with the recommender by specifying their unique tastes.

Thus, I advise future researchers to be cautious and not overestimate the value of explanation solutions, but instead carefully investigate a specific explanation solution in combination with other design elements in different fields. Research has indicated that explanations may not always lead to improved understanding or user outcomes [154, 177]. In some cases, excessive or

complex explanations can overwhelm users or lead to cognitive overload [197, 167].Cognitive load theory, proposed by John Sweller, suggests that instructional materials should be designed to manage the cognitive load imposed on learners [224]. This theory emphasizes the importance of carefully evaluating the value and impact of explanations.

7.4 Taste Clarification and Self-actualization

The traditional top N recommendations created by recommender systems that focus on accuracy inevitably ignore the items that the system thinks the user will not like, which creates a "filter bubble" that prevents users from discovering new and unknown areas of their own taste, and limits the diversity of presented content. The idea of recommender systems for self-actualization (RSSA) can help overcome the "filter bubble" problem by supporting rather than replace decision making so as to help users develop and express their preferences and focusing on exploration rather than consumption, which motivates the alternative recommendation lists (see Study III in Chapter 5), however, the proposed alternative lists do not contribute a better user experience. Instead of taking the taste Clarification and self-actualization as a motivation, introducing new item features that are reflective on users (such as emotion) for diversification in a recommender system (see Study IV in Chapter 6)) provides a potential solution to mitigate the "filter bubble" problem: participants who received the diverse N recommendations perceived significant higher diversity than participants who received the traditional top N recommendations, the perceived diversity in turn has a significant positive effect on participants' perceptions of taste coverage, which in turn increases participants' perceptions of taste clarification potential and self-actualization.

Arguably, being exposed to a wide range of content from the perspective a novel reflective item feature (such as emotion), users have the opportunity to discover new interests, expand their horizons, and challenge their existing preferences based on their experiences and reactions to the diverse content [254, 140]. This plays a causal role in the taste clarification potential. Combining diversification together with the interactivity, users are able to both actively engage with content and actively seek out new experiences, through which they acquire knowledge, develop new skills, and evolve their understanding of themselves and the world around them. This ongoing growth and learning process leads to a deeper understanding of personal preferences and contributes to taste clarification and self-actualization.

7.5 The Dual-route Approach Supporting both User Experience and Self-actualization

The idea of "Recommender Systems for Self-Actualization (RSSA)" provides a potential solution to the "filter bubble" problem [124], which aims not to optimize recommendation quality or user experience with the system, but instead cover the crucial process of users discovering their own unique tastes and preferences. This motivated me to come up with the alternative recommendation lists (i.e., RSSA features, see Study III in Chapter 5) that go beyond the traditional Top-N list (which purely concentrates on the algorithm accuracy) to keep the user "in-the-loop" with the goal of supporting rather than replace decision making so as to help users develop and express their preferences and focusing on exploration rather than consumption. The study results turn out that the alternative recommendation lists do not contribute to a better user experience with the system. This is not surprising, on one hand, users often have ingrained expectations: users expect a recommender system to give them good recommendations, not items that "make them think";. on the other hand, machine learning (including recommendation algorithms [184]) often remains a "black box" for the end-users, hiding its inner workings from its users due to the complexity of the algorithms; and thus in the design of study III (Chapter 5) we implicitly embedded the solutions motivated by self-actualization (the RSSA features) in the back-end algorithms; besides, in order to particularly examine the effect of those RSSA features (which serves an initial cautionary step in the investigation of "Recommender Systems for Self-Actualization (RSSA)"), we chose not to introduce the explanation element yet — which is able to potentially reveal the implicitly embedded essence of the RSSA features to the user — to avoid confounding effects. The

While in Study IV (Chapter 6), introducing emotions for diversification, visualization, and interactivity increases both user experiences and self-actualization. Arguably, compared to study III with implicitly embedded RSSA features for taste developing and thus for self-actualizing, the role of emotion in the algorithm is for diversification in study IV, similarly, it is also implicitly embedded in the back-end algorithm; however, other than the increased user engagement and understandability, the design of interactivity and visualization (visualized explanation) also plays a part in revealing users the novel attribute of items (i.e., emotion), which helps users to become reflective. Thus, the reasonable fusion of those elements (i.e., emotion, diversification, visualization, and interactivity) finally contributes to both user experience and self-actualization.

Research has demonstrated that when designing explanations and interactivity in recommender systems, it is important to take into account not only the specific application domains but also the individual differences among stakeholders [135, 179, 39, 160, 148]. Ribera and Lapedriza argue that, for domain experts, providing interactive visualizations allow the experts to lead the self-discovery [204]; while for lay users, providing briefer explanations as well as allowing users to select one argument that is most interesting to their case [204]. From this perspective, combining the findings of these four studies in this dissertation, I hereby argue that deliberate and thoughtful designs of explanation with the application domains and individual differences fully considered, along with appropriate direct interactivity, which are able to arouse user reflection or resonance, would potentially promote both user experience and user self-actualization in adaptive experiences.

7.6 Contribution in Recommender Systems

Bibliography

[1] How news feed works. *https://www.facebook.com/help/1155510281178725*, Accessed: 2022-05-03.

[2] Behnoush Abdollahi and Olfa Nasraoui. Transparency in fair machine learning: the case of explainable recommender systems. *Human and machine learning: visible, explainable, trustworthy and transparent*, pages 21–35, 2018.

[3] Mohaned Abu Dalffa, Bassem S Abu-Nasser, and Samy S Abu-Naser. Tic-tac-toe learning using artificial neural networks. 2019.

[4] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[5] Gediminas Adomavicius and YoungOk Kwon. Optimization-based approaches for maximizing aggregate recommendation diversity. *INFORMS Journal on Computing*, 26(2):351–369, 2014.

[6] Darius Afchar, Alessandro B Melchiorre, Markus Schedl, Romain Hennequin, Elena V Epure, and Manuel Moussallam. Explainability in music recommender systems. *arXiv preprint arXiv:2201.10528*, 2022.

[7] Elizabeth Aguirre, Dominik Mahr, Dhruv Grewal, Ko De Ruyter, and Martin Wetzels. Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness. *Journal of retailing*, 91(1):34–49, 2015.

[8] Mohammed Z Al-Taie and Seifedine Kadry. Visualization of explanations in recommender systems. *Journal of Advanced Management Science Vol*, 2(2):140–144, 2014.

[9] Joseph Alba, John Lynch, Barton Weitz, Chris Janiszewski, Richard Lutz, Alan Sawyer, and Stacy Wood. Interactive home shopping: consumer, retailer, and manufacturer incentives to participate in electronic marketplaces. *Journal of marketing*, 61(3):38–53, 1997.

[10] Mohammed F Alhamid, Majdi Rawashdeh, Haiwei Dong, M Anwar Hossain, and Abdulmotaleb El Saddik. Exploring latent preferences for context-aware personalized recommendation systems. *IEEE Transactions on Human-Machine Systems*, 46(4):615–623, 2016.

[11] Öznur Alkan, Dennis Wei, Massimiliano Matteti, Rahul Nair, Elizabeth M Daly, and Diptikalyan Saha. Frote: Feedback rule-driven oversampling for editing models. *arXiv preprint arXiv:2201.01070*, 2022.

[12] Jorge A Alvarado-Valencia and Lope H Barrero. Reliance, trust and heuristics in judgmental forecasting. *Computers in human behavior*, 36:102–113, 2014.

[13] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120, 2014.

[14] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. Investigating ad transparency mechanisms in social media: A case study of facebook's explanations. In *NDSS 2018-Network and Distributed System Security Symposium*, pages 1–15, 2018.

[15] Guy Aridor, Duarte Goncalves, and Shan Sikdar. Deconstructing the filter bubble: User decision-making and recommender systems. In *Fourteenth ACM Conference on Recommender Systems*, pages 82–91, 2020.

[16] Gaurav Arora, Ashish Kumar, Gitanjali Sanjay Devre, and Amit Ghumare. Movie recommendation system based on users' similarity. *International journal of computer science and mobile computing*, 3(4):765–770, 2014.

[17] Daniar Asanov et al. Algorithms and methods in recommender systems. *Berlin Institute of Technology, Berlin, Germany*, 2011.

[18] Amos Azaria, Avinatan Hassidim, Sarit Kraus, Adi Eshkol, Ofer Weintraub, and Irit Netanely. Movie recommender system for profit maximization. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 121–128, 2013.

[19] Nadeem Bader, Osnat Mokryn, and Joel Lanir. Exploring emotions in online movie reviews for online browsing. In *Proceedings of the 22nd international conference on intelligent user interfaces companion*, pages 35–38, 2017.

[20] Lisa Barnard. *The cost of creepiness: How online behavioral advertising affects consumer purchase intention*. PhD thesis, The University of North Carolina at Chapel Hill, 2014.

[21] Susanne Barth and Menno DT De Jong. The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. *Telematics and informatics*, 34(7):1038–1058, 2017.

[22] Nanci Bell and Phyllis Lindamood. *Visualizing and verbalizing: For language comprehension and thinking*. Academy of Reading Publications Paso Robles, CA, 1991.

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

[24] Benedikt Berger, Martin Adam, Alexander Rühr, and Alexander Benlian. Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, 63(1):55–68, 2021.

[25] Elisa Bertino, Shawn Merrill, Alina Nesen, and Christine Utz. Redefining data transparency: A multidimensional approach. *Computer*, 52(1):16–26, 2019.

[26] Émilie Bigras, Pierre-Majorique Léger, and Sylvain Sénécal. Recommendation agent adoption: how recommendation presentation influences employees' perceptions, behaviors, and decision quality. *Applied Sciences*, 9(20):4244, 2019.

[27] Mustafa Bilgic and Raymond J Mooney. Explaining recommendations: Satisfaction vs. promotion. In *Beyond personalization workshop, IUI*, volume 5, page 153, 2005.

[28] Daniel Billsus and Michael J Pazzani. A personal news agent that talks, learns and explains. In *Proceedings of the third annual conference on Autonomous Agents*, pages 268–275, 1999.

[29] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

[30] Dirk Bollen, Bart P Knijnenburg, Martijn C Willemsen, and Mark Graus. Understanding choice overload in recommender systems. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 63–70, 2010.

[31] Philip Bonhard, Clare Harries, John McCarthy, and M Angela Sasse. Accounting for taste: using profile similarity to improve recommender systems. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1057–1066, 2006.

[32] Hilda Borko and Carol Livingston. Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American educational research journal*, 26(4):473–498, 1989.

[33] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. Tasteweights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 35–42, 2012.

[34] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239, 2020.

[35] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, 2019.

[36] Josh Catone. Why web personalization may be damaging our world view. *Marshable*, Access: June 2011.

[37] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232, 2018.

[38] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. How serendipity improves user satisfaction with recommendations? a large-scale user evaluation. In *The world wide web conference*, pages 240–250, 2019.

[39] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 765–774, 2019.

[40] Li Chen Cheng and Ming-Chan Lin. A hybrid recommender system for the mining of consumer preferences from their reviews. *Journal of Information Science*, 46(5):664–682, 2020.

[41] Sung J Choi and M Eric Johnson. Understanding the relationship between data breaches and hospital advertising expenditures. *The American journal of managed care*, 25(1):e14–e20, 2019.

[42] Miki Cohen-Kalaf, Joel Lanir, Peter Bak, and Osnat Mokryn. Movie emotion map: an interactive tool for exploring movies according to their emotional signature. *Multimedia Tools and Applications*, 81(11):14663–14684, 2022.

[43] Fabio Colella, Pedram Daee, Jussi Jokinen, Antti Oulasvirta, and Samuel Kaski. Human strategic steering improves performance of interactive optimization. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 293–297, 2020.

[44] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction*, 18:455–496, 2008.

[45] Elizabeth M Daly, Massimiliano Mattetti, Öznur Alkan, and Rahul Nair. User driven model adjustment via boolean rule explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5896–5904, 2021.

[46] Linda Darling-Hammond, Maria E. Hyler, and Madelyn Gardner. *Effective Teacher Professional Development*. Learning Policy Institute, June 2017. ISSN: ISSN- Publication Title: Learning Policy Institute.

[47] Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. Boolean decision rules via column generation. *Advances in Neural Information Processing Systems*, 31:4655–4665, 2018.

[48] Edward L Deci, Haleh Eghrari, Brian C Patrick, and Dean R Leone. Facilitating internalization: The self-determination theory perspective. *Journal of personality*, 62(1):119–142, 1994.

[49] Edward L Deci and Richard M Ryan. The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry*, 11(4):227–268, 2000.

[50] Edward L Deci and Richard M Ryan. Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian psychology/Psychologie canadienne*, 49(3):182, 2008.

[51] Mustafa Demir, Nathan J. McNeese, and Nancy J. Cooke. Understanding human-robot teams in light of all-human teams: Aspects of team interaction and shared cognition. *International Journal of Human-Computer Studies*, 140:102436, 2020.

[52] Nic DePaula, Kaja J Fietkiewicz, Thomas J Froehlich, AJ Million, Isabelle Dorsch, and Aylin Ilhan. Challenges for social media: Misinformation, free speech, civic engagement, and data regulations. *Proceedings of the Association for Information Science and Technology*, 55(1):665–668, 2018.

[53] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pages 592–603, 2018.

[54] Jaap J Dijkstra, Wim BG Liebrand, and Ellen Timminga. Persuasiveness of expert systems. *Behaviour & Information Technology*, 17(3):155–163, 1998.

[55] Justin H Dion and Nicholas M Smith. Exploring private causes of action for victims of data breaches. *W. New Eng. L. Rev.*, 41:253, 2019.

[56] Wenjing Duan, Bin Gu, and Andrew B Whinston. Do online reviews matter?—an empirical investigation of panel data. *Decision support systems*, 45(4):1007–1016, 2008.

[57] Aidan Duane and Pat Finnegan. Managing empowerment and control in an intranet environment. *Information Systems Journal*, 13(2):133–158, 2003.

[58] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 2873–2882, 2015.

[59] Michael D Ekstrand. Lenskit for python: Next-generation software for recommender systems experiments. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2999–3006, 2020.

[60] Michael D Ekstrand, F Maxwell Harper, Martijn C Willemsen, and Joseph A Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 161–168, 2014.

[61] Michael D Ekstrand and Martijn C Willemsen. Behaviorism is not enough: better recommendations through listening to users. In *Proceedings of the 10th ACM conference on recommender systems*, pages 221–224, 2016.

[62] Aaron C Elkins, Norah E Dunbar, Bradley Adame, and Jay F Nunamaker. Are users threatened by credibility assessment systems? *Journal of Management Information Systems*, 29(4):249–262, 2013.

[63] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45, 2003.

[64] Ramon A Suarez Fernandez, Jose Luis Sanchez-Lopez, Carlos Sampedro, Hriday Bavle, Martin Molina, and Pascual Campoy. Natural user interfaces for human-drone multi-modal interaction. In *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 1013–1022. IEEE, 2016.

[65] Marina Fiori, Alessandra Lintas, Sarah Mesrobian, and Alessandro E. P. Villa. Effect of Emotion and Personality on Deviation from Purely Rational Decision-Making. In Tatiana V. Guy, Miroslav Karny, and David Wolpert, editors, *Decision Making and Imperfection*, volume 474, pages 129–161. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. Series Title: Studies in Computational Intelligence.

[66] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.

[67] Kenneth R Fleischmann and William A Wallace. A covenant with transparency: Opening the black box of models. *Communications of the ACM*, 48(5):93–97, 2005.

[68] Kristopher Floyd, Ryan Freling, Saad Alhoqail, Hyun Young Cho, and Traci Freling. How online product reviews affect retail sales: A meta-analysis. *Journal of retailing*, 90(2):217–232, 2014.

[69] Jodi Forlizzi and Katja Battarbee. Understanding experience in interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 261–268, 2004.

[70] APJ Anne Frederix. Investigating the effect of using personalized critiques in a decision-making tool.

[71] Alex A. Freitas. Comprehensible classification models – a position paper. *ACM SIGKDD Explorations*, 15(1):1–10, 2014.

[72] Gerhard Friedrich and Markus Zanker. A Taxonomy for Generating Explanations in Recommender Systems. *AI Magazine*, 32(3):90–98, June 2011.

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

[74] Alireza Gharahighehi and Celine Vens. Diversification in session-based news recommender systems. *Personal and Ubiquitous Computing*, pages 1–11, 2021.

[75] Alireza Gharahighehi and Celine Vens. Personalizing diversity versus accuracy in session-based recommender systems. *SN Computer Science*, 2(1):1–12, 2021.

[76] Giorgos Giannopoulos, George Papastefanatos, Dimitris Sacharidis, and Kostas Stefanidis. Interactivity, fairness and explanations in recommendations. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 157–161, 2021.

[77] Alexander Gilmanov. Here's why personalization algorithms are so efficient, Jun 2021.

[78] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[79] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236, 2008.

[80] R Gopinath. Prominence of self-actualization in organization. 2020.

[81] Kelly Goto. The era of adaptive experiences: Rethinking universal and inclusive design. *Design Management Review*, 30(3):28–33, 2019.

[82] Anne-Britt Gran, Peter Booth, and Taina Bucher. To be or not to be algorithm aware: a question of a new digital divide? *Information, Communication & Society*, 0(0):1–18, 2020.

[83] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Christopher Hall, and Tobias Höllerer. Smallworlds: visualizing social recommendations. In *Computer graphics forum*, volume 29, pages 833–842. Wiley Online Library, 2010.

[84] Ulrike Gretzel and Daniel R Fesenmaier. Persuasion in recommender systems. *International Journal of Electronic Commerce*, 11(2):81–100, 2006.

[85] James J Gross and Robert W Levenson. Emotion elicitation using films. *Cognition & emotion*, 9(1):87–108, 1995.

[86] Quentin Grossetti, Cedric du Mouza, Nicolas Travers, and Camelia Constantin. Reducing the filter bubble effect on twitter by considering communities for recommendations. *International Journal of Web Information Systems*, 2021.

[87] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[88] Siddharth Gulati, Sonia Sousa, and David Lamas. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10):1004–1015, 2019.

[89] Junius Gunaratne, Lior Zalmanson, and Oded Nov. The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems*, 35(4):1092–1120, 2018.

[90] David Gunning. Darpa's explainable artificial intelligence (xai) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page ii, New York, NY, USA, 2019. Association for Computing Machinery.

[91] Lijie Guo. Beyond the top-n: algorithms that generate recommendations for self-actualization. In *Proceedings of the 12th acm conference on recommender systems*, pages 573–577, 2018.

[92] Lijie Guo, Elizabeth M Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. Building trust in interactive machine learning via user contributed interpretable rules. In *27th International Conference on Intelligent User Interfaces*, pages 537–548, 2022.

[93] Lijie Guo, Christopher Flathmann, Reza Anaraky, Nathan McNeese, and Bart Knijnenburg. The effect of recommendation source and justification on professional development recommendations for high school teachers. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 175–185, 2022.

[94] Karthik S Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. Efficient data representation by selecting prototypes with importance weights. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 260–269. IEEE, 2019.

[95] Marc Hassenzahl. The thing and i: understanding the relationship between user and product. In *Funology 2*, pages 301–313. Springer, 2018.

[96] Chen He, Denis Parra, and Katrien Verbert. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56:9–27, 2016.

[97] Michael Hemphill. A note on adults' color–emotion associations. *The Journal of genetic psychology*, 157(3):275–280, 1996.

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

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

[100] Ai Thanh Ho, Ilusca LL Menezes, and Yousra Tagmouti. E-mrs: Emotion-based movie recommender system. In *Proceedings of IADIS e-Commerce Conference. USA: University of Washington Both-ell*, pages 1–8, 2006.

[101] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery.

[102] Fred Hohman, Minsuk Kahng, Robert S. Pienta, and Duen Horng Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25:2674–2693, 2019.

[103] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[104] Bryan Horlingm and Robby Bryant. Personalized search for everyone. *https://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html*, December 2009.

[105] Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. *Management Science*, 60(4):805–823, 2014.

[106] Li-tze Hu and Peter M Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1):1–55, 1999.

[107] Rong Hu and Pearl Pu. A comparative user study on rating vs. personality quiz based preference elicitation methods. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 367–372, 2009.

[108] Rong Hu and Pearl Pu. A study on user perception of personality-based recommender systems. In *International conference on user modeling, adaptation, and personalization*, pages 291–302. Springer, 2010.

[109] Yoel Inbar, Jeremy Cone, and Thomas Gilovich. People's intuitions about intuitive insight and intuitive choice. *Journal of personality and social psychology*, 99(2):232, 2010.

[110] Anthony Jameson, Bettina Berendt, Silvia Gabrielli, Federica Cena, Cristina Gena, Fabiana Vernero, Katharina Reinecke, et al. Choice architecture for human-computer interaction. *Foundations and Trends® in Human–Computer Interaction*, 7(1–2):1–235, 2014.

[111] Milena Janic, Jan Pieter Wijbenga, and Thijs Veugen. Transparency enhancing tools (tets): an overview. In *2013 Third Workshop on Socio-Technical Aspects in Security and Trust*, pages 18–25. IEEE, 2013.

[112] Yucheng Jin, Nava Tintarev, Nyi Nyi Htun, and Katrien Verbert. Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction*, 30(2):199–249, 2020.

[113] Leslie K John, Tami Kim, and Kate Barasz. Ads that don't overstep. *Harvard Business Review*, 96(1):62–69, 2018.

[114] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. 2020.

[115] Ashish Kapoor, Bongshin Lee, Desney Tan, and Eric Horvitz. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1343–1352, 2010.

[116] Sunil Karamchandani, Parth Gandhi, Omkar Pawar, and Shruti Pawaskar. A simple algorithm for designing an artificial intelligence based tic tac toe game. In *2015 International Conference on Pervasive Computing (ICPC)*, pages 1–4. IEEE, 2015.

[117] Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. Exploring the value of personality in predicting rating behaviors: a study of category preferences on movielens. In *Proceedings of the 10th ACM conference on recommender systems*, pages 139–142, 2016.

[118] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[119] Tami Kim, Kate Barasz, and Leslie K John. Why am i seeing this ad? the effect of ad transparency on ad effectiveness. *Journal of Consumer Research*, 45(5):906–932, 2019.

[120] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*, RecSys '12, pages 43–50, New York, NY, USA, 2012. ACM.

[121] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 43–50, 2012.

[122] Bart P. Knijnenburg, Nikhil Rao, and Alfred Kobsa. Experimental Materials Used in the Study on Inspectability and Control in Social Recommender Systems. Institute of Software Research, UC Irvine: Technical Report UCI-ISR-12-4, 2012.

[123] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 141–148, 2011.

[124] Bart P Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. Recommender systems for self-actualization. In *Proceedings of the 10th acm conference on recommender systems*, pages 11–14, 2016.

[125] Bart P. Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. Recommender Systems for Self-Actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems - RecSys '16*, pages 215–218, 2016.

[126] Bart P Knijnenburg and Martijn C Willemsen. Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. In *Proceedings of the third ACM conference on Recommender systems*, pages 381–384, 2009.

[127] Bart P Knijnenburg and Martijn C Willemsen. Evaluating recommender systems with user experiments. In *Recommender systems handbook*, pages 309–352. Springer, 2015.

[128] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.

[129] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. Explaining the user experience of recommender systems. *User Modelling and User-Adapted Interaction*, 22(4-5):441–504, 2012.

[130] Bart Piet Knijnenburg and Alfred Kobsa. Increasing sharing tendency without reducing satisfaction: Finding the best privacy-settings user interface for social networks. In *ICIS*, 2014.

[131] Sebastian Köhler, Thomas Wöhner, and Ralf Peters. The impact of consumer preferences on the accuracy of collaborative filtering recommender systems. *Electronic Markets*, 26(4):369–379, 2016.

[132] Joseph A Konstan and John Riedl. Recommender systems: from algorithms to user experience. *User modeling and user-adapted interaction*, 22(1):101–123, 2012.

[133] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[134] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. User preferences for hybrid explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, pages 84–88, 2017.

[135] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 379–390, 2019.

[136] Alex H Krist, Sebastian T Tong, Rebecca A Aycock, and Daniel R Longo. Engaging patients in decision-making and behavior change to promote prevention. *Information Services & Use*, 37(2):105–122, 2017.

[137] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems–a survey. *Knowledge-based systems*, 123:154–162, 2017.

[138] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1675–1684, New York, NY, USA, 2016. Association for Computing Machinery.

[139] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

[140] Lior Lansman, Osnat Mokryn, Lijie Guo, Mehtab Iqbal, and Bart P. Knijnenburg. Emotion diversification to improve user experience in movie recommender systems. *submitted to 16th ACM Conference on Recommender Systems*.

[141] CheonSol Lee, DongHee Han, Keejun Han, and Mun Yi. Improving graph-based movie recommender system using cinematic experience. *Applied Sciences*, 12(3):1493, 2022.

[142] Michael Leyer and Sabrina Schneider. Me, you or ai? how do we feel about delegation. 2019.

[143] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[144] Bin Liu, Anmol Sheth, Udi Weinsberg, Jaideep Chandrashekar, and Ramesh Govindan. Adreveal: Improving transparency into online targeted advertising. In *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, pages 1–7, 2013.

[145] George Loewenstein and Jennifer S. Lerner. The role of affect in decision making. 2003. ISBN: 0195377001 Publisher: Oxford University Press.

[146] Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

[147] Tania Lombrozo. Explanation and categorization: How "why?" informs "what?". *Cognition*, 110(2):248–253, 2009.

[148] Yichao Lu, Ruihai Dong, and Barry Smyth. Why i like it: multi-task learning for recommendation and explanation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 4–12, 2018.

[149] Gabriel Machado Lunardi, Guilherme Medeiros Machado, Vinicius Maran, and José Palazzo M de Oliveira. A metric for filter bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing*, 97:106771, 2020.

[150] Jan-Bert Maas, Paul C van Fenema, and Joseph Soeters. Erp system usage: The role of control and empowerment. *New Technology, Work and Employment*, 29(1):88–103, 2014.

[151] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. A grounded interaction protocol for explainable artificial intelligence. *arXiv preprint arXiv:1903.02409*, 2019.

[152] Frank H Mahnke. *Color, environment, and human response: an interdisciplinary understanding of color and its use as a beneficial element in the design of the architectural environment.* John Wiley & Sons, 1996.

[153] Jeff C. Marshall and Daniel M. Alston. Effective, Sustained Inquiry-Based Instruction Promotes Higher Science Proficiency Among All Groups: A 5-Year Analysis. *Journal of Science Teacher Education*, 25(7):807–821, November 2014.

[154] Millecamp Martijn, Cristina Conati, and Katrien Verbert. "knowing me, knowing you": personalized explanations for a music recommender system. *User Modeling and User-Adapted Interaction*, 32(1-2):215–252, 2022.

[155] F MARTIN. Top 10 lessons learned developing, deploying, and operating real-world recommender systems, rec-sys 2009 industry keynote. *http://blog. strands. com/2009/10/23/recsys-2009-keynote-top-10-lessons-learned-developing-deploying-and-operating-real-world-recommender-systems/*, 2009.

[156] Abraham Maslow. Self-actualization and beyond. 1965.

[157] Abraham Harold Maslow. A dynamic theory of human motivation. 1958.

[158] Maciej A Mazurowski. Estimating confidence of individual rating predictions in collaborative filtering recommender systems. *Expert Systems with Applications*, 40(10):3847–3857, 2013.

[159] Aleecia McDonald and Lorrie Faith Cranor. Beliefs and behaviors: Internet users' understanding of behavioral advertising. Tprc, 2010.

[160] James McInerney, Benjamin Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, and Rishabh Mehrotra. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM conference on recommender systems*, pages 31–39, 2018.

[161] Sean M McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K Lam, Al Mamunur Rashid, Joseph A Konstan, and John Riedl. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 116–125, 2002.

[162] Sean M McNee, Shyong K Lam, Joseph A Konstan, and John Riedl. Interfaces for eliciting new user preferences in recommender systems. In *International Conference on User Modeling*, pages 178–187. Springer, 2003.

[163] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101, 2006.

[164] Sean M McNee, John Riedl, and Joseph A Konstan. Making recommendations better: an analytic model for human-recommender interaction. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1103–1108, 2006.

[165] Michael D. McNeese and Nathaniel J. McNeese. Chapter 9 - humans interacting with intelligent machines: at the crossroads of symbiotic teamwork. In Richard Pak, Ewart J. de Visser, and Ericka Rovira, editors, *Living with Robots*, pages 165–197. Academic Press, 2020.

[166] Stuart E Middleton, Harith Alani, and David C De Roure. Exploiting synergy between ontologies and recommender systems. *arXiv preprint cs/0204012*, 2002.

[167] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 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*, pages 397–407, 2019.

[168] Bill Moggridge and Bill Atkinson. *Designing interactions*, volume 17. MIT press Cambridge, 2007.

[169] Osnat Mokryn, David Bodoff, Nadim Bader, Yael Albo, and Joel Lanir. Sharing emotions: determining films' evoked emotional experience from their online reviews. *Information Retrieval Journal*, 23(5):475–501, 2020.

[170] Osnat Mokryn, David Bodoff, Nadim Bader, Yael Albo, and Joel Lanir. Sharing emotions: determining films' evoked emotional experience from their online reviews. *Information Retrieval Journal*, 23(5):475–501, October 2020.

[171] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977, 2018.

[172] Alan L Montgomery and Michael D Smith. Prospects for personalization on the internet. *Journal of Interactive Marketing*, 23(2):130–137, 2009.

[173] Raymond J Mooney and Loriene Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204, 2000.

[174] Yashar Moshfeghi, Benjamin Piwowarski, and Joemon M Jose. Handling data sparsity in collaborative filtering using emotion and semantic based features. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 625–634, 2011.

[175] SM Mudambi and D Schuff. What makes a helpful review? a study of customer reviews on amazon. com (ssrn scholarly paper no. id 2175066). *Social Science Research Network, Rochester, NY*, 2010.

[176] Sayooran Nagulendra and Julita Vassileva. Understanding and Controlling the Filter Bubble through Interactive Visualization : A User Study. In *Proceedings of the 25th ACM conference on Hypertext and social media - HT '14*, pages 107–115, 2014.

[177] Mohammad Naiseh, Nan Jiang, Jianbing Ma, and Raian Ali. Personalising explainable recommendations: literature and conceptualisation. In *Trends and Innovations in Information Systems and Technologies: Volume 2 8*, pages 518–533. Springer, 2020.

[178] M. Narayanan, Emily Chen, Jeffrey He, Been Kim, S. Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *ArXiv*, abs/1902.00006, 2018.

[179] Sidra Naveed, Tim Donkers, and Jürgen Ziegler. Argumentation-based explanations in recommender systems: Conceptual framework and empirical results. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 293–298, 2018.

[180] NCSL. 2019 consumer data privacy legislation, March 2020.

[181] Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. Exploring mental models for transparent and controllable recommender systems: a qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 183–191, 2020.

[182] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686, 2014.

[183] Donald A Norman. *The psychology of everyday things.* Basic books, 1988.

[184] Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27:393–444, 2017.

[185] Heather L O'Brien and Elaine G Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology*, 59(6):938–955, 2008.

[186] Eli Pariser. Beware online "filter bubbles.". *https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles*, March 2011.

[187] Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

[188] Javier Parra-Arnau, Jagdish Prasad Achara, and Claude Castelluccia. Myadchoices: Bringing transparency and control to online advertising. *ACM Transactions on the Web (TWEB)*, 11(1):1–47, 2017.

[189] Ellen Peters et al. The functions of affect in the construction of preferences. *The construction of preference*, pages 454–463, 2006.

[190] Hans-Rüdiger Pfister and Gisela Böhm. The multiplicity of emotions: A framework of emotional functions in decision making. *Judgment and decision making*, 3(1):5, 2008. ISBN: 1930-2975 Publisher: Society for Judgment & Decision Making.

[191] Robert Plutchik. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier, 1980.

[192] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[193] Marianne Promberger and Jonathan Baron. Do patients trust computers? *Journal of Behavioral Decision Making*, 19(5):455–468, 2006.

[194] Andrew K Przybylski, Kou Murayama, Cody R DeHaan, and Valerie Gladwell. Motivational, emotional, and behavioral correlates of fear of missing out. *Computers in human behavior*, 29(4):1841–1848, 2013.

[195] Kristen Purcell, Lee Rainie, and Joanna Brenner. Search engine use 2012. 2012.

[196] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

[197] Alexander Renkl and Robert K Atkinson. Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational psychologist*, 38(1):15–22, 2003.

[198] Paul Resnick, R. Kelly Garrett, Travis Kriplean, Sean A. Munson, and Natalie Jomini Stroud. Bursting Your (Filter) Bubble : Strategies for Promoting Diverse Exposure. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work - CSCW '13*, pages 95–100, 2013.

[199] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186, 1994.

[200] Paul Resnick and Hal R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, March 1997.

[201] Paul Resnick and Hal R. Varian. Recommender Systems. *Communications of the ACM*, 40(3):56–58, 1997.

[202] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[203] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 18, pages 1527–1535, 2018.

[204] Mireia Ribera and Agata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. In *IUI workshops*, volume 2327, page 38, 2019.

[205] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530, 2007.

[206] Carl R. Rogers. Client-centered therapy, its current practice, implications, and theory. 1951.

[207] David Rogers. How business can gain consumers' trust around data, Nov 2015.

[208] Matthew Rowe. Semanticsvd++: incorporating semantic taste evolution for predicting ratings. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 213–220. IEEE, 2014.

[209] Ian Ruginski. Structural equation modeling in r tutorial 6: Confirmatory factor analysis using lavaan in r, October 2019.

[210] Koustuv Saha, Yozen Liu, Nicholas Vincent, Farhan Asif Chowdhury, Leonardo Neves, Neil Shah, and Maarten W Bos. Advertiming matters: Examining user ad consumption for effective ad allocations on social media. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2021.

[211] J Ben Schafer, Joseph A Konstan, and John Riedl. E-commerce recommendation applications. *Data mining and knowledge discovery*, 5(1):115–153, 2001.

[212] Markus Schedl, Peter Knees, Brian McFee, Dmitry Bogdanov, and Marius Kaminskas. Music recommender systems. In *Recommender systems handbook*, pages 453–492. Springer, 2015.

[213] Kamran Shafi and Hussein A Abbass. Biologically-inspired complex adaptive systems approaches to network intrusion detection. *information security technical report*, 12(4):209–217, 2007.

[214] Donghee Shin and Yong Jin Park. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, 98:277–284, 2019.

[215] Donghee Shin, Bouziane Zaid, and Mohammed Ibahrine. Algorithm appreciation: Algorithmic performance, developmental processes, and user interactions. In *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, pages 1–5. IEEE, 2020.

[216] Jaehyun Shin. Amazon personalize. *https://aws.amazon.com/personalize/*, Access: 2022-05-03.

[217] H Jeff Smith, Tamara Dinev, and Heng Xu. Information privacy research: an interdisciplinary review. *MIS quarterly*, pages 989–1015, 2011.

[218] Hyeonjin Soh, Leonard N Reid, and Karen Whitehill King. Measuring trust in advertising. *Journal of advertising*, 38(2):83–104, 2009.

[219] Changsoo Song and Jooho Lee. Citizens' use of social media in government, perceived transparency, and trust in government. *Public Performance & Management Review*, 39(2):430–453, 2016.

[220] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In *Conference on Fairness, Accountability and Transparency*, pages 5–19. PMLR, 2018.

[221] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009, 2009.

[222] Jianshan Sun, Jian Song, Yuanchun Jiang, Yezheng Liu, and Jun Li. Prick the filter bubble: A novel cross domain recommendation model with adaptive diversity regularization. *Electronic Markets*, pages 1–21, 2021.

[223] Kirsten Swearingen and Rashmi Sinha. Beyond algorithms: An hci perspective on recommender systems. In *ACM SIGIR 2001 workshop on recommender systems*, volume 13, pages 1–11. Citeseer, 2001.

[224] John Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier, 2011.

[225] P. Symeonidis, A Nanopoulos, and Y. Manolopoulos. Providing Justifications in Recommender Systems. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 38(6):1262–1272, November 2008.

[226] Kar Yan Tam and Shuk Ying Ho. Understanding the impact of web personalization on user information processing and decision outcomes. *MIS quarterly*, pages 865–890, 2006.

[227] Ed S Tan. *Emotion and the structure of narrative film: Film as an emotion machine*. Routledge, 2013.

[228] Muh-Chyun Tang and I-Han Liao. Preference diversity and openness to novelty: Scales construction from the perspective of movie recommendation. *Journal of the Association for Information Science and Technology*, 2022.

[229] Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245, 2019.

[230] Sharon Y Tettegah and Dorothy L Espelage. *Emotions, technology, and behaviors*. Academic Press, 2015.

[231] Nava Tintarev. Explanations of recommendations. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 203–206, 2007.

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

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

[234] Marko Tkalcic and Li Chen. Personality and recommender systems. In *Recommender systems handbook*, pages 715–739. Springer, 2015.

[235] Marko Tkalčič, Berardina De Carolis, Marco De Gemmis, Ante Odić, and Andrej Košir. Introduction to emotions and personality in personalized systems. *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, pages 3–11, 2016.

[236] Marko Tkalčič, Berardina De Carolis, Marco de Gemmis, Ante Odić, and Andrej Košir. Introduction to Emotions and Personality in Personalized Systems. In Marko Tkalčič, Berardina De Carolis, Marco de Gemmis, Ante Odić, and Andrej Košir, editors, *Emotions and Personality in Personalized Services*, pages 3–11. Springer International Publishing, Cham, 2016. Series Title: Human–Computer Interaction Series.

[237] Helma Torkamaan, Catalin-Mihai Barbu, and Jürgen Ziegler. How can they know that? a study of factors affecting the creepiness of recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 423–427, 2019.

[238] Jutta Treviranus and Stephen Hockema. The value of the unpopular: Counteracting the popularity echo-chamber on the web. In *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*, pages 603–608. IEEE, 2009.

[239] Daniel Tunkelang. Are ads really that bad?, Dec 2020.

[240] Joseph Turow, Jennifer King, Chris Jay Hoofnagle, Amy Bleakley, and Michael Hennessy. Americans reject tailored advertising and three activities that enable it. *Available at SSRN 1478214*, 2009.

[241] Rahat Ullah, Naveen Amblee, Wonjoon Kim, and Hyunjong Lee. From valence to emotions: Exploring the distribution of emotions in online product reviews. *Decision Support Systems*, 81:41–53, 2016.

[242] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *proceedings of the eighth symposium on usable privacy and security*, pages 1–15, 2012.

[243] M. Sree Vani. A recommender system for online advertising. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(2):599–604, Feb 2016.

[244] Jesse Vig, Shilad Sen, and John Riedl. Tagsplanations: explaining recommendations using tags. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 47–56, 2009.

[245] Serena Villata, Guido Boella, Dov M Gabbay, and Leendert Van Der Torre. A socio-cognitive model of trust using argumentation theory. *International Journal of Approximate Reasoning*, 54(4):541–559, 2013.

[246] Kristen L Walker. Surrendering information through the looking glass: Transparency, trust, and protection. *Journal of Public Policy & Marketing*, 35(1):144–158, 2016.

[247] Ruijie Wang, Reece Bush-Evans, Emily Arden-Close, Elvira Bolat, John McAlaney, Sarah Hodge, Sarah Thomas, and Keith Phalp. Transparency in persuasive technology, immersive technology, and online marketing: Facilitating users' informed decision making and practical implications. *Computers in Human Behavior*, page 107545, 2022.

[248] Weiquan Wang and Izak Benbasat. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, 23(4):217–246, 2007.

[249] Kangning Wei, Jinghua Huang, and Shaohong Fu. A survey of e-commerce recommender systems. In *2007 international conference on service systems and service management*, pages 1–5. IEEE, 2007.

[250] Joe Whittaker, Seán Looney, Alastair Reed, and Fabio Votta. Recommender systems and the amplification of extremist content. *Internet Policy Review*, 10(2):1–29, 2021.

[251] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

[252] Daricia Wilkinson, Moses Namara, Karishma Patil, Lijie Guo, Apoorva Manda, and Bart Knijnenburg. The pursuit of transparency and control: A classification of ad explanations in social media. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 763, 2021.

[253] Daricia Wilkinson, Saadhika Sivakumar, Pratitee Sinha, and Bart P Knijnenburg. Testing a Recommender System for Self-Actualization. In *2nd Workshop on Engineering Computer-Human Interaction in Recommender Systems (EnCHIReS 2017)*, pages 1–5, 2017.

[254] Martijn C Willemsen, Mark P Graus, and Bart P Knijnenburg. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction*, 26(4):347–389, 2016.

[255] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, pages 261–270, 2009.

[256] Chan Yun Yoo. Effects beyond click-through: Incidental exposure to web advertising. *Journal of Marketing Communications*, 15(4):227–246, 2009.

[257] Nadia Yusuf, Nisreen Al-Banawi, and Hajjah Abdel Rahman Al-Imam. The social media as echo chamber: The digital impact. *Journal of Business & Economics Research (JBER)*, 12(1):1–10, 2014.

[258] Brahim Zarouali, Koen Ponnet, Michel Walrave, and Karolien Poels. "do you like cookies?" adolescents' skeptical processing of retargeted facebook-ads and the moderating role of privacy concern and a textual debriefing. *Computers in Human Behavior*, 69:157–165, 2017.

[259] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. Bad news: Clickbait and deceptive ads on news and misinformation websites. In *Workshop on Technology and Consumer Protection (ConPro). IEEE, New York, NY*, 2020.

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