Interactive Explainable Artificial Intelligence in Movie Recommender Systems

Candidate Number GJYW8
HCI MSc Final Project Report 2020
UCL Interaction Centre, University College London
Supervisor: Dr. Enrico Costanza

ABSTRACT
Video on Demand platforms employ artificially intelligent movie recommender systems to support users in finding suitable content. However, prior work suggests that users still face several problems. Users reportedly feel overwhelmed by the high number of recommendations and find it hard to make sense of these suggestions and how these systems work. At the same time, current systems only employ few or no possibilities to control and make sense of the recommendation algorithms. This study investigates interactive explanations as an approach to improve sense-making and user-system trust. An initial study (based on 12 semi-structured interviews) examined the acceptability of expressive feedback on watched movies to derive highly personalised explanations. While this study found extensive feedback to be only moderately accepted because of high interaction effort, explanations of the eventual recommendations were found to be in high favour. Therefore, a follow-up within-subject online user study (N=30) compared three prototypes of interactive explanations (feature-weighting, timeline, and genre-swapping) with a non-interactive/non-explainable baseline. These prototypes used the participant’s Netflix viewing activity to make personalised recommendations. In particular, each prototype’s effect on different levels of sense-making, user-system trust, and other variables was measured. The results show that interactive explanations are an effective way to enable users to make sense of recommender systems and their output. Moreover, all interactive systems were found to improve user-system trust significantly. An overreaching demand for more control and increased transparency in movie recommender systems furthermore suggests that Video on Demand platforms should start to employ interactive explanations.

Author Keywords
Movie recommender systems; explainable artificial intelligence; Video on Demand platforms

CCS Concepts
• Human-centered computing → Ubiquitous computing;
Please use the 2012 Classifiers and see this link to embed them in the text: https://dl.acm.org/ccs/ccs_flat.cfm

MSc Contribution Type
Empirical

INTRODUCTION
The worldwide Video on Demand (VOD) market is booming: In 2020, nearly one billion users are streaming videos on various VOD Platforms like Netflix or Amazon Prime Video and can access tens of thousands of movies and television series [27, 61, 62]. Artificially intelligent movie recommender systems are integrated on these platforms to help users to discover suitable movies and series.

However, to date, users face a variety of problems with current movie recommender systems (MRS) [19, 45, 51]. The vast choice of movies being recommended to a user often results in choice overload, which can make choosing a movie an overwhelming and time-intensive task of indecisiveness [30, 51]. A phenomenon called the ‘Netflix Effect’ [45]. Furthermore, users might consider not all of these recommendations to be a match, and this state of choice overload is often accompanied by a lack of understanding of how these recommendations were formed [45]. Users reportedly find it hard to make sense of the suitability of the provided recommendations, they lack insight into how MRS calculated the recommendations, and they cannot influence these systems according to their needs [68, 44]. This lack of transparency can cause distrust in a system, and it makes these systems behave like a "black box" [56, 31, 8, 47]. Researchers argue, that current recommender systems of intelligent everyday applications often fail to adequately support users to make sense of and control recommender systems, and more attention must be paid to the user to explain these artificially intelligent systems and their output [44, 48, 19].

Initial research found interactive explanations a promising technique to control and make sense of recommender systems [55, 9, 47]. In parallel, Zürn et al. suggest investigating interfaces that allow users to ask “what if?” questions, as this could be an auspicious way to foster the understanding of an Artificial Intelligence (AI)’s decision-making process [68]. Past research has conceptualised various forms of interactivity, but to date, these interfaces are neither widely distributed nor were their effects on sense-making and user-system trust holistically investigated [68, 52, 44].

Against this background, this thesis took a two-fold approach to improve recommender systems of VOD platforms. Therefore, an initial study investigated whether users could potentially overcome these challenges through highly personalised
recommendations based on their input. This study based on semi-structured interviews and feedback on prototype videos (N=12 participants) shed light on the following research question:

**RQ1:** How acceptable is it to users to give extensive feedback on watched movies as a way to derive highly personalised explanations for movie recommendations?

Giving extensive feedback turned out to be only of mediocre acceptance, but the findings of the initial study suggested that explanations on the eventual recommendations are in high favour. Based on this, the main user study investigated different concepts of interactive explanations which were inspired by past research and the initial study. This empirical study was conducted to answer the following research question:

**RQ2:** How can interactive explanations improve user-system trust and sense-making in movie recommender systems?

In an online user study with 30 participants, three fully-functioning interactive explainable prototypes (based on feature-weighting, a timeline-approach, and genre-swapping) were compared to a non-interactive/non-explainable baseline. The underlying MRS used the participant’s Netflix viewing activity to show meaningful and personalised recommendations. In contrast to other studies [55, 52], basing recommendations on real user data allowed for more natural and valid interactions with the system. In particular, the prototypes’ effect on user-system trust, sense-making, and satisfaction was investigated together with their suitability for VOD portals.

The results of the study show that VOD platform users demand more control over movie recommender systems. Moreover, the evaluated concepts of interactive explanations significantly increased user-system trust and satisfaction and empowered users to make sense of the MRS and its recommendations. These findings suggest that VOD platforms can substantially improve their User Experience (UX) and yield a real “practical value [for] a more interactive approach to explanation” [19, p. 21]. This could help such media services to differentiate from competitors in the highly saturated VOD market [62].

This thesis is structured as follows: An in-depth literature review introduces the status quo and current challenges of MRS. Following this, related work in the field of explainable artificial intelligence in recommender systems is critically discussed and findings are derived for this study. Subsequently, the methodology and results of the initial qualitative study on RQ1 are presented together with design implications that inspire the prototypes of the main study. Based on that, the main study’s methodology and prototypes used to respond to RQ2 are described in detail. Then, the quantitative and qualitative results of the main study are reported. A subsequent discussion recaps the research questions and interprets the findings. Finally, a conclusion briefly lists the main findings and the study’s limitations are presented together with promising fields for future research.

**RELATED WORK**

This section lays the theoretical foundation for the empirical user study on interactive explainable artificial intelligence in movie recommender systems (MRS). After a brief introduction to MRS and their current challenges, explainable artificial intelligence (XAI) is discussed as one potential but promising solution. In particular, current research on the aspect of interactivity in XAI is critically reviewed.

**Movie Recommender Systems: Status Quo**

Recommender systems are an elementary component of e-commerce websites, consumer-faced applications, and expert systems to guide users through “a large space of possible options” [11, p. 331] through personalised recommendations. In the field of Video on Demand platforms, these systems suggest suitable movies and series based on a variety of data features. Distinguished by their knowledge source, MRS can be either classified as collaborative filtering-based (social and individual data), content-based, or hybrid recommender systems [12, 21, 59]. While research on recommender systems comprised a magnitude of directions, substantial attention has been drawn to algorithmic research to improve the accuracy of recommendations [12, 39]. Here, collaborative filtering (CF) recommender systems that base movie recommendations on user ratings and demographic information were of major research interest [16, 12]. This technique is highly accurate yet simple, but it is affected by the “cold start” problem [39, 12]. To approach the cold start problem, content-based (CB) algorithms that appropriate data features like keywords and genres became of great research interest [59, 63, 36]. This research area gained further momentum through the application of artificial intelligence (AI). Data mining and machine learning together with the growing availability of big (semantic) datasets and computing power have enabled researchers to improve content-based recommender systems in the past few years [36, 49]. While these content-based approaches were found to be faster, successfully overcame the cold start problem, and allowed serendipitous movie recommendations, collaborative-filtering algorithms remained superior in recommendation accuracy [39, 49].

To approach the shortcomings of both techniques, hybrid recommendation systems which utilise and combine various data sources became a robust MRS standard yielding higher coverage and accuracy [39, 9]. Current VOD platforms like Netflix, therefore, employ hybrid recommendation systems that allow nearly unlimited movie suggestions that match a user’s interests [60, 7]. This sheer number of highly-fitting recommendations provided by an AI, however, contribute to the current challenges users face on VOD platforms today – users reportedly feel overwhelmed by the number of recommendations – a phenomenon called the ‘Netflix Effect’ [45]. In this state of choice overload, it is hard to make sense of which titles are suitable and why certain movies were recommended as recommendations are often made without providing information on the knowledge source [47, 30]. This uncertainty and lack of transparency can lead users to distrust a system and cause what Jhaver et al. coined “algorithmic anxiety” [31, 8, 47, 56].

1New data items without previous ratings or new user profiles without existing preferences (cold start) rule collaborative filtering-based recommendations out [59] and large datasets cause data processing performance issues [39, 21]
Furthermore, the user experience may be harmed by the feeling of being in a “filter bubble”\(^2\). As specified within the DIN EN ISO 9241-11 “[usability is the] extent to which a system, product or service can be used by specific users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use” [29]. Hence, these issues can be considered counterproductive towards a movie recommender system’s usability if a user experiences difficulty (effectiveness, efficiency, and satisfaction) reaching their goal of finding movies of interest. This is a promising field for solutions based on explainable artificial intelligence.

**Explainable Artificial Intelligence**

With a growing prevalence of artificially intelligent algorithms in expert systems and “intelligent everyday applications” [19, p. 22] users benefit from robust predictions and classification. However, these systems often lack transparency and do not allow users to build an understanding of how these algorithms work and what data features influence their decision [4]. Accordingly, these systems are often considered to have a “black-box nature” leading to decreased user-system trust [56, 28]. The research domain of explainable artificial intelligence, therefore, has become of major importance and interest in recent years to make an AI’s reasoning more transparent and interpretable [4].

The transparency and interpretability of an AI’s functioning and its reasoning process are of utmost importance in safety-critical and high-risk applications in the medical domain, autonomous transportation, or algorithmic trading. It can be considered crucial in intelligent everyday applications too to allow users to build a mental model and gain insight into how their data is processed [38]. This effort became a legal requirement with the “right to explanation” through the EU’s General Data Protection Regulation [56, 1].

For several years, research focused mainly on improving the interpretability of machine learning models from a technical perspective to debug models [4]. These explanations aimed to inform experts through techniques like saliency maps, rule extraction, or decision trees. However, Miller has argued that such explanations are not suitable for keeping end-users of intelligent everyday applications in the loop [48]. This research on explainable AI in movie recommender systems, therefore, contributes to an understanding of methods used to produce explanations for non-expert users. This is crucial for improving the transparency of AI in MRS and guaranteeing the long-lasting adoption and user acceptance of AI in intelligent everyday applications [19, 48].

**Interactive Explainable Artificial Intelligence**

Explainable AI is a wide research area covering different techniques for explanations in various domains and looks into the effects on individuals using XAI-enabled systems [4]. The following section gives an introduction to the motivation of interactivity in XAI. Previous work is also examined. Furthermore, the constructs sense-making and user-system trust are discussed as their improvement is considered the main goal in XAI [8, 13, 18, 56].

Together with the discussion of the affinity for technology interaction as a key facet of user personality [40], these components form the methodological basis of this study which looks into the effects of interactivity in XAI in the field of MRS.

**Motivation for XAI in Movie Recommender Systems**

Previous Human-Computer Interaction studies shed light on different types of explanation techniques in recommender systems. Predominantly, researchers focused on static textual explanations as a way to explain recommendations [17, 37, 8]. These explanations help users to build an understanding through reading the respective explanation only. Other authors investigated the effectiveness of static and interactive visualisations as a way to explain recommendations. Here, the authors found that interactive visualisations can yield higher user satisfaction, recommendation accuracy, and better user experience [55, 9].

Contrary to these findings, Kouki et al. found no visual explanation to be more persuasive compared to textual explanations in an empirical study on three different visual explanations [37]. This extreme contrast is most likely due to a few reasons as follows: (1) Odonovan et al. and Bost et al. investigated interactive visualisations as a way to control the input of a system (=knowledge source). Whereas Kouki et al. only made the output (=recommendations) interactively explorable. (2) Visual explanations require users to have sufficient (data) visualisation literacy [47].

In line with several studies that show a great need for explanations in intelligent applications, current research points out that interactive explanations must be investigated as potential solutions too. Einband et al. analysed about 45,000 Google Play Store reviews of intelligent everyday applications and concluded that users reported a "desire for more control" and "[a] potential practical value of a more interactive approach to explanation" [19, p. 21]. Furthermore, they suppose that recommender systems in VOD platforms could benefit from interactive controls to influence and correct MRS algorithms.

**Interactive Explanations**

Whilst textual and non-interactive visual explanations involve the user only by looking and reading explanations, interactive explanations add a layer of active examination to the reasoning process. This active component is not only known to improve sense-making [68, 50], but it also allows users to adjust a system depending on their current mood or changed preferences over time [12].

Within this thesis, controlling algorithms through the following concepts based on previous research were investigated (for extensive descriptions of all prototypes see *MovieLand Prototypes*):

- **Timeline-Based Controls**
  
  Zürn et al. suggested allowing users to manipulate a recommender system’s input based on a time-axis [68]. They supposed that users could make sense by asking “what if?”
questions like “What if my recommendations would be based on my viewing history five years ago?”. While this thesis builds on the general idea of a timeline to explain, the approach was used in a slightly different way. Users are not only allowed to simulate what recommendations would look like at a certain time but can also include/exclude/combine individual movies that were watched over time.

- **Feature-Weighting Sliders**
  Nguyen et al. developed a set of software tools to interact with machine learning algorithms and presented feature-weighting sliders as a way to control algorithms in MRS as a usage example [52]. These sliders allowed users to influence the algorithm and see how their input influences recommendations. The presented work does not involve an empirical user study. Therefore, no assumptions about the applicability of feature-weighting sliders in MRS can be made, and no insights into how users might want to use these sliders were given. Similar to this idea, Bost and Höllerer used sliders in a music recommender system called TasteWeights to allow users to mix different knowledge sources [9]. Their findings showed that feature-weighting sliders were found to be a sufficient explanation technique in music recommender systems. These studies motivate the work at hand to investigate feature-weighting sliders in MRS in an empirical user study that must identify requirements and evaluate the resulting user interactions.

- **Mood/Genre-Based Clusters**
  Eiband et al. found that users expressed a need for recommendation clusters based on moods/genres [19]. These clusters explain themselves through semantic grouping. Building on the idea of personalized explanations by [37], this thesis investigates whether extensive feedback on watched movies can be used to derive highly personalised clusters. Retrieving extensive feedback on movies to derive explanations as tag-clouds has been done before by Chen et al. [15]. They explained that this approach is valued as users can specifically indicate what they like or dislike about a certain title. However, Chen et al.’s approach required users to tag movies based on free text input which involves high interaction effort. Therefore, this thesis’ initial study evaluated whether the interaction effort can be lowered through pre-generated tags derived from the meta-information of a movie. Moreover, the initial study looks into mood boards derived from these tags as a form of explanation.

**Sense-making and Reasoning**
Sense-making describes the cognitive process of constructing an understanding from a given situation [5]. Explanations allow users to reason about a recommender system’s way of functioning and help to build a mental model [68]. According to Hoffmann and Klein, XAI allows users to understand causality in different ways [25]. First, MRS users build an understanding of which data features influence recommendations on a global and local level and how the MRS works in general. Also, an understanding based on conditions is formed, and users can comprehend which circumstances would lead to a particular set of recommendations. This reasoning process involves asking “but why?” [48] and “what if?” [25] questions. Zürn et al. stress the point that current recommender systems still relatively often fail to answer these questions [68]. They propose interactivity (direct manipulation) of input features to allow users to investigate how certain changes affect the output of a recommender system. This not only contributes to a better understanding of a system but helps to establish user-system trust [18].

Despite improved sense-making being a major goal of XAI, only a few studies so far have investigated this construct in the context of certain explanations. However, there is a recent uptake in studies that (suggest to) investigate sense-making as a direct effect of explanations [42, 64, 22]. The absence of studies measuring sense-making may also be due to the lack of scales that quantify sense-making, which was often investigated from a qualitative perspective [5]. Building on that, Alsufiani et al. introduced a questionnaire to measure sense-making. Their instrument measures sense-making on the five subscales – “comprehension and insight, understanding connections, gap discovering and bridging, structuring, plus reducing confusion, uncertainty, and ambiguity” [5, p. 3]. Given the importance of sense-making and an abundance in comparative studies investigating it, this study focuses on interactive explanations and their effect on sense-making.

**User-System Trust**
User-system trust (trust in the following) has long been known to be critical for the adoption of information systems [66]. Explanations in AI-enabled systems help users to build trust in a system’s decisions and recommendations [56]. However, it has been argued that trust in information systems is a complex construct which consists of multiple factors such as competence, integrity, benevolence, and transparency/honesty [66, 32]. Interestingly, the general literature on XAI often concludes that explainable AI seemingly automatically fosters users trusting artificially intelligent applications [4, 24, 48].

However, actual studies on explanations with humans in the loop have shown that trust can be fostered but is no necessity. In a study on different forms of explanations for recommended movies, Berkovsky et al. have found, that explanations yielded higher trust in the system but differed greatly among different forms of explanations [8]. They suggested that user-system trust underlies a personality component. This idea motivated this study to investigate the effect of a user’s affinity for technology interaction (described hereafter).

Besides a personality component, trust needs may differ from domain to domain. Bussonne et al. conducted an explanatory study in which healthcare professionals used a clinical decision support system and tested several variations of explanations [13]. They found that extensive explanations led to increased user-system trust but at the same time led the staff to over-rely on the system’s decisions. These findings stand in contrast to Kizilcec’s study on varying explanation extents given on grades provided by an online system for peer assessment [35]. Here, users trusted a system less if their received score was lower than expected and at the same time provided more extensive explanations for the system’s results. Given that, the influence of explanations on trust must be further
investigated and inspires the work at hand to measure trust as a multi-factor construct as done by [8].

**Affinity for Technology Interaction**

As introduced earlier, the trust in certain explanations underlies the personality of different users. Therefore, personality traits make an important covariate to investigate when XAI-enabled systems are compared. Berkovsky et al. investigated the effect of personality on trust based on the “Big Five” personality traits [8]. Their findings suggested that different personality characteristics (e.g. neuroticism vs. agreeableness) favour different types of static movie recommendation explanations and ways to present these. However, this approach only allows assumptions on why different personality characteristics may lead to favouring certain explanations.

Further studies tried to investigate more definite personality traits like the affinity of technology interaction (ATI). In a study on music recommender systems, Millecamp et al. clustered their participants by tech-savviness while investigating interactive explanations in a music recommender system [47]. Unfortunately, they did not report back any results on the influence of tech-savviness and trust. Besides a potential effect of ATI on user-system trust, one’s affinity to interact with technology might be of particular interest when it comes to the perception of a particular system. Jin et al. found that more tech-savvy users perceive higher effort when searching for good songs using a chatbot-based recommender system [33]. Hence, investigating the affinity for technology interaction (“tech-savviness”) can be considered an important variable in XAI research.

In contrast to studies that researched tech-savviness based on fixed defined clusters [47], new findings suggest a less pre-determined approach. Franke et al. undertook extensive research on the affinity for technology interaction and found that ATI is a continuous variable and does not allow for artificial clustering [20]. They explain that this is because ATI precisely defines one’s likelihood to interact with technologies and does not suddenly change. This thesis, therefore, considered tech-savviness based on Franke et al.’s proposed Affinity for Technology Interaction scale and investigated the effect of ATI on how users perceived different types of (interactive) explanations.

**METHODOLOGY**

Building on the previously described related work, an initial study (semi-structured interviews) was conducted and informed the subsequent main study (empirical user study). The methodology of both studies is described in the following sections.

**Initial Study**

The following initial study was carried out to get a deep understanding of how movie recommender systems are currently used on VOD platforms. Furthermore, giving extensive feedback on movies was investigated to find out whether this could be an accepted way to generate highly-personalised explanations based on mood-modes inspired by the research of [37, 19]. The following section describes the methodology and discusses the results.

**Methodology**

The researcher chose to conduct a qualitative study utilising semi-structured interviews to obtain rich but at the same time comparable insights [14]. Next, users were introduced to the idea of giving more extensive feedback to get better recommendations and their views were obtained. Finally, the participants were asked to provide feedback on common and future recommender system prototypes. This user-centred approach was chosen to identify user needs and pain points and ultimately derive requirements for a novel recommender system that is investigated in the following main study.

**Participants**

The participants (N=12, see Figure 6) were recruited through personal contacts of the researcher. To get a broad range of insights, the sample consisted of a heterogeneous selection of participants: Six female participants and six male participants (M = 28.75 years, SD = 5.63 years) were actively sourced to resemble Netflix’s largest user group (people ageing between 25–34 years) [34]. Furthermore, the participants were chosen based on their different views on data privacy. Participants were not compensated for participation.

**Procedure and Data Analysis**

Each semi-structured interview lasted between 25–45 minutes, depending on the participant’s talkativeness. The interviews were conducted via audio or video calls, while participants were required to look at the different prototypes on a laptop/PC display. All participants got a short introduction to the research goal and were then asked to sign a consent form complying with GDPR and University College London’s ethic standards. After giving consent, the interviews were recorded and subsequently transcribed. 8 out of 12 interviews were held in German and subsequently translated to English using deepL3. After double-checking the transcripts for validity, a thematic analysis was conducted using Dovetail4 [10]. Therefore, initial codes were generated before the analysis as well as codes emerged progressively during the text analysis. These codes fit in the pre-defined main themes building on the semi-structured nature of the interviews. After coding the interviews, codes were refined, condensed, and grouped into matching sub-themes (see appx.).

**Results and Discussion**

The following sections reports on the main findings by theme and discusses the results.

**Using Video on Demand Platforms**

The first part of the interview shed light on the general usage of VOD platforms and how people find the content of their interest.

All users used multiple VOD platforms on a weekly basis while Netflix was predominant (11), followed by Amazon Prime Video (10). Other platforms (9) comprised BCC iPlayer (UK), TVNOW (Germany), and YouTube (4). This initial study, therefore, mainly reports on using Netflix and Amazon Video Prime. Reportedly, these platforms were used in parallel

3https://www.deepl.com/translator
4https://dovetailapp.com/
“because they offer different kinds of films and series” (P11) and enable users to “have a broader range” (P11). In terms of general usability and content offering, three participants emphasized Netflix’s advantage over Amazon Prime Video.

**Typical User Journey**

A broad consensus was found among the typical user journeys when searching for contents of interest on both platforms. Eleven participants reported that they start to browse for new content among their recommendations which are displayed on a platform’s start screen. This involves both personal recommendations based on a viewer’s history and more general recommendations such as trending content. This often leads users already to the final content they chose to watch. If the recommendations did not contain any suitable material, participants browsed the genre-based categories.

Half of the users felt stressed and overwhelmed (choice overload) by the amount of content which is offered to them. “Actually, we’re really searching for 20 minutes and don’t know which film to pick [...] [in those] 500 suggestions we have to choose one” (P5). A small fraction mentioned using the search function when having a particular title, actor, or director in mind. Content is chosen based on the thumbnail and meta information (title, genre, actors, length, etc.) while half of the participants indicated to be review-dependant: Here, external platforms such as IMDB, Google, or Rotten Tomatoes are used to gain insights into “others’ opinions” (P12) in the form of reviews and ratings and whether “a film is worth your time” (P8).

**Rating a Movie**

Only two participants actively used Netflix’s “thumbs up / thumbs down”-rating function to “influence the algorithm to show me more things that I’m interested in” (P1). Interestingly, about half of the participants had an idea of how an underlying algorithm could improve their recommendations by either learning from one’s watching history or their explicit ratings. For example, the low-frequency user P2 showed his understanding by reporting “we don’t watch enough [...] so the algorithm can’t be good enough”. However, current rating systems are not commonly used for several reasons: Five participants reported simply not knowing a built-in rating functionality exists, three had misconceptions about such rating functions, while two did not want to spend time on rating content, or they knew about it but still did not want to use it. The three latter reasons are partially based on the same misconception that ratings only help other people but not influence their own recommendations. P5 imagined Netflix’s rating system to be “[not] relevant for me [...] but for other people”. This misconception is closely tied to the participant’s prevalent usage of ratings/reviews to inform their own search for suitable content: About half of the participants made use of these indicators themselves but disliked providing this information to other users made one feel “totally selfish” (P2).

**Receiving Recommendations**

The interviews revealed great variance in terms of the subject’s satisfaction with their current personal recommendations. Three participants indicated to be fully satisfied with the recommendations, while, in contrast, three reported to be not satisfied with recommendations at all. Six participants experienced recommendations to be partly satisfying. Strikingly, all three of the participants who were not satisfied with the recommendations at all were the users who used shared profiles (see Figure 4) together with another person. Given that, MRS must satisfy the needs of people using shared profiles. Two of the three participants who were fully satisfied with the recommendations were “power users” and either used the rating system (P1) or consumed large amounts of media on VOD platforms. Participants that indicated to be only partially satisfied thought the recommendations were often too rough/broad or did not take a user’s need for varying content at different times into account. The latter turned out to be a salient theme and was self-reportedly described with the term “filter bubble” (P3, P4,
P10). More precisely, a filter bubble describes “A situation in which an Internet user encounters only information and opinions that conform to and reinforce their own beliefs, caused by algorithms that personalize an individual’s online experience.” [2]. Hence, MRS must allow for more fine-granular controls to influence recommendations and allow exploring diverse content.

Providing and Receiving Extensive Feedback
Within this part of the interview, participants were asked to put themselves into the situation of providing more feedback to a VOD platform about what they have liked or disliked about a movie. First, emerging themes on data privacy, rating criteria, and time investment in ratings are discussed. Secondly, the participant’s feedback on current and future recommender systems (input and output) which rely on more extensive feedback are presented.

Data Privacy
As providing richer feedback on what users liked or disliked about certain titles requires more data, participants were asked to share their views on data privacy in this context. The majority of them (8) reported having no serious concerns about data privacy, while one was slightly concerned. In contrast, three participants made reservations to their data privacy. The majority did not consider ratings as personal or sensitive data and therefore “don’t have a problem with that and [...] don’t have any concerns” (P5). It was often reported that “so much data is already collected everywhere [...] and it brings me an advantage” (P3) as recommendations would be improved. On the other side, concerns about data privacy were caused by one’s “general attitude” (P2) or a specific fear of “them selling [my data] to third parties” (P8).

Investing Time and Making Effort
Time is a scarce resource in general, which was also reflected by the participants. Therefore, three users indicated that they would not like to provide extensive feedback for time reasons or because “it’s too much work for me” (P6). The majority “would give feedback quickly” (P10) and would mostly only want to have “very, very little effort” (P5). However, “browsing Netflix for a very long time without even finding anything I would like to watch” (P12) motivates users to “invest a little more time in rating a movie to get better recommendations” (P7). Eight users mentioned seeing a ‘return on rating time investment’ which first requires more time, but finally leads to better recommendations (faster search) and less frustrating user experience. Few users mentioned, “if I liked the movie, I think I would spend more time rating the movie than if I didn’t like the movie at all” (P7).

Rating Criteria
The participants were asked to describe the qualities on which they assess good and bad movies to derive meaningful feedback categories. The following categories were most salient (descending): Cast (7), genre (7), ratings/reviews (6), individual dramaturgic/mood preferences (6), cinemagraphic style/quality (6), subject (3), director (3), length (2), language (1), music (1). It must be noted that a movie’s genre and the subject was often referred to in a mixed way, and the subject might be of higher importance than actively mentioned [41]. However, attention should be paid to individual preferences like “[no] excessive violence [and] a feminist representation of sex scenes” (P4), “happy ends” (P6), “realistic [plots]” (P4/5/10), and “[low] predictability” (P3/10).

Rating Systems
Within the interview, three animated video prototypes (see appx.) of different rating systems were presented to the participants, and their current experiences and opinions on these different input modalities were obtained.

Input: Rating Movies – Short Videos
- IP1 Like/dislike buttons: rating content by clicking thumbs up or thumbs down, e.g. Netflix
- IP2 10 stars: rating content by assigning star-rating, e.g. IMBD
- IP3 Extensive feedback (see Figure 2): rating content by liking/disliking certain movie criteria

Most participants were familiar with thumbs-rating systems (IP1) and star-rating systems (IP2), but none had encountered an extensive feedback system based on multiple rating criteria as per IP3. In general, the subjects rated IP3 to allow most accurate recommendations and therefore would trust recommendations based on IP3 the most. IP3 was valued for increasing individuality (”reducing mass taste” (P10)) and finding niches. IP1/IP2 were assessed to give a “rough direction” (P10), but trust is notably lower as the system “of course [doesn’t] know for what reason [the] film is rated good” (P7) and “not much information could be drawn from that” (P12). However, IP3 was considered to be too time-consuming for rating multiple movies in a row.

While IP3 was valued for its capability of collecting highly detailed feedback which is supposed to make recommendations more precise and trustworthy, three users found the detailed approach questionable. P7 stated “I can imagine it is difficult because the answer possibilities are already very detailed [...] meaning I can only choose what is offered to me” while P3 “don’t want to ‘dissect’ the film into [its different categories because] I find these things play into it subconsciously”. Lastly, IP3 appeared to be “long-winded [and] it feels elaborate [going] through the three or four steps” (P9).

However, as time is considered scarce and participants value low interaction efforts [63], providing extensive feedback through a prototype such as IP3 was found to be acceptable by only five participants, while three would try the system.

Getting Recommendations
After showing the participants the systems to input their ratings, two static prototypes to show recommendations on a VOD platform were presented and should be compared:

Output: Receiving Recommendations – Screenshots
- OP1 Plain recommendations: no further categorization and information
- OP2 Explainable recommendations: clustered by matching movie criteria as obtained per IP3
While comparing OP1 and OP2, three user groups emerged. One group preferred no further information on recommendation sources (OP1), while the second user group preferred the data-intensive approach OP2. Lastly, a third group indicated wanting more details on demand. The plain prototype OP1 was preferred by five participants who found the displayed cluster information in OP2 too noisy or not relevant. Reportedly, OP2 “is too much for me to read” (P9) and “I don’t want to know all that [...] a system should show me movies I like and not bore me with that” (P4). In contrast, five subjects valued OP2 for two main reasons. OP2 supports users to make sense of the selection as P1 reported: “the second one is great because it is already showing me why [the movies are suggested] [...] I don’t have to think too much, and I can make a decision quicker”. Furthermore, participants believed “[having] more trust in [...] suggestions” (P2) of OP2. And lastly, some participants desired to be able to “get even deeper into it” (P10) and manipulate the recommendations. These findings go hand in hand with previous research findings which show that a need for (a certain type of) explanation is dependent on personal characteristics [8, 57, 37]. “[Getting] even deeper into it” was also connected with approaching the filter-bubble and a need for exploration. Here, ten participants pointed towards the fact that recommendations must take different moods and social situations into account. P11 highlighted “you know, sometimes you want to watch a movie with good humour [and sometimes] you want to watch a really historical movie” as well as P10 said “I like these things [...] but wouldn’t watch it if I wanted to have an evening for two”.

Conclusion
Within the initial study, current user journeys of VOD users were investigated, resulting in the following pain points and design directions:

Pain Points and Design Directions
- Users feel overwhelmed by the amount of content (‘Netflix Effect’ [45]) ⇒ Support users to find suitable movies more easily
- Users with shared profiles may be unsatisfied with recommendations due to imprecise movie suggestions ⇒ Allow shared profile users to find recommendations based on their individual needs

Mixed
- Users have misconceptions about the rating functionality ⇒ Improve feature communication
- Users often cannot make sense of recommendations ⇒ Consider explanations (and allow to adjust the system to personal explanation preferences)
- Users find recommendations only partially satisfying and some experience a filter-bubble ⇒ Allow users to discover diverse content
- Users consider giving extensive feedback to be of high interaction-cost (but often see benefit) ⇒ Ensure that interactions have a high interaction-benefit

However, sizeable potential was salient in the direction of (interactive) explanations and findings were put in action in the prototypes of the following main study.

Main Study
Building on the initial study, which found explanations and further exploration of the eventual recommendations and MRS in high favour, the main study was set out to investigate definite forms of interactive explanations in practice. In particular, three different types of interactive explanations (described in the following section) were compared to a non-interactive/non-explaining baseline and their effect on various dimensions like user-system trust, sense-making, and others was measured. For every type of explanation, an independent prototype was developed. All prototypes were based on a hybrid recommender system that used the participant’s Netflix viewing activity to show meaningful, personalised recommendations. In contrast
to other studies [55, 52], this study’s recommender system used real user data and therefore allowed for more natural interaction and is of higher ecological validity.

The following section introduces the implemented system and prototypes, the researcher’s hypotheses, participants, and experimental design.

MovieLand Prototypes
Based on the identified design directions and related work, three concepts for interactive explanations were derived and implemented in a fully-working movie recommender system called MovieLand.

Hybrid Recommender System
MovieLand is a hybrid movie recommender system web-application. To offer personalised recommendations, the participants were asked to upload their Netflix viewing activity CSV file5 within the onboarding process of the study. The Netflix viewing activity was matched against the openly available MovieLens dataset by GroupLens. For performance reasons, the latest “MovieLens 100k / Small” dataset (100.000 ratings by 600 users on 9.000 movies, 9/2018) was used to identify titles in one’s Netflix viewing activity. This was achieved by exact string matching as Netflix did only provide the item title and date of consumption (as of June 2020). To keep the generation of recommendations short and to avoid edge cases, only the latest 30 matched items were considered for further data processing. The resulting set of recommendations was no subject of further analysis as every user got highly personalised and different recommendations.

To get an accurate but yet rich set of recommendations per participant, multiple knowledge sources were combined within a hybrid recommender system algorithm following the work of [9, 37]. Therefore, the following sources were used to generate recommendations:

- Collaborative Filtering
  A CF-based approach was implemented using a Keras model based on embeddings. Based on Banerjee’s work [6], the recommendation model was trained on the MovieLens 100k dataset and saved as a pre-trained TensorFlow model. For every matched movie, similar users and their top recommendations on other movies were used to retrieve meaningful, personalised recommendations. Furthermore, redundant recommendations between similar users were flagged to be more important.

- Content-Based Recommendations
  Content-based recommendations were retrieved on a genre/story keywords and actor basis as the initial study revealed these factors to be the most important rating factors for the participants. Two individual sets of recommendations were generated using the “The Movie Database” API (TMdb). The vendors of TMdb do not provide any further information about the implementation of their algorithms [3]. Similar to the CF approach, a recommendation importance score was calculated for redundant recommendations.

5https://help.netflix.com/en/node/101917

The three different knowledge sources were limited to the 100 most important recommendations per source. Subsequently, the individual recommendation sets were merged to one weighted hybrid set of recommendations [9] based on 100 recommendations adhering to the following fractions: 50% CF-based, 35% CB-genre-based, 15% CB-actor-based.

Basic Prototype Functionality
Independent of each explanation type, the prototypes offered the user a basic set of functionalities (see Appendix Figure 13): All recommendations were listed in a seven-column grid of movie covers which listed one item below the other. With hovering over a movie cover, an overlay showed meta-information about the movie such as the title, genres, length, and original language. Furthermore, two buttons allowed to either retrieve more information about a movie in a modal dialog or adding the movie to a list of favourites. The latter function was necessary for the main task of the study (see Procedure).

Tech-Stack and Data Source
The Netflix activity matching algorithm and hybrid recommender system were implemented using Python and generated individual JSON files for each source of recommendations and their respective raw movie data (title, genres, etc.). MovieLand’s frontend was implemented using HTML5, CSS3, and jQuery. All movie data was retrieved by TMdb.

Types of Explanations (Prototype)
The following section lists the different types of explainable interfaces (independent variables) which were implemented into the MovieLand prototype (see Figure 3). To test interactivity as an explanation method, a Hybrid Wizard of Oz (HWoO) approach was followed. According to Viswanathan et al., HWoO prototypes are high-fidelity prototypes that simulate intelligent behaviour programmatically without requiring a human wizard. Although MovieLand was based on an artificially intelligent algorithm, changing the data input did not actually trigger generating a new set of recommendations but resembled the effect. The HWoO approach was chosen to avoid unpleasant waiting times in the user interface. Each prototype is based on the findings of related work and the initial study.

1) Timeline
The timeline prototype (see Appendix Figure 15) listed all matching movies of a participant’s uploaded Netflix history ascending chronologically. A text informed the participants that their recommendations were based on N movies between the starting date and today. Each movie in the timeline came with a checkbox (initially activated) allowing the user to include/exclude a movie from the MRS’s input data. With toggling a movie, the user was able to make sense of how their recommendations changed on a particular title: A percentage number indicating the match-quality depending on the current selection was updated. And the set of recommendations may have changed depending on each item’s connection to the toggled movie. The numeric indicator was based on the calculated importance of a movie, based on the current selection of movies.

As inspired by [68], this prototype was set out to allow users to make sense of their recommendations by asking “what if?” questions about their recommendations (e.g. “What
Figure 3. This figure shows the feature-weighting prototype as an example for one of the four implemented fully-functioning prototypes. Screenshots of all prototypes can be found in the appendix.

would be recommended to me if I’d only include these movies?”). Furthermore, this prototype encouraged users to fine-granularly control which movies should be considered for generating recommendations. This approach was chosen to yield higher recommendation relevancy and satisfy users that want to influence recommendations in shared user profiles.

2) Feature-Weighting
The feature-weighting prototype (see Appendix Figure 16) aimed to enable participants in influencing the hybrid recommender system’s actor and genre fractions through a user interface (UI) slider component as introduced by [52]. Both sliders introduced their scope on an importance scale: Participants were able to increase/decrease the importance of the actor/genre-similarity affecting the number of recommendations per knowledge source. To satisfy the need for diversity and exploration, the resulting set of recommendations contained at least 20% CF-based movies. Hence, the actor/genre-based weights influenced up to 80% of the recommendations depending on the chosen importance. With decreasing the importance of actor/genre-based recommendations, CF-based recommendations increased. To allow fine-granular control, each slider control was divided into six steps while labels offered guidance ("not important"/"somewhat important"/"very important"). While this form of interactivity allows users to understand how different qualities affect their recommendations, this concept was also chosen to investigate its effectiveness in terms of the filter-bubble problem. This is because decoupling recommendations from certain similarities (e.g. actors) may allow participants to overcome the feeling of being in a filter-bubble.

3) Genre-Swapping
The genre-swapping prototype (see Appendix Figure 17) was based on the idea of providing mood-related recommendations [19] and allowing users to “get even deeper into it” (P10, Initial Study) through manipulating these clusters. Initially, the interface shows participants a mood cluster based on their two most common genre categories. The subjects were able to add and delete certain genres and combine them with other genres (“swapping”). Here, the participants were able to combine up to three genres at a time. While adding/deleting genres allowed users to control complexity, combining genres aimed to foster the exploration of (diverse) recommendations (filter-bubble problem). Furthermore, the interactive genre-swapping control was accompanied by a textual description to support sense-making.

4) Non-Interactive/Non-Explaining Baseline
To effectively investigate the aspect of interactivity in explanations, a fourth non-interactive prototype (see Appendix Figure 14) without explanations was added as a baseline. This prototype did not allow the users to control the algorithm in any way and simply showed a set of recommendations based on the previously described hybrid MRS algorithm.

All interactive prototypes supported animated sorting to improve a participant’s understanding of how their input changed the output of the algorithm [46].

The remainder of this thesis uses the following terminology to speak about the investigated prototypes:

- **Baseline Prototype:**
  - Non-Interactive/Non-Explaining (NI_PRO)
One Netflix Account

Profile N
Profile 1
SINGLE usage
Profile 2
SHARED usage

Figure 4. VOD platform profiles (N per account) may be used by SINGLE users or are SHARED between multiple users. Recommendations in SHARED accounts are therefore based on the mixed preferences of many users.

- Interactive-explainable Prototypes:
  Timeline (TL_PRO), Feature-Weighting (FW_PRO), Genre-Swapping (GS_PRO)

Hypotheses
Each concept was examined in the following user study, whereby their effect on user-system trust, sense-making, and other constructs (see Design and Measurements) was measured. Based on previous work in the field of interactive explanations and the initial study’s findings, the following hypotheses about the prototypes were derived:

Concerning Interactive Explanations in General
- H1: Interactive explanations support the user’s sense-making process.
- H2: Interactive explanations increase user-system trust.
- H3: Interactive explanations yield higher user satisfaction with an MRS.
- H4: Interactive explanations increase cognitive demand.
- H5: ATI has an effect on the perceived cognitive demand of a prototype.

Concerning the Timeline
- H6: The timeline prototype satisfies SHARED profile users significantly more than SINGLE profile users.

Concerning Feature-Weighting / Genre-Swapping
- H7: Interactive exploration helps users to overcome the filter-bubble problem.

Each hypothesis is tested and discussed alongside the analysed data in the discussion part of this thesis.

Participants
The researcher recruited the participants directly and via social network posts. This direct approach was chosen to balance gender (18 females, 12 males), age (M = 28.9 years, SD = 7.17), and profile usage types (see Figure 4). Table 1 shows a distribution of the subjects across the demographic and highlights user characteristics. To be eligible for participation, the subjects had to fulfil several requirements: The setup of the study required the participants to be Netflix users and have a viewing activity that yields at least seven matched movies in the study’s MRS. This requirement was checked programmatically within the onboarding of the study (see Procedure). Furthermore, all participants were required to have a good level of English as the study prototype was in English language. Lastly, access to a laptop/PC with an Internet connection was necessary to qualify for participation. Out of 38 recruited participants, 30 individuals were eligible to participate in the study. Eight participants were not considered as their Netflix viewing activity yielded too few results, or they aborted the study before completion. After successfully completing the study, the participants were able to take part in a lottery and win £100. A total of 29 people participated in the lottery and three times £100 were raffled.

<table>
<thead>
<tr>
<th>Demog./characteristic</th>
<th>Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>20–24</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>25–29</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>30–34</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>35–39</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>40–44</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>45–49</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>50–54</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>55–60</td>
<td>1</td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>12</td>
</tr>
<tr>
<td>Profile Usage Type</td>
<td>SINGLE</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>SHARED</td>
<td>18</td>
</tr>
<tr>
<td>Used VOD platforms</td>
<td>Netflix</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>YouTube</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>Am. Prime Video</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>BBC iPlayer</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Disney Plus</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>6</td>
</tr>
<tr>
<td>VOD-platform usage</td>
<td>0–5 h/week</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>6–10 h/week</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>11–15 h/week</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>16+ h/week</td>
<td>2</td>
</tr>
<tr>
<td>ATI</td>
<td>M = 4.1, SD = 0.91</td>
<td></td>
</tr>
<tr>
<td>ATI distribution</td>
<td>[1,2): Low affection</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>[2,3)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>[3,4): Med. affection</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>[4,5)</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>[5,6): High affection</td>
<td>5</td>
</tr>
<tr>
<td>Native language</td>
<td>German</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 1. This table shows the distribution of the participants (N=30) across the demographic together with their user characteristics.

Apparatus
The user study was conducted online and accessed through a custom-developed web application running on Django. This study portal was based on a Django application implemented
The above-noted setup and variables combined with 30 participants represent a total of 120 trials. To counterbalance order effects, the participants were randomly assigned to one of four Latin Square groups. While doing so, the study portal ensured a balanced number of participants in all Latin Square groups. The study setup is summarised in Figure 5.

**Measurements**

The following section describes the three different questionnaire types (see appx.) and their respective measurements and questions which the participants had to fill out during the study. For an extensive theoretical background about the standardised scales, look at the Related Work section. If not indicated differently, all used 5-point Likert scales measured the participant’s level of agreement from “strongly disagree” to “strongly agree”.

- **Pre-Study Questionnaire**
  The pre-study questionnaire was used to collect the participant’s native language, age, gender, and VOD platform usage details like used portals, time of usage/week, and profile usage type.

- **Post-Test Questionnaire**
  Each of the four conditions was connected to a post-test questionnaire which evaluated the participant’s experience with the previously tested explanation type. Therefore, the following standardised scales were used, and partially adapted, as well as open questions were asked.

  - **Preferences**: Two separate open questions surveyed the participants on what they liked and disliked about the recommender system. These questions were used to build a deeper understanding of the user experience and support the objective measures.
  
  - **User-System Trust**: To effectively measure the multi-factor construct of user-system trust, Berkovsky et al.’s trust scale was used and adapted towards the context of this study [8]. The trust scale measures user-system trust on the six dimensions presented in Table 2.
  
  - **Satisfaction**: The subject’s satisfaction with the MRS was measured through three items on the dimensions of recommendation quality [38], diversity-need satisfaction, and exploration-need satisfaction.
  
  - **Sense-Making**: The influence of the explanation type on a participant’s sense-making process was measured using an adapted version of Alsufiani et al.’s sense-making scale [5]. The original sense-making scale (5 subscales/16 questions) was reduced to the relevant subscales and to decrease the demand on the subjects. The eventual questionnaire consisted of the three dimensions and six questions: Gaining insight (build understanding), connections (between user input and recommendations), reducing uncertainty and ambiguity.

  - **Cognitive Demand**: The resulting cognitive demand of a prototype was calculated from a subset of five items. In particular, the following three items were combined with the “reducing uncertainty and ambiguity” dimension of the sense-making scale: “Using this recommender system was NOT mentally demanding.”, “The recommender system’s complexity was reasonable.”, “The effort put into using this
Figure 5. Experimental design of the main study: All participants tested all prototypes (_PRO) in sequence (orders were determined by Latin Square Design, e.g. Group 1: TL-GS-FW-NI).

### Dimension Description

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRUST_C</td>
<td>Competence: system’s knowledgeableness</td>
</tr>
<tr>
<td>TRUST_I</td>
<td>Integrity: honest and unbiased recommendations</td>
</tr>
<tr>
<td>TRUST_B</td>
<td>Benevolence: ability to reflect a user’s interests</td>
</tr>
<tr>
<td>TRUST_T</td>
<td>Transparency: comprehensibility of MRS’s data source</td>
</tr>
<tr>
<td>TRUST_R</td>
<td>Reuse: desire to use MRS again</td>
</tr>
<tr>
<td>TRUST_O</td>
<td>Overall: MRS’s overall trustworthiness</td>
</tr>
<tr>
<td>TRUST</td>
<td>Average of all dimensions</td>
</tr>
</tbody>
</table>

Table 2. MRS trust dimensions according to Berkovsky et al. [8]

A recommender system was worth the result” (Interaction-Cost-Benefit).

- **Interaction Logs**
  
  While conducting the task (adding at least five movies to the favourite list), the task completion time and the number of bookmarked movies were logged. Saved items on the favourite list were not considered for analysis due to potential bias by the respective prototype.

- **Post-Study Questionnaire**
  
  The post-study questionnaire surveyed the participants on different dimensions to allow a direct and conclusive comparison of each explanation type. Therefore, three open questions (see Table 3) shed light on the dimensions of a system’s ease of use, supportiveness and its ability to counteract the filter-bubble. Furthermore, the participants were asked to rank each prototype on a place between 1 (best prototype) and 4 (worst prototype) together with a justification for their choice. Finally, the previously described Affinity for Technology Interaction scale by Franke et al. was used to measure how likely each participant interacts with technology in general [20].

### RESULTS

The results of the main study are divided into two sections: The first section covers the results of the statistical analysis of the questionnaire scales and interaction logs. Subsequent, the results of the open questionnaire questions are reported.

**Analysis of Scales and Interaction Logs**

All data was pre-processed using Python and evaluated in R. The pre-processing of the data involved selecting only participants that have completed the whole study resulting in a data set of one pre-study questionnaire response, four post-task questionnaire responses, four interaction logs (four prototypes), and one post-study questionnaire response. Following, the internal consistency of the TRUST, SAT, SENSE, and DEMAND scales’ sub-items was tested by calculating Cronbach’s alpha. None of the scales violated the internal reliability, which enabled the calculation of their scale means. Furthermore, each dependent variable’s distribution was checked for normality through Shapiro-Wilk’s and visual tests.

Any potential main effects resulting from the influence of the independent variable (prototype) on the dependent variables (see Design and Measurements) were investigated using an analysis of variance. For normally distributed data, a One-Way ANOVA for repeated measures and Mauchly’s Test for Sphericity with automatic Sphericity corrections for violations was conducted. Non-normally distributed data were investigated using Friedman’s test. Next, pairwise comparisons of all prototypes were carried out. While all post-hoc tests applied a Bonferroni correction, pairwise t-test were used for normally distributed data, and Wilcoxon tests were used for non-normally distributed data.

To investigate the relationship between profile usage type on satisfaction and one’s affinity for technology interaction (ATI)
on a prototype’s perceived demand, linear mixed-effects models\textsuperscript{9} were set up as previously done by [47]. Each model was fitted with individual intercepts for the subjects as this study involved repeated measures following the advice of [65]. In total, two linearmixed-effect models were created to investigate potential main effects of profile usage type/ATI (Formula 1) and their interaction effects on the different prototypes (Formula 2). For the latter, the \textit{Factor} and \textit{Fixed Effect} were swapped (profile_type or ATI) depending on the effect of interest.

\textit{DV} \sim \textit{ATI} + \textit{profile_type} + \textit{Prototype} + (1|\textit{subject}) \quad (1)

\textit{DV} \sim \textit{Factor} \times \textit{Prototype} + \textit{FixedEffect} + (1|\textit{subject}) \quad (2)

To identify potential effects, a likelihood-ratio-test was conducted and compared the respective models in two ways according to [65]: Model 1 (Formula 1) was fitted with the \textit{Factor} and compared against a null model without the \textit{Factor}. Model 2 was fitted with the interaction effect between the \textit{Factor} \times \textit{Prototype} and compared against model 1 in which the \textit{Factor} was a fixed-factor. The p-values of both likelihood-ratio-tests were obtained by t-tests using Satterthwaite’s method.

All results are reported as statistically significant with \(p < 0.05\) if not explicitly stated otherwise.

\textit{Sense-Making}

Each prototype’s ability to support sense-making was measured on three different subscales (see Design and Measurements) and an overall \textit{SENSE} score (\(\alpha = 0.89\)) was calculated. All results can be found in Table 4.

The prototype’s ability to support constructing an understanding and make sense of the available information within the movie recommender system (\textit{SENSE}\textsubscript{G}I) was found to significantly differ among the prototypes (\(\chi^2(3) = 14.7, p = 0.002\)). All interactive-explainable prototypes improved the participants’ ability to construct an understanding and make sense of the information compared to the non-interactive and non-explaining baseline (M = 3.13, SD = 1.02). The weighting features (M = 4.00, SD = 0.49) was found to be of supremacy in terms of its ability to support sense-making and significantly differed from the baseline (Z = 203, \(p = 0.014\)). Similarly, the genre-swapper (M = 3.93, SD = 0.83, Z = 134, \(p = 0.038\)) and timeline (M = 3.90, SD = 0.65, Z = 14, \(p = 0.011\)) significantly differed from the baseline.

Also, the prototypes’ ability to help users to understand how their input influenced the MRS (\textit{SENSE}\textsubscript{CO}) yielded a significant difference between the tested systems (\(\chi^2(3) = 19, p < 0.001\)). On average, the feature-weighting (M = 3.80, SD = 0.76) and timeline prototype (M = 3.80, SD = 0.89) were equally helpful to understand the relation between user input and MRS output. And both, the feature-weights (Z = 278, \(p = 0.012\)) and timeline (Z = 34.5, \(p < 0.001\)) enabled users significantly better understand this connection compared to the non-interactive baseline (M = 2.77, SD = 1.02).

Lastly, the analysis revealed a significant difference among the prototypes’ qualities to reduce uncertainty and ambiguity (\textit{SENSE}\textsubscript{RE}, \(\chi^2(3) = 16.8, p = 0.001\)). The feature-weighting prototype (M = 3.60, SD = 0.80) was most helpful to reduce uncertainty and significantly differed from the baseline (M = 2.83, M = 0.83, Z = 178, \(p = 0.04\)). Also, genre-swapping (M = 3.55, SD = 0.89, Z = 180, \(p = 0.032\)) and the timeline (M = 3.48, SD = 0.91, Z = 22.5, \(p = 0.036\)) enabled users to significantly reduce ambiguity and uncertainty compared to the baseline.

In line with these findings, also the overall satisfaction score (\textit{SAT}) differed significantly between the UIs (F(3,87) = 10.11, \(p < 0.001\)). Mauchly’s test for sphericity indicated no violations (p = 0.379). The overall \textit{SENSE} scores were transformed to a normal-distribution using Yeo-Johnson transformation (Shapiro-Wilk test: W = 0.98, p-value = 0.073). All interactive-explainable prototypes were found to be significantly more satisfactory compared to the non-interactive/non-explaining baseline: The feature-weighting prototype (M = 3.5, SD = 0.84) was perceived to be the most satisfactory prototype and differed significantly from the baseline (M = -0.76, SD = 0.91, \(p = 0.002\)). Followingly, the timeline (M = 0.27, SD = 0.96, p

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|}
\hline
Prototype & NI_PRO & TL_PRO & FW_PRO & GS_PRO & Significance \\
\hline
TRUST & *** & ** & & & \\
\hline
SENSE & & ** & *** & ** & \\
\hline
SAT & & & & & \\
\hline
\end{tabular}
\caption{Each prototype’s \textit{TRUST}, \textit{SENSE}, and \textit{SAT} scores including significant post-hoc tests.}
\end{table}
Table 4. Results of the satisfaction and sense-making scales including significant post-hoc tests.

User-System Trust

The user-system trust was measured on six different scales (see Table 2). On that basis, a summarising TRUST score ($\alpha = 0.87$) was calculated from the average of the sub-scales (see Table 4.1.2).

A significant difference in competence was found ($\chi^2(3) = 6.79$, $p = 0.079$) assuming a confidence level of $p = 0.1$. However, only the genre-swapping prototype ($M = 3.73$, $SD = 0.98$) was perceived as significantly more competent than the baseline ($M = 3.13$, $SD = 0.94$, $Z = 118$, $p = 0.047$).

Moreover, a significant difference between the integrity of the prototypes was found using a Friedman test ($\chi^2(3) = 10.3$, $p = 0.016$). But Wilcoxon tests for pairwise comparisons did not show any significant differences after the Bonferroni-correction was applied.

The ability to reflect a user’s interests (benevolence) was found to differ significantly among the prototypes ($\chi^2(3) = 15.2$, $p = 0.002$). In comparison, the genre-swapping prototype ($M = 3.70$, $SD = 0.65$) was found to be significantly more benevolent than the baseline ($M = 2.87$, $SD = 0.82$, $Z = 244$, $p = 0.006$). On average, the timeline ($M = 3.70$, $SD = 0.99$) was perceived as equally benevolent as the genre-swapping prototype and also differed significantly from the baseline ($Z = 33$, $p = 0.022$).

Highly significant differences were perceived in terms of the prototype’s transparency ($\chi^2(3) = 28.6$, $p < 0.001$). All interactive-explainable prototypes differed significantly from the baseline. The feature-weighting prototype ($M = 3.80$, $SD = 1.06$) was perceived most transparent about its data sources that affect its decisions and differed significantly from the baseline ($M = 2.17$, $SD = 1.12$, $Z = 346$, $p = 0.001$). Besides, the timeline ($M = 3.70$, $SD = 1.26$) was found to be significantly more transparent than the non-interactive and non-explaining baseline ($Z = 17$, $p = 0.002$). And lastly, the genre-swapping prototype ($M = 3.40$, $SD = 1.10$) differed significantly in terms of its transparency compared to the baseline ($Z = 265$, $p = 0.006$).

Next, a significant difference was found for the participants’ desire to use certain prototypes again beyond the study ($\chi^2(3) = 11.4$, $p = 0.010$). The majority of the participants indicated that they would like to use any of the interactive-explainable prototypes, while the baseline was considerably less desired. In particular, the feature-weighting prototype ($M = 3.43$, $SD = 1.25$) was found to be significantly more likely to be used again compared to the baseline ($M = 2.43$, $SD = 0.86$, $Z = 272$, $p = 0.017$). Similarly, a significantly higher desire to use the genre-swapping prototype ($M = 3.43$, $SD = 0.94$) was expressed compared to the baseline ($Z = 229$, $p = 0.004$). Also the timeline ($M = 3.30$, $SD = 1.21$) was significantly more desired in future scenarios than the baseline ($Z = 40$, $p = 0.015$).

Lastly, also the perceived overall trustworthiness was significantly disparate ($\chi^2(3) = 16.6$, $p = 0.001$). Compared to the baseline ($M = 2.93$, $SD = 0.94$), the genre-swapping prototype ($M = 3.67$, $SD = 0.84$, $Z = 148$, $p = 0.037$) was considered to be significantly more trustworthy overall. Also, the feature-weighting prototype ($M = 3.67$, $SD = 1.09$) differed significantly in terms of its trustworthiness comparing it to the non-explaining baseline ($Z = 41$, $p = 0.03$).

The summarising TRUST scores were found to be normally distributed (Shapiro-Wilk test: $W = 0.98$, $p$-value = 0.120). All interactive-explainable prototypes were found to be significantly more trustworthy than the non-interactive/non-
explaining baseline (F(3, 87) = 10.75, p < 0.001). Since the sphericity was found to be violated (p = 0.020), sphericity corrections were applied. Overall, the timeline prototype (M = 3.65, SD = 0.81) was found to be most trustworthy while there was a significant difference in comparison to the baseline (M = 2.77, SD = 0.63, p < 0.001). Also, TRUST in the genre-swapping prototype (M = 3.59, SD = 0.53) was significantly higher compared to the non-interactive/non-explaining prototype (p < 0.001). And lastly, participants trusted the feature-weighting prototype significantly more than the baseline (p = 0.006).

### Satisfaction

The satisfaction was measured on the three different scales (see Design and Measurements) and a summarising average SAT score (α = 0.84) was calculated. All results can be found in Table 4.

A significant difference in satisfaction concerning the general recommendation quality among the prototypes was found (SAT_R, χ²(3) = 14.6, p = 0.002). Here, the weighting features (M = 3.80, SD = 0.89) yielded the most satisfactory movie recommendations compared to the static baseline (M = 2.97, SD = 0.76, Z = 179, p = 0.032). Also, the recommendation quality obtained through swapping genres (M = 3.73, SD = 0.74) was found to yield significantly more satisfactory recommendations compared to the baseline (Z = 203, p = 0.011).

Furthermore, a significant difference was found between the resulting satisfaction of a prototype in terms of its ability helping the participants to explore movies of their interest (SAT_E, χ²(3) = 18, p < 0.001). Here, the feature-weighting prototype (M = 3.70, SD = 0.84) was found to satisfy the exploration need best and a significant difference was perceived in contrast to the non-interactive/non-explaining baseline (M = 2.80, SD = 1.00, Z = 205, p = 0.009). Also, the genre-swapping prototype (M = 3.60, SD = 0.81) yielded a significantly higher exploration-need satisfaction compared to the baseline (Z = 159, p = 0.006).

Also, the systems' ability to support the user to find a diverse range of movies of their interest was found to differ significantly (SAT_FB, χ²(3) = 12.5, p = 0.006). Again, the feature-weighting prototype (M = 3.77, SD = 0.82) was most satisfactory to approach a filter-bubble and was perceived to be significantly more satisfactory in that matter compared to the static baseline (M = 3.00, SD = 1.05, Z = 182, p = 0.021). Similarly, the genre-swapping prototype (M = 3.73, SD = 0.91) satisfied the participants significantly more in the matter of finding diverse movies compared to the baseline (Z = 192, p = 0.042).

To approximate the summarising SAT scores to a Gaussian distribution for further statistical analysis, SAT data was transformed using a Box-Cox Transformation. While the Shapiro-Wilk test did not reveal a statistically significant normal distribution (W = 0.97, p-value = 0.003), a visual test showed a close-to-normality distribution. An analysis of variance revealed a significant difference in overall satisfaction between the prototypes (χ²(3) = 20.8, p < 0.001). In particular, the feature-weighting prototype (M = 0.34, SD = 0.89) was found to be significantly more satisfactory than the baseline (M = -0.64, SD = 0.85, Z = 331, p = 0.004). Furthermore, the participants found the genre-swapping prototype (M = 0.25, SD = 0.85) also significantly more satisfactory than the baseline (Z = 292, p = 0.003).

### Cognitive Demand

The cognitive demand of a prototype was transformed to a normal distribution using Box-Cox’s method (Shapiro-Wilk test: W = 0.98, p-value = 0.133). As demand was based on a

<table>
<thead>
<tr>
<th>df = 3</th>
<th>TRUST_C</th>
<th>TRUST_I</th>
<th>TRUST_B</th>
<th>TRUST_T</th>
<th>TRUST_R</th>
<th>TRUST_O</th>
<th>TRUST</th>
</tr>
</thead>
<tbody>
<tr>
<td>χ² / F(df, df_e)</td>
<td>6.79</td>
<td>10.3</td>
<td>15.2</td>
<td>28.6</td>
<td>11.4</td>
<td>16.6</td>
<td>F(3, 87) = 10.75</td>
</tr>
<tr>
<td>p</td>
<td>0.079</td>
<td>0.016</td>
<td>0.002</td>
<td>&lt; 0.001</td>
<td>0.010</td>
<td>0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>NI_PRO – M</td>
<td>3.13</td>
<td>3.10</td>
<td>2.87</td>
<td>2.17</td>
<td>2.43</td>
<td>2.93</td>
<td>2.77</td>
</tr>
<tr>
<td>SD</td>
<td>0.94</td>
<td>0.61</td>
<td>0.82</td>
<td>1.12</td>
<td>0.86</td>
<td>0.94</td>
<td>0.63</td>
</tr>
<tr>
<td>TL_PRO – M</td>
<td>3.73</td>
<td>3.63</td>
<td>3.70</td>
<td>3.70</td>
<td>3.30</td>
<td>3.83</td>
<td>3.65</td>
</tr>
<tr>
<td>SD</td>
<td>0.98</td>
<td>0.93</td>
<td>0.99</td>
<td>1.26</td>
<td>1.21</td>
<td>0.99</td>
<td>0.81</td>
</tr>
<tr>
<td>FW_PRO – M</td>
<td>3.63</td>
<td>3.47</td>
<td>3.47</td>
<td>3.80</td>
<td>3.43</td>
<td>3.67</td>
<td>3.58</td>
</tr>
<tr>
<td>SD</td>
<td>0.93</td>
<td>0.90</td>
<td>0.90</td>
<td>1.06</td>
<td>1.25</td>
<td>1.09</td>
<td>0.81</td>
</tr>
<tr>
<td>GS_PRO – M</td>
<td>3.73</td>
<td>3.60</td>
<td>3.70</td>
<td>3.40</td>
<td>3.43</td>
<td>3.67</td>
<td>3.59</td>
</tr>
<tr>
<td>SD</td>
<td>0.58</td>
<td>0.67</td>
<td>0.65</td>
<td>1.10</td>
<td>0.94</td>
<td>0.84</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 5. Statistical results of the TRUST scale including significant post-hoc tests.
set of positively formulated questions, the resulting variable LD ($\alpha = 0.84$) indicated an actually lower demand for higher values. An analysis of variance revealed significant differences among the prototypes LD scores ($F(3,87) = 4.32$, $p = 0.007$). No violations of the sphericity were found ($p = 0.634$). However, only the genre-swapping prototype ($M = 0.22$, $SD = 1.01$) was found to be significantly less demanding than the non-interactive baseline ($M = -0.51$, $SD = 0.82$, $p = 0.023$).

**Interaction-Cost-Benefit and Interaction Logs**

The results of the analysed Interaction-Cost-Benefit (ICB) and interaction logs on Task Completion Time ($M = 164.59$ seconds, $SD = 114.95$ sec), Bookmarked Movies (number of movies added to the list of favourites), and Time per Movie Selection (task completion time / bookmarked movies) are reported in Table 6.

**Ranking**

After using all prototypes, the participants were asked to rank the prototypes. The rankings were investigated individually by profile usage type (SINGLE/SHARED) as well as the overall rankings (BOTH profile usage types mixed) were analysed. The non-interactive/non-explaining baseline did rank lowest in any condition. Great variances were found for the respective rankings by profile usage type (see Figure 7).

**Affinity for Technology Interaction**

To check whether ATI has an effect on the participant’s ability to TRUST certain prototypes, a linear mixed-effect model analysis was conducted using a likelihood-ratio test. Within this test, ATI did not have a significant main effect on the perceived TRUST among the prototypes ($\chi^2(1) = 0.516$, $p = 0.476$). In line with that, no significant interaction effect between ATI and particular prototypes was found ($\chi^2(1) = 1.667$, $p = 0.644$).

In a subsequent analysis, a significant main effect of ATI on the perceived cognitive demand caused by a prototype was found (LD, $\chi^2(1) = 7.01$, $p = 0.008$). ATI increased LD by $0.29 \pm 0.10$ (standard error). Recalling that LD was inversely coded, a higher LD value actually indicates lower cognitive demand perceived by the user. Participants with a higher affinity for technology interaction, therefore, perceived significantly lower cognitive demand using any of the prototypes. A subsequent covariate analysis of a potential interaction effect between ATI and a particular prototype (see Figure 8) and its effect on LD did not yield any significant differences ($\chi^2(1) = 1.68$, $p = 0.644$). Hence, this study did not find ATI to yield a significant difference in cognitive demand between particular prototypes.

**Profile Usage Type**

No significant main effect of the profile usage type on one’s satisfaction with any of the tested prototypes was found in a likelihood-ratio test (SAT, $\chi^2(1) = 0.0019$, $p = 0.9657$). Consequently, whether participants SHARED a Netflix profile or used it on an individual basis did not have a significant effect on their satisfaction with the tested MRS in this study. Following these findings, also no significant interaction effect between the profile usage type and a particular prototype was found ($\chi^2(1) = 5.41$, $p = 0.066$).

**Insights into User Preferences**

The following section reports on the open text responses to the questions asked in the post-task and post-study questionnaires after using all fully-functioning prototypes. The answers (N=360) were pseudonymised, categorised by question, and imported to Dovetail10. Subsequently, a random data sample was analysed to generate codes inspired by the dependent variables of the quantitative study. An initial set of eleven codes was reduced to the nine most salient codes:

- **Cognitive demand** while using a prototype, e.g. “I like that the system has narrowed down the selection” (P19)
- **Control** over the used movie recommender system, e.g. “I can decide what genre to include” (P24)
- **Recommendation quality**, e.g. “some recos were really off” (P26)
- **Interaction effort** while using a prototype, e.g. “[..] was easy to use” (P30)
- **Fun & novelty** concerning the user-interface, e.g. “I thought this was novel and interesting.” (P18)

---

10https://dovetailapp.com/
<table>
<thead>
<tr>
<th>df = 3</th>
<th>LD</th>
<th>Interaction-Cost/Benefit</th>
<th>Completion Time</th>
<th>Time per Selection</th>
<th>Bookmarked Movies</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \chi^2 / F(df, df_e) )</td>
<td>F(3,87) = 4.32</td>
<td>12.1</td>
<td>F(3,87) = 11.34</td>
<td>19.20</td>
<td>4.27</td>
</tr>
<tr>
<td>p</td>
<td>0.007</td>
<td>0.007</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.234</td>
</tr>
<tr>
<td>NI_PRO – M</td>
<td>-0.51</td>
<td>3.03</td>
<td>-0.45</td>
<td>21.80</td>
<td>6.33</td>
</tr>
<tr>
<td>SD</td>
<td>0.82</td>
<td>0.81</td>
<td>1.08</td>
<td>19.91</td>
<td>2.72</td>
</tr>
<tr>
<td>TL_PRO – M</td>
<td>0.08</td>
<td>3.57</td>
<td>0.32</td>
<td>34.45</td>
<td>6.43</td>
</tr>
<tr>
<td>SD</td>
<td>1.01</td>
<td>0.97</td>
<td>0.96</td>
<td>25.26</td>
<td>2.58</td>
</tr>
<tr>
<td>FW_PRO – M</td>
<td>0.20</td>
<td>3.77</td>
<td>0.17</td>
<td>28.63</td>
<td>6.80</td>
</tr>
<tr>
<td>SD</td>
<td>1.02</td>
<td>1.10</td>
<td>0.88</td>
<td>19.61</td>
<td>2.63</td>
</tr>
<tr>
<td>GS_PRO – M</td>
<td>0.22</td>
<td>3.73</td>
<td>-0.04</td>
<td>28.90</td>
<td>5.87</td>
</tr>
<tr>
<td>SD</td>
<td>1.01</td>
<td>0.98</td>
<td>0.95</td>
<td>21.96</td>
<td>1.61</td>
</tr>
<tr>
<td>Post-hoc Z</td>
<td>p_{bonf}</td>
<td>GS &gt; NI 3.15</td>
<td>0.023</td>
<td>–</td>
<td>FW &gt; NI 3.79</td>
</tr>
</tbody>
</table>

Table 6. Results of the lower cognitive demand LD (higher values indicate lower cognitive demand, see Cognitive Demand) and interaction logs including significant post-hoc tests.

Figure 9. This figure shows the obtained rankings according to the qualitative data analysis of the questionnaire responses (double-mentions considered).

- **Exploration** of diverse content, e.g. “I find movies with the same actors in a different genre” (P3)
- **Sense-making** of the MRS functioning, e.g. “I could not see what the recommendations were based on” (P14)
- **Miscellaneous/No reason**, e.g. “Genre-swapping” (P22, answer on which prototype the easiest was and why)

The report on the qualitative data focuses on the codes that appeared most frequently under a given theme (statement-based count). The seven reported themes hereafter are based on the individual survey questions. Figure 9 shows an overview of the participants’ preferences in the ranking-based questions. The categories about what was liked/disliked about a certain prototype were excluded as these questions did not have a ranking-character (e.g. “most supportive” vs “what did you like”). The responses were checked for recency effects whereby no violations have been found. While the actual ranking numbers as reported under the statistical results are based on a web form with one choice per ranking place only, this section reports on the written justifications for these rankings. Please note, that the qualitative analysis respected double-mentions and therefore slightly differs from the numbers given in the form-based ranking.

**Most Liked Prototype**

Overall, the feature-weighting prototype (12) was ranked highest, followed by the timeline (10). The subjects justified their top-rankings with an increased sense of control (9) over the movie recommender system: A timeline user stressed the point that “no interaction is no longer a solution. I want to influence my surface.” (P6) as well as another user reported “[Through the feature-weights] I was able to influence my recommendations based on factors I wanna emphasise in my selection.” (P19).

Followingly, certain filter (9) qualities were crucial for high rankings, too, as for example found for the slightly less favoured genre-swapper (8): “I like to be able to filter via genre types based on my mood instead of looking through a lot of movies of different genres [...]” (P2, genre-swapping). This enhanced control over the MRS and made subjects appreciating the improved resulting recommendations (7).

Lastly, few users (2) ranked the non-interactive baseline on first place as they appreciated the prototype for its low interaction effort and simplicity. P29 reported “It’s nice to [...] adjust your preferences, but also sometimes you just want to watch something random and not think too much about options.”.

**Positive Qualities**

Across all prototypes, a sense of control (41) and good movie recommendations (26) were most appreciated. Furthermore, it became salient that the participants valued aspects of prototypes which improved their ability to make sense of the
Figure 10. This figure shows the number of statements made per code (double-mentions considered) per question.
MRS (25). This became overly striking (19) with the timeline prototype: “I really liked the fact that I can filter through my timeline and deselect movies that were not watched by me but [by] the person I share my Netflix [profile] with.” (P15). P21 “[…] liked being able to see the past films [they] had watched in the past year. It allowed [them] to better reflect on what [they] enjoyed and didn’t enjoy.”.

Furthermore, novel and fun ways to interact with the recommendations (25) pleased the participants in many cases. This was especially true for the timeline (13) and feature weights (10). P16 highlighted “I really liked the slider options of “how important [are] the genre or actors?” I loved how I could see [the impact] in real time [that it had on recommendations].”.

Connectedly, it has been reported that low interaction effort (24) is a factor of high importance, too: “[the slider was] very easy and fast to use” (P26).

Lastly, lowering the cognitive demand (18) a user has to face while searching for movies was also favoured: P19 reported to “[like] that the [genre-swapping prototype] has narrowed the range of [their] selection”. In line with that, P15 briefly noted: “Very simple to use, low effort. […] I liked having fewer recommendations to evaluate.”.

Least Liked Prototype
The lowest-ranked prototypes – predominantly the non-interactive baseline – failed to please the user for lack of sense (10) and control (5). The comments included: “[The non-interactive prototype], the worst prototype, provided no insights into why the listed films were being recommended to me and [I had] no control over changing them.” (P18), “Incomprehensible where the recommendations come from – not trustworthy – no filtering possible” (P1), or “Not any indication why something was recommended and no possibility to alter my recos, very time consuming and just too much. While the range of [recommendations was] good, it was just too confusing.” (P26).

However, interactive-prototypes were also ranked on the lowest place as individual filters (5) and functions happened to not be suitable. The feature-weighting or genre-swapping prototype, for example, did allow the users adjustments on qualities that may be not relevant to them as for P10 who said “Actors are not interesting for me.”. Or P30, who stated that “[I wouldn’t use the genre-swapping so the filter would be useless]”.

Negative Qualities
While unsuitable filters only rarely were a reason to rank a prototype lowest, 39 statements were made on missing, too few, or inappropriate filters to influence (15) the movie recommender system concerning all prototypes. Furthermore, the recommendations (32) itself were sometimes considered to be “too wide” (P1, NI_PRO) or “a bit random” (P5, NI_PRO). This predominantly was the case for the baseline (10). And lastly, some participants demanded a higher number of recommendations in general.

Some more specific problems were found in connection to the actual nature of the prototypes: With the non-interactive prototype, the comments included: “I don’t feel [like] it is under my control. It all depends on the algorithm.” (P7) and “No individualization visible” (P1).

The timeline prototype made some users demanding more filter options: “It would’ve been nice to filter the recommendations that were provided.” (P21). Or the match-indicator (see MovieLand Prototypes) caused confusion “The percentage presented did not match my taste or the movies I had watched before.” (P9)

Similarly, having only two weightable features was not enough for some users, and they reported: “there could be more of those sliders” (P4) or missed “[a] filter option regarding the time period” (P23).

Lastly, some users “disliked that [they] could only choose up to 3 genres [in combination],“ (P17) or did not like the logic of the genre-swapper: “only OR, not AND-combination of genres” (P26).

Easiest to Use Prototype
As seen in Figure 9, “all of the prototypes were easy [to use]” (P9). The subjects indicated that high ease of use is connected to low general interaction effort (19) or did not further explain (17) their ranking decision. Some participants also justified their rankings by mentioning certain prototype qualities like particular filters (8).

The non-interactive prototype was perceived to be easy to use (10) because it did not require the user to dispute with certain (novel) UI controls. Therefore, low interaction effort (7) was the main reason to rank the NI_PRO highest as highlighted by P24: “[…] I don’t need to do anything […] I don’t need to think. […] I just want to browse through without thinking too much, and pick whatever looks good for me.”.

Nevertheless, also the interactive prototypes were considered to be of low interaction effort (12), as their UI controls were perceived to be ‘straightforward’ (P11) and ‘intuitive’ (P14). In particular, the feature-weighting (12) and genre-swapping prototype (12) were most often ranked as the easiest prototypes. The participants commented “The feature-slider was pretty straightforward. I like that I had control and knew what to do immediately” (P11). Besides, P9 claimed that “[…] [FW_PRO and GS_PRO] were the most intuitive [prototypes] to play with and explore.”. Lastly, P18 “[…] thought that adding or subtracting genres with large breadcrumb-like buttons is fairly intuitive and common in other filter system platforms.”.

The timeline was perceived to be slightly less easy to use (9).

Most Supportive Prototype
Not surprisingly, no user considered the non-interactive baseline to be the most supportive prototype (0). In total, swapping genres (13) was found to be most supportive, followed by the timeline (12) and weighting features (8). Overall, these UIs were found to be most supportive as they gave the user control (16) over their recommendations and “produced more accurate results” (P21) (14).
“Eliminate certain movies” (P3) and have “more control over the data [a user was] feeding into the algorithm” (P21) was found to be most supportive when using the timeline. This not only gave the users more control (5) but also helped them together with the “percentage match-bars” (P22) to make sense of the recommendations (6). Some even mentioned that “next time when I want to select a movie to watch, I would like to have the option to include or exclude [certain movies]. So this is quite important to me.” (P24).

Choosing the most supportive prototype was often connected to speaking about finding recommendations matching a certain ‘mood’ (P24, P29, P22). Here, swapping genres (8) and weighting certain features (5) was found to yield the best-matching recommendations. Here, participants commented: “I sometimes go through a phase where I watch all the movies from the same actor, but sometimes I just want to watch something in a certain genre. I think [the feature-weighting prototype is] useful to adjust the recommendations to my mood.” (P29). Or “[The genre-swapper is] useful as I look for films according to my mood. Choosing a genre allows for this need to be met directly.” (P22).

Prototype With the Least Perceived Filter-Bubble
The non-interactive prototype (13) gave users most often the feeling of not being in a filter-bubble. While some (9) did not give any particular reasons for this, others explained this perception more detailed. Here, two main reasons were salient: A lack of transparency (“not seeing what the recommendations are based on”, P14) can cause users to believe that their viewing activity does not influence recommendations. Also, some participants stated that these recommendations (same data source as other prototypes) had a certain amount of randomness and variety (“The results showed quite a certain randomness. But that’s not what I expect from a recommendation system...”, P5). However, the non-interactive prototype was assessed to produce the least present filter-bubble, this was not always appreciated:

The interactive features were reportedly a way to be in control (6) and escape the filter-bubble, especially for the feature-weighting prototype which ranked second (7) and the genre-swapping prototype which ranked third (6). Users reported to take action in the following ways: “I can change the [feature-weights] when I feel am in a filter-bubble.” (P7) or “I could influence the selection myself [through using the feature-slider]” (P1). P19 noted “[with the genre-swapping prototype] I was able to see different movie types and explore some other genres. I had the feeling my recommendations were quite rich.”.

The timeline which allowed the user to make recommendations more specific, ranked lowest (1) that category.

DISCUSSION
Nowadays, users of Video on Demand platforms often experience the ‘Netflix Effect’ [45] – a feeling of being overwhelmed by the sheer number of recommendations – and a lack of understanding about why these recommendations were made to them. This is a feeling that can cause distrust in AIs [31]. This thesis explored interactive explanations that help to increase the transparency of these artificially intelligent “black box” recommender systems.

A two-fold research approach investigated interactivity in explainable artificial intelligence. In an initial study based on semi-structured interviews, the following research question was investigated:

RQ1: How acceptable is it to users to give extensive feedback on watched movies as a way to derive highly personalised explanations for movie recommendations?

This research found extensive feedback through rating movies on certain qualities to only have mediocre acceptance. While some users liked the idea and saw the value in tagging movies to get highly personalised explanations and recommendations based on their input, a large number of participants was sensitive to the seemingly high interaction effort required to eventually get these recommendations. Strong pointers to misconceptions about rating systems in general also suggest that many users see rating movies as something not relevant to themselves but more of an act of kindness for other users that profit from ratings. While increased user input was not found to be of major acceptance, the idea of having more explained guidance on the output of recommendations as well as being able to further explore recommendations was favoured by many. This desire together with a growing demand for more control in intelligent everyday applications [19] and a gap in research [4, 48, 52, 19, 68], motivated this thesis to explore the benefit of interactive explanations in MRS:

RQ2: How can interactive explanations improve user-system trust and sense-making in movie recommender systems?

Building on the initial study, a subsequent main study investigated three different fully-functioning types of interactive explanations (timeline, feature-weighting, and genre-swapping) in comparison to a baseline prototype that allowed no reasoning and interactivity (extensive description of all prototypes: MovieLand Prototypes). The underlying MRS used the participant’s previously uploaded Netflix viewing activity to show personalised, and therefore more relevant recommendations. The following sections critically discuss the results alongside the hypotheses and draw the line to current research in the field.

Making Sense and Trusting Movie Recommender Systems
This study found that all interactive explanations enabled users to make more sense of the movie recommender system’s internal functioning and its eventual recommendations compared to the non-interactive and non-explaining baseline. Its impact on an MRS’s success is immense. A lack of provision of explanations was found to be the foremost reason why the baseline was ranked lowest, followed by a demand for more control. The improvements of sense-making can be most likely attributed to two intertwined reasons:

1. First, the static text nature of explanations given by all explainable prototypes (e.g. “Recommendations are based on your most-watched genres ‘Comedy’ and ‘Romance’”) allowed users to gain insight and construct an understanding
of how these systems work and use their data (SENSE_GI). These explanations helped to significantly reduce uncertainty and confusion in comparison to the non-explaining baseline (SENSE_RE). The results also suggest that allowing users to make sense of a system’s way of functioning is also a proxy for user-system trust concerning the transparency (TRUST_T), as both items had substantial overlap regarding the questions and responses. This connection was also recognised by a participant who criticised the non-explaining baseline: “Incomprehensible where the recommendations come from - not trustworthy” (P1). These findings stand in the same light with previous work that found static text explanations to be an effective way to increase user-system trust [8, 35].

2. Second, the interactivity component of explanations allowed users to make sense as they were able to interactively influence the MRS and see how it affects the recommendations in real-time. In particular, weighting movie features and including/excluding certain movies on the timeline helped to significantly better understand the connection between one’s (viewing) activity and the eventual recommendations (SENSE_CO). This form of sense-making is exclusive to prototypes that allow the interactive experience of causality (influence and following effect) and was expectedly not found to be supported by the genre-swapping prototype (due to its filter-nature) or the non-interactive baseline. As proposed by Zürn et al., this cause-effect form of interactivity through feature weights and the timeline was found to be an effective way to answer “what if?” questions in the context of movie recommender systems [68]. Users reportedly liked this form of exploration and appreciated sense-making itself – “I really liked the slider options of ‘how important [are] the genre or actors?’. I loved how I could see [the impact] in real-time [that] it had on recommendations.” (P16). Given that, interactivity itself likely also contributed to the perceived transparency of a system (TRUST_T) as it reveals its internal way of working and responding to data.

As all interactive-explainable prototypes significantly improved sense-making in comparison to the non-interactive and non-explainable baseline, the hypothesis H1 (Interactive explanations support the user’s sense-making process) can be accepted. These findings expand the knowledge of other studies [8, 35] and show the importance of a symbiosis of interactivity and static textual explanations to make sense and build trust in MRS.

Trust in Competence, Benevolence and Overall Trustworthiness
While all interactive-explainable prototypes gave the user a feeling of the MRS being more knowledgeable about their interests in movies, only the genre-swapping prototype significantly differed from the baseline. This difference was not much bigger than the other prototypes, yet it was significantly better than the baseline. The genre-swapper’s unique agency of narrowing down the recommended movies and informing the user about its most-watched set of genres (e.g. Comedy and Romance) let one assume that this could be the reason for an increased perception of competence. This assumption builds on the fact that all other prototypes required the user to take agency (timeline/feature-weights) themselves or did not allow any control (baseline). Interestingly, the timeline was considered as competent as the genre-swapper, but this perception varied vigorously between users. The qualitative findings suggest that this might be due to the match-percentage indicator as some felt “[the indicator] didn’t seem to get it right exactly” (P7).

Furthermore, all interactive-explainable prototypes caused the users to experience higher levels of trust in a system’s benevolence while only the timeline and genre-swapping prototypes significantly differed from the baseline. Using the timeline and swapping genres influenced trust in a way that participants saw their interests reflected by the system and is not surprising. The timeline allowed users to (de)select movies that should influence their recommendations. Also, the genre-swapping prototype respected that the majority of users take mood-related decisions based on the genre to select movies as found in the initial study. This finding also became salient again in the main study where participants reported that they “look for films according to [their] mood. Choosing a genre allows for this need to be met directly.” (P22). In contrast, neither the baseline nor the feature-weighting prototype allowed a user to yield recommendations for certain moods as the former only allowed to yield more/fewer results across all genres of particular actors. Therefore, these results did not directly resemble mood-related decisions as users then still had to go through the list and filter recommendations themselves – an understandable reason to trust a system’s benevolence less. Finding the timeline and genre-swapping to support mood-based searches is likely also the reason why these prototypes were perceived to be the most supportive prototypes while the feature-weighting prototype did slightly worse here.

In general, the integrity of all prototypes did not differ significantly. However, the interactive prototypes were perceived to yield slightly more honest and unbiased recommendations which presumably builds on the increased transparency (TRUST_T). Overall, all interactive-explainable prototypes caused users to trust the system significantly more than they trusted the non-interactive and non-explainable baseline. Here, the timeline was the most entrusted form of interactive explanation and users reported to “like[…] the novelness of the timeline approach, seeing my viewing history made the connection between my viewing data and resulting recommendations really explicit and made it feel more personal.” (P15). Based on the overall improvement of user-system trust for all interactive explanation types in comparison to the baseline, H2 (Interactive explanations increase user-system trust) can be accepted.

To conclude, it can be said that interactive explanations greatly help users to make sense of an MRS and therefore yield high user-system trust. This closely-knit connection between sense-making and trust was expected and found in prior work [18, 48, 17]. As pointed out before, interactive explanations improved user-system trust in general, but certain prototypes influenced various levels of trust differently. This furthermore stresses
the point that user-system trust is a complex construct which must be measured on multiple different factors rather than on a single dimension as pointed out by Berkovsky et al. [8].

Satisfying Users Through More Control
Besides absence of transparency, lack of control was found to be the second most important reason to rank a prototype lowest. At the same time, the most frequently mentioned negative aspects about all – even the interactive – prototypes were about missing or unsuitable possibilities to filter recommendations. This demand for control forms a seemingly strong case for the autonomous user of intelligent everyday applications like MRS. This study found that giving users more control through the provision of feature-weights or the ability to swap genres yielded more satisfying recommendations. “[Through the feature-weights. I was able to influence my recommendations based on factors I wanna emphasise in my selection.” (P19).

However, when looking at this finding more closely, it is somewhat surprising that the genre-swapping prototype was perceived to yield significantly better recommendations compared to the baseline but not the timeline. This is because both prototypes accessed the same set of recommendations and their controls were allowed to filter these and therefore make a selection more specific to the user’s current needs. Adjusting the controls in both cases, however, did not trigger the generation of new, more specific recommendations in comparison to the feature-weights that revealed more recommendations based on the weighted criteria (actors/genres). A possible explanation for this could be that the genre-swapping prototype was positively affected by a ‘halo effect’ [54] which made the user believe that this prototype yields better recommendations in general as it seems to know more about the user’s preferences (TRUST_C).

Even though all interactive prototypes yielded more satisfying recommendations compared to the baseline (not significant for GS_PRO), the average number of bookmarked movies did not significantly differ between the prototypes and the participants only bookmarked 6.36 movies on average. This is only slightly higher than the required minimum of five movies as defined for the main task (for details see Procedure).

Interestingly, the feature-weighting prototype was the most liked and highest-ranked prototype and yielded the most satisfying recommendations, but it was the least supportive prototype among its interactive competitors. This might be due to the following two reasons: (1) As mentioned earlier, finding suitable movies is often mood-based. The feature-weighting slider did not allow as much control as the timeline/genre-swapper that enabled the user to narrow the selection to specific genres/moods. (2) The particular weightable features turned out to be irrelevant to some users as indicated by “I have a hard time remembering actors and I don’t care, I prefer to select the movies through a genre filter instead of this slider.” (P2) or “Actors are not interesting for me.” (P10). Hence, limited supportiveness does not necessarily yield more inadequate recommendations but makes an interface less relevant to a user and, therefore, presumably negatively influences the user experience. However, it must be considered that users were able to unweight less relevant features (e.g. adjusting the slider to ‘actor: not important’) to consider their disinterest in a certain feature, but still, these irrelevant features were reported on. This may suggest that increasing certain feature weights is perceived to be of higher relevance than decreasing it.

Filter-Bubble and Exploration
Whether one’s perception of being in a filter-bubble changes with one of the prototypes was investigated in a two-fold way which yielded unexpected results: Participants considered the feature-weighting prototype and genre-swapping prototype to deliver the most diverse recommendations (SAT_FB) and be most explorative. Therefore, both were significantly different from the baseline and timeline. However, when participants were asked which prototype leads to the least-perceived filter-bubble, more than one-third of the participants mentioned the non-interactive/non-explaining baseline more often than the feature-weighting/genre-swapping prototypes and the timeline (see Figure 9).

Qualitative insights suggest that (1) no explanations for the data source (transparency) causes a (2) “certain randomness and [impersonality]” (P6). However, this randomness was not necessarily what [participants] expect from a recommendation system.” (P5). Participants found influencing the algorithm through filters like the feature-weights or genre-swapping to be most satisfactory as it enabled them to “[…] change the filter when I feel am in a filter-bubble.” (P7). In particular, unweighting features (lowering importance of certain actors/genres of one’s history) can be considered most-effective as “the various extremes produce very different results.” (P18). The timeline led to the highest perceived filter-bubble (highest transparency), while the resulting recommendations were not perceived to be much different from the feature-weighting and genre-swapping prototype. Hence, interactive exploration may reduce the feeling of being in a filter-bubble, but increased transparency through (interactive) explanations can coincidentally worsen the perception.

In accordance with previous work [53, 51, 67] and the findings at hand, the actual improvement of being in a filter-bubble is based on more diverse recommendations – a requirement that was met by the feature-weighting and genre-swapping prototype (not significant for the timeline). Therefore, H7 (Interactive exploration helps users to overcome the filter-bubble problem.) can only be accepted if interactive exploration at the same time yields more diverse results.

This study found strong evidence for the “desire for more control” and “[a] potential practical value of a more interactive approach to explanation” [19] as identified by a large-scale analysis of user needs concerning intelligent everyday applications. Allowing users more control over an MRS and enabling them to make sense of the system and its recommendations were found to improve the level of satisfaction significantly. Consequently, the results prove H3 (Interactive explanations yield higher user satisfaction with an MRS.) to be true.

\[1\]non-diverse, bias-limited recommendations
Interaction Effort and Demand

Giving users control is of utmost importance to satisfy them. This thesis’ initial and main study’s results suggest that users are sensitive to interaction effort but are not afraid to invest effort provided that it comes with a high benefit. While the measured Interaction-Cost-Benefit (ICB) did not lead to any significant differences between all prototypes, interactive-explainable prototypes still yielded a higher ICB. This is most likely because the baseline did not allow much interaction but also showed that all interactive prototypes were about to be equally beneficial to the participants. However, findings on the easiest to use prototype suggest that low interaction effort is of the highest importance in general. At the same time, more effortful interactive controls (other than the non-interactive prototype) were perceived to be easy to use because of their ‘innovative’ (P9) and ‘straightforward’ (P11) way of functioning and their ‘practical value’ [19].

In an environment characterised by choice overload [51, 45], it is furthermore of paramount importance to lower the cognitive demand a user faces while using an MRS. The study found that the genre-swapping prototype was the only prototype to be significantly less demanding than the baseline, while the other interactive prototypes were not significantly disparate. The genre-swapping prototype lowered the cognitive demand by taking agency and narrowing the number of recommendations made to the user by selecting only recommendations from the two most favourite genres. This finding suggests that future systems could consider agency and reducing choice overload to assist users in a complex and noisy environment. However, decreasing choices may be only beneficial to some users as prior studies found that choice overload and resulting indecisiveness is subject to individual differences [58, 43].

Coincidentally, the perceived demand was significantly correlated with the participant’s Affinity for Technology Interaction which caused less tech-savvy users to perceive higher demand over all prototypes. Less tech-savvy users supposedly experience higher demand as they engage less (intensively) with technology and are less familiar with interactive technology. Therefore, H5 can be accepted (ATT has an effect on the perceived cognitive demand of a prototype.). Finding one’s affinity for technology interaction – “a key personal resource” [20] – a significant predictor for the perception of explanations in recommender systems contributes to the knowledge about a mix of affecting personality traits in XAI like the need for cognition [47], The Big Five [8], or desire for control [33].

Surprisingly, the added interactive explanations did not increase the cognitive demand faced when using an MRS but instead lowered the cognitive demand. However, it was lowered significantly for only one prototype. H4 (Interactive explanations increase cognitive demand.), therefore, cannot be accepted.

Lastly, it must be considered that adding interactive-components can cause significantly longer interaction times as seen for the feature-weighting prototype and timeline. However, this is not necessarily considered to be negative, as the participants “[... enjoyed playing with different configurations and seeing the recommendations being added and shuffled” (P15, FW_PRO). Furthermore, they “liked being able to dig through [their] past viewing [activity]” (P21, TL_PRO), which likely improved their sense-making process and general UX.

Shared Profile Users

The qualitative findings of the initial and main study suggest that some SHARED profile users (see Figure 4) are less satisfied with their recommendations which is most likely to a diverse set of titles in their viewing history. Therefore, the timeline was predicted to “work well with a shared profile” (P15) as one shared-profile user pointed out. However, the statistical analysis did not find any significant differences when comparing SHARED/SINGLE profile users and their satisfaction with any of the prototypes. However, the overall ranking indicates that some SHARED profile users consider the timeline as the best prototype while, coincidentally, another fraction does not seem to see the value in this prototype (see Figure 7). This might be because SHARED profiles do not necessarily yield unsuitable recommendations as users that use one profile together may have similar preferences in movies. Therefore, H6 (The timeline prototype satisfies SHARED profile users significantly more than SINGLE profile users.) cannot be accepted and must be rejected as it is not unconditionally true.

Lastly, it must be noted that all interactive-explainable prototypes did not significantly differ from one another in any dimension and their improvements seemed to be of almost equal extent. However, as the rankings and qualitative findings suggest, a notable personal preference for specific interfaces was sensible and suggests that different users benefit in various ways.

This study found interactive explanations to be a suitable – but moreover – a strongly demanded way to allow users more control over movie recommender systems. Coincidentally, these interactive explanations helped users to effectively build mental models of how these systems work and therefore they trusted these MRS significantly more. Consequently, the participants expressed their wish to make use of these systems again in future scenarios. Future systems should therefore build on the findings of the presented prototypes that applied feature-weighting, a timeline approach, and genre-swapping as a way to satisfy their users with more control and higher transparency. While this work contributed to an understanding of user-system trust and sense-making in interactive explanations, it also raised new questions for further scientific work which are presented in Limitations and Future Work.

LIMITATIONS AND FUTURE WORK

The implemented hybrid recommender system was limited in its movie matching performance and the generation of recommendations. As the processed Netflix viewing activities only included the movie title, exact string matching against the MovieLens dataset reportedly led to a small number of improper matches. Furthermore, the timeline/genre-swapping prototypes did not trigger the generation of new recommendations after altering the input as the generation of new titles lasted several seconds. In comparison, the feature-weighting prototype was able to pull slightly more recommendations
from the pre-generated actor/genre recommendations. However, this was not the focus of this study. Future work could therefore look into prototypes with a higher recommendation accuracy and the generation of recommendations in real-time based on user input.

Furthermore, the affinity of technology interaction in participants was not normally distributed as the majority of the participants was relatively tech-savvy. Future work could, therefore, investigate a higher sample size and a broader range of ATI among the participants. Also, prospective studies could investigate the effect of profile usage types (SINGLE/SHARED) as a primary goal and investigate whether heterogeneity in taste or a varying number of users that share a profile influences the suitability of certain prototypes.

Lastly, this study found high variance in preferences on certain movie qualities. Therefore, feature-weights that allow weighting criteria which are of personal importance to individuals (e.g. similarity of directors, epochs, etc.) seems to be a promising research area.

CONCLUSION
This study investigated interactive explanations as a measure to enhance sense-making and build user-system trust in recommender systems on Video on Demand platforms. Two subsequent studies investigated different types of interactive explanations in empirical user studies. Strong evidence was found that users not only benefit from interactive explanations, but they also demand more control and transparency over movie recommender systems. The main takeaways of the initial study and the preceding main study are as follows:

- Users who are asked to give extensive feedback on watched-movies to derive personalised explanations report moderate acceptance, but they report high favour on explained-recommendations
- Greater transparency through interactive explanations fosters sense-making and yields higher user-system trust
- Users desire to have more control over movie recommender systems and, as a product, can obtain better recommendations: “No interaction is no longer a solution. I want to influence my surface.” (P6)
- Users with a higher affinity for technology interaction (tech-savviness) experience reduced cognitive demand when interacting with these systems

Video on Demand platforms must increasingly implement interfaces that allow users to explore and make sense of their recommendations interactively. In a highly saturated market that continues to increase AI-usage, returning some autonomy to the user can enable services like Netflix or Amazon Prime Video to ease the user-journey, and make their services more satisfying.

ACKNOWLEDGEMENTS
I am incredibly thankful for the huge methodological guidance, provision of the previously implemented Django application, and deployment management by my supervisor Dr. Enrico Costanza. Furthermore, I want to express my gratitude to all 42 participants of the initial and the main study – without you, this research would not have been possible. And lastly, I want to thank my parents for their long-standing emotional and material support during my Bachelor and Master studies.

REFERENCES


### User Details
- has mental model about RS algorithm (4)
- is review/rating-focused (6)
- is review/rating-independent (1)
- uses shared profile (3)
- NF better than APV (3)
- APV better than NF (1)

### Current Experiences
- information overload (5)

### Ratings
- rating misconception (2)
- doesn't know about rating fn. (4)
- uses ratings (2)

### Recommendations
- fully satisfied with rec. (3)
- partly satisfied with rec. (6)
- not satisfied with rec. (3)

### Future Systems
- requirement (7)
- idea towards future systems (4)

### Ratings
- pros of providing ext. feedback (9)
- cons of providing ext. feedback (7)
- investing time / making effort (11)
- sees return of feedback time-invest (8)
- rate movie after watching it (8)
- rate movie at a later point (6)
- ext. feedback feedback (11)
- star-rating feedback (11)
- thumbs feedback (10)

### Rating Criteria
- exploration/mood-related need (10)
- show (some) rec. sources (9)
- rec. sources help making sense (8)
- rec. sources foster trust (5)
- rec. sources are unnecessary (5)
- avoid filter-buble (4)
- show content details (3)

### Future Systems
- requirement (7)
- idea towards future systems (4)

### Rating Criteria
- cast (7)
- genre (7)
- indiv. dramaturgic/mood pref. (6)
- cinematographic style (6)
- ratings/reviews (6)
- subject (3)
- director (3)
- length (3)
- language (1)
- music (1)
Input: Rate a Movie

Please rate the following series you have watched

Output: Get Recommendations

Movie/series recommendations for you

This selection matches your preferences in

Movie/series recommendations for you

This selection matches your preferences in
Figure 13. Main Study: Basic interaction patterns supported by all four prototypes.

Figure 14. NI_PRO: Non-interactive/Non-explaining Prototype (Baseline)
Figure 15. TL_PRO: Timeline Prototype

Your recommendations are based on 9 movies you've watched between 2017 till today.

2017 ™ 2018

Lammbock
RIDICULOUS
PIXELS
SANDY WEXLER
DEEDS
DCÉVER
WHEN WE FIRST MET

Your Movie Recommendations (84)

Figure 16. FW_PRO: Feature-Weighting Prototype

Your Movie Recommendations (98)

Recommend videos which are similar to your watched movies based on...

Actors

not important somewhat important very important

Genres

not important somewhat important very important
Figure 17. GS_PRO: Genre-Swapping Prototype
Figure 18. Pre-Study Questionnaire

Please answer the following questions before you start with the actual study:

1. What is your native language?

2. How old are you?

3. How do you identify?
   - Female
   - Male
   - Other
   - Prefer not to say

4. Which Video on Demand platforms do you use regularly?
   - Netflix
   - Amazon Prime Video
   - YouTube
   - Hulu
   - HBO
   - Disney Plus
   - BBC iPlayer
   - Now TV
   - Sky Go
   - Others

5. How many hours do you roughly spend on average per week on Video on Demand platforms (Netflix or other)?

6. Please read the description (below) and answer which of the following options describes your situation best?
   
   One Netflix ACCOUNT (e.g. shared by a family) can consist of multiple Profiles (e.g. single family members)
   
   - Single usage: I have an own Netflix Profile and almost exclusively watch on my own
   - Shared usage: I share my Netflix Profile with others and/or often watch movies/series together with others

Submit
Figure 19. Post-Task Questionnaire

Please answer the following questions based on your experience with the previous prototype:

1. Please name one or more things you liked about the recommender system.

2. Please name one or more things you disliked about the recommender system.

Trust

Please indicate the degree to which you agree/disagree with the following statements:

<table>
<thead>
<tr>
<th>Statement</th>
<th>strongly disagree</th>
<th>disagree</th>
<th>neutral</th>
<th>agree</th>
<th>strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>This recommender system is knowledgeable about the movies I'd like to see.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This recommender system provides honest and unbiased suggestions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This recommender system reflects my interests.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This recommender systems helps me to understand what the recommendations are based on.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For selecting my next movie to watch I would like to use this recommender system.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This recommender system is trustworthy.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Satisfaction

Please indicate the degree to which you agree/disagree with the following statements:

<table>
<thead>
<tr>
<th>Statement</th>
<th>strongly disagree</th>
<th>disagree</th>
<th>neutral</th>
<th>agree</th>
<th>strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>This recommender system provides satisfactory recommendations.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This recommender system satisfies my need to explore movies I'm interested in.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This recommender system helps me to find a broad range of different movies I'm interested in.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The effort put into using this recommender system was worth the result.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using this recommender system was NOT mentally demanding.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The recommender system's complexity was reasonable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sense-Making

Please indicate the degree to which you agree/disagree with the following statements about whether the recommender system has helped you to perform the following processes successfully:

<table>
<thead>
<tr>
<th>Process</th>
<th>strongly disagree</th>
<th>disagree</th>
<th>neutral</th>
<th>agree</th>
<th>strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I was able to construct understanding from the available information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was able to make sense of the available information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was able to understand how my user behaviour influences my recommendations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was able to understand the connection between my viewing activity and my recommendations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was able to reduce confusion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was able to reduce uncertainty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 20. Post-Study Questionnaire

Almost done! Just some last questions :-) 
Please answer the following questions based on your experience with all prototypes:

Certain Prototype Qualities
1. Which of the prototypes above did you find the easiest to use and why?

2. Which of the prototypes above did you think you most enjoyed using and why?

3. Which of the prototypes above did you feel least being in a filter-bubble (recommendations too similar) and why?

Ranking
Please rank the prototypes (if needed, rank their names on the screenshots above):

<table>
<thead>
<tr>
<th>Place 1 (Best Prototype)</th>
<th>Place 2</th>
<th>Place 3</th>
<th>Place 4 (Worst Prototype)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Non-Interactive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(B) Timeline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(C) Feature-Slider</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D) Genre-Swapping</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please specify your reasons why you picked the respective prototype in the question above as the best and the worst prototype?

Technology Interaction
Please indicate the degree to which you agree/disagree with the following statements.

<table>
<thead>
<tr>
<th>Statement</th>
<th>completely disagree</th>
<th>largely disagree</th>
<th>slightly disagree</th>
<th>slightly agree</th>
<th>largely agree</th>
<th>completely agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I like to occupy myself in greater detail with technical systems.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I like testing the functions of new technical systems.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I predominantly deal with technical systems because I have to.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>When I have a new technical system in front of me, I try it out extensively.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I enjoy spending time becoming acquainted with a new technical system.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>It is enough for me that a technical system works, I don't care how or why.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I try to understand how a technical system exactly works.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>It is enough for me to know the basic functions of a technical system.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I try to make full use of the capabilities of a technical system.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Study in General
Please leave your comments about the overall experience of this study or your suggestions for improvement.

Submit