Identifying Group-Specific Mental Models of Recommender Systems: A Novel Quantitative Approach

Johannes Kunkel, Thao Ngo, Jürgen Ziegler, and Nicole Krämer

University of Duisburg-Essen, 47057 Duisburg, Germany
{johannes.kunkel, thao.ngo, juergen.ziegler,
nicole.kraemer}@uni-duisburg-essen.de

Abstract. How users interact with an intelligent system is determined by their subjective *mental model* of the system's inner working. In this paper, we present a novel method based on card sorting to identify such mental models of recommender systems quantitatively. Using this method, we conducted an online study ($N = 170$). Applying hierarchical clustering to the results revealed distinct user groups and their respective mental models. Independent of the recommender system used, some participants held a strict procedural-based, others a concept-based mental model. Additionally, mental models can be characterized as either technical or humanized. While procedural-based mental models were positively related to transparency perception, humanized models might influence the perception of system trust. Based on these findings, we derive three implications for the consideration of user-specific mental models in the design of transparent intelligent systems.

Keywords: Mental models · Transparency · Recommender systems · Card sorting · Hierarchical clustering

1 Introduction

Mental models of intelligent systems are subjective, typically incomplete and flawed understandings of the system's inner working [38, 45]. They are shaped through system interaction [38]. Studying mental models can, thus, explain how users perceive a system and how they interact with it, e.g. by identifying superstitions or misconceptions. This is also a crucial prerequisite to explain elements of intelligent systems better and to increase their transparency [13, 54].

To investigate subjective mental models, research has focused thus far on qualitative approaches that characterize single mental models in greater detail using small samples (typically smaller than $N = 20$; e.g. [13, 34, 36]). While such qualitative studies describe single, overarching models and are valuable for a general comprehension of what is included in such a model, they struggle to capture their full *diversity* and lack the ability to reliably identify and systematically compare different mental models that might coexist in large samples. Such comparisons, most importantly, offer systematic insights into relationships

between specific mental models and user-centered aspects. We argue that this needs to be addressed through a quantitative approach.

In fact, quantitative methods might reveal individual and reappearing structures of mental models in *large samples* and *across* different systems. Hence, they could allow comparisons of diverse mental models among individuals and groups. Studying mental models quantitatively might also lead to practical implications for the design of user-friendly interfaces. Specific themes and visual perspectives could be designed for certain user groups, commonalities among models could foster general design of transparent systems. However, the application of quantitative methods still poses a serious challenge.

In this work, we aim to close this gap and explore the users' mental models of intelligent systems quantitatively. For this, we applied a novel card sorting setting which captured the entire processing chain of an intelligent system. The card sorting setting provided typical functional steps of intelligent systems (e.g. data acquisition) for users to reconstruct their mental model. The method allows us (1) to identify user groups and characterize their mental models, and (2) to explore the relationship of these user groups and mental models with system perceptions (e.g. transparency).

We applied this novel card sorting setting in the domain of recommender systems (RS) as RS are a mainstay in today's online environment. Furthermore, their decisions are often perceived as subjective which are met with more distrust by users than other systems that make more objective decisions (e.g. route planners) [6]. We asked RS users of a broad sample ($N = 170$) to sort different actions according to how they think the RS works internally. Hence, we aim at answering the following research questions:

RQ1: Which different mental models do users hold across RS?
RQ2: How do these mental models relate to the perception of RS?
RQ3: Based on these findings, which implications can we derive for the design of transparent intelligent systems?

With this work, we contribute to the advancement of research on users' assumptions and knowledge about intelligent systems in three ways: (1) We captured the entire processing chain of an intelligent system (i.e. RS) in a detailed way through a novel card sorting setting. (2) This allowed us to demonstrate and uncover which mental models are prevalent in a broad sample of RS users, thus forming a baseline for future research in this domain. (3) We derive practical implications for system designers regarding how the knowledge of such mental models can be leveraged to increase user-centric qualities of intelligent systems, such as transparency and trustworthiness.

2 Background & Related Work

Recommender systems typically appear as *black box* to users, i.e. their internal reasoning and functioning remain hidden. This can affect users negatively, e.g. it can cause feelings of creepiness towards recommendations [47]. Furthermore, users may distrust algorithmic decision making and reject its results [12, 40].

This *algorithm aversion* seems to pertain especially to situations of subjective compared to objective decision making [6]. As a result, RS that recommend subjective items (e.g. music or movies) are more affected by distrust than objective systems (e.g. route planners that give directions) [6]. To tackle this potential distrust in subjective decision support systems, transparency appears to be a central factor. Studies indicate that transparency can increase the users' trust in and satisfaction with a system [27, 49] and recommendation acceptance [11, 20].

A clear understanding of the users' knowledge and interpretation of the system's functioning is a key prerequisite for determining how to improve transparency and which parts of the system to focus on [13, 55]. A holistic depiction of such knowledge can be conceptualized as *mental models* [38, 39].

2.1 Mental Models of Intelligent Systems

Mental models (closely related to *folk theories*[1]) can be defined as cognitive knowledge representations of technological systems that serve users to cognitively simulate system behavior and predict its outcomes [45]. They are subjective in nature, and thus, may be parsimonious and flawed [38]. Mental models are developed through system interaction, especially when confronted with anomalies and unexpected behavior [18]. In other words, mental models represent *what* users know about a system and determine *how* they interact with it.

A field study by Tullio et al. [50] demonstrated that this also holds for intelligent systems. They found that, without prior knowledge about the system, users showed a basic understanding of machine learning methods when confronted with an intelligent agent. In particular users' mental models included decision trees and statistic predictions based on "patterns" and "averages".

Other research has highlighted the impact of mental models on the users' task performance. For example, in a qualitative study Muramatsu and Pratt [34] showed that flaws in mental models of search engines may cause confusion regarding the interpretation of search results. Despite the familiarity and daily use of search engines, many participants did not fully understand how search queries are processed. This is supported by a study of Kulesza et al. [26] which showed that improved soundness of mental models was positively related to the effectiveness of interaction with the system.

Most studies on mental models of intelligent systems focused a single general mental model, e.g. [13, 19, 36]. Eiband et al. [13] highlighted the importance of identifying one "overarching user mental model" of a target group and indicated that within this model, several group-specific mental models may exist.

Indeed, some findings indicate a diverse landscape of mental models. In the domain of RS, Ghori et al. [19] showed that users mostly explain technical concepts, such as collaborative filtering, in their own words. In an interview study, Ngo et al. [36] revealed that mental models of RS might be technical or metaphorical. The study also suggests that users had different views on the importance of themselves in the recommendation process.

[1] For a detailed discussion on *folk theories*, e.g. see [15, 16].

To summarize, while the elaboration of an overarching mental model for a system is useful, there is also strong support for the existence of diverse mental models within a population. To find a balance between one overarching mental model and an individual mental model for each user, we therefore argue to identify group-specific mental models. Even though qualitative approaches may provide some insights into the diversity of mental models, a quantitative approach is required to more precisely identify and classify these diverse models.

2.2 Methods for Eliciting Mental Models

Few studies have applied a quantitative approach to explore the mental models of intelligent systems. They mostly studied *effects* of mental models on the perception of a system. For instance, Kulesza et al. [26] induced different mental models and captured their "soundness" through multiple-choice questions. Thus, they did not directly investigate the structure and characteristics of mental models but the users' capacity of using them to simulate certain system outputs.

Other studies have used mixed-method approaches: Xie et al. [53] investigated the effects of mental model similarity on web page interaction performance in an experimental study. They combined a card sorting and a path diagram of web navigation and calculated different similarity measures based on these methods. A recent example studied mental models of cooperative AI agents in a game setting [18]. The researchers first applied a think-aloud task to explore the mental models. Then, a large-scale survey was conducted. We encourage such *informed quantitative studies* that exploit insights from former qualitative work.

Conceptual techniques, such as the repertory grid, pairwise rating, or card sorting [10, 30] can be used to study mental models quantitatively. They are based on an existing body of concepts which needs to be explored before, e.g. through interviews. Thus, they do not rely on direct verbalization [10].

In repertory grid and pairwise rating, users rate different concepts on a certain scale or compare them with one another. This leads to a similarity matrix between the concepts representing the user knowledge. The data can be analyzed through e.g. multidimensional scaling [30]. While both methods have different advantages, they are either time-consuming or are limited in the number of concepts that can be studied. Therefore, Cooke [10] recommends to apply card-sorting techniques, if the number of concepts is higher than 25–30.

In card sorting, users assign certain cards, representing concepts, into categories. The method is often used in usability studies to determine navigation structures [9]. There are different types of settings: In closed card sorting, the content of the cards and the label as well as number of categories are fixed. In open card sorting, participants can label the cards themselves [9]. The method allows for the identification of common themes and differences in samples [44].

Table 1: Overview of the four general categories and their associated action cards we used in the card sorting task.

Acquisition of user data	*Comparing items or users*
[01] Recording my mouse clicks	[10] Comparing items regarding their content
[02] Asking me for my age	[11] Matching rating data of items
[03] Recording my dwell time on an item's detail page	[12] Calculating a similarity score between items
[04] Asking me to explicitly rate items	[13] Calculating a similarity score between users
Inference and aggregation	*Presenting recommendations*
[05] Determining my interest in item categories	[14] Suggesting items that are new to me
[06] Analyzing my current mood	[15] Showing items, I might like
[07] Combining all data about me to an abstract user profile	[16] Presenting items that other users liked in the past
[08] Adding additionally item data that users cannot see	
[09] Analyzing content of items	

3 Identifying Diversity and Commonalities of Mental Models

We developed a new setting based on card sorting. Card sorting is suitable for large online studies [4], allows open and closed settings [9], and can be used to include a wide range of concepts [10]. Our setting considers the subjectivity of mental models by providing a diverse range of pre-defined cards, and allowing participants to formulate their own thoughts using open cards and as many actions and steps as they find appropriate to describe their mental model.

In our card sorting setting, participants are presented with a set of cards representing typical RS actions and are asked to assign them to up to seven sequential steps. Our method assumes a procedural structure of the inner workings of a RS. This is in line with how these systems typically work and with observations in previous qualitative user studies [36, 38, 50]. The resulting card sorts of each participant represents their mental model of RS. Through hierarchical clustering, card sorts can be aggregated into groups, allowing us to characterize the differences and commonalities between mental models in a larger sample.

3.1 Cards Used as Actions of RS

We carefully created 35 cards for participants to express their mental model:
- 16 *action cards*, represent actions of the recommendation process (Table 1)
- 12 *distractor cards*, represent actions that are not part of the central recommendation process
- 4 *question mark cards*, provide the possibility to express uncertainty
- 3 *open cards*, let users express self formulated actions

The *action cards* correspond to typical paradigms used by RS while still "*speaking the language of the user*". We extracted concepts of mental models from former qualitative mental model studies [13, 19, 26, 36, 50] and contributed our own technical expertise on RS functioning. In particular, we followed the four general categories provided by Ngo et al. [36]: (1) *acquisition of user data*, (2) *inference and aggregation*, (3) *comparing items or users*, and (4) *presentation of recommendations*. For each category, we designed up to five cards (Table 1). We describe the rationale behind the action cards in the following.

(1) Acquisition of User Data (Cards 01–04): For any personalized RS, elicitation of user data and their preferences is a necessary prerequisite [42, 46]. While these data can take various forms (e.g. ratings, purchases, clicks), the underlying concept appears to be well known by RS users and was mentioned in many in-depth qualitative user studies [13, 19, 36, 50].

(2) Inference and Aggregation (Cards 05–09): In almost all cases RS do not perform their recommending on raw user data, but aggregate them or infer further (e.g. situational) data [1, 3]. A similar concept can also be found in many user responses of prior interview studies. Users, for instance, mentioned (statistical) inference [50], or construction of a personal interest profile [19].

(3) Comparing Items or Users (Cards 10–13): Relating users or items is one of the most common techniques in RS design [25, 37]. Such techniques, e.g. the commonly used *collaborative filtering*, are apparently well understood by users. In many prior mental models identified, the similarity between users or items was mentioned or played a central role [19, 36].

(4) Presenting Recommendations (Cards 14–16): While the form of presenting recommendations seems to play an inferior role in users' mental models [36], it is very relevant for RS research [23, 48]. We thus decided to also include three actions for the presentation of RS outcome.

To further diversify answers and to enable analysis of the extent to which the mental models of participants diverge from a "*ground truth*", we added 12 *distractor cards*. These cards were chosen as misconceptions of RS as well as actions that are not part of the main personalization process. Distractor cards were collected by identifying such actions in results of a previous qualitative user study to which we had access (i.e. [36]). All distractor cards can be found in the supplementary material. Examples are: "*Employees suggest items for me*", "*Evaluating my satisfaction of recommendations*", and "*Blocking advertisement*".

Finally, we added *question mark* and *open cards*. The question mark cards account for uncertainties in participants' mental models, i.e. to indicate that there might be an unknown action performed in a certain step. Open cards account for any missing actions that were not part of the pre-labeled action or distractor cards, but are part of the subjective mental model.

4 User Study

Our online study consisted of three parts: instruction, mental model task, and measurement of technical knowledge and perception of RS. At the end, partici-

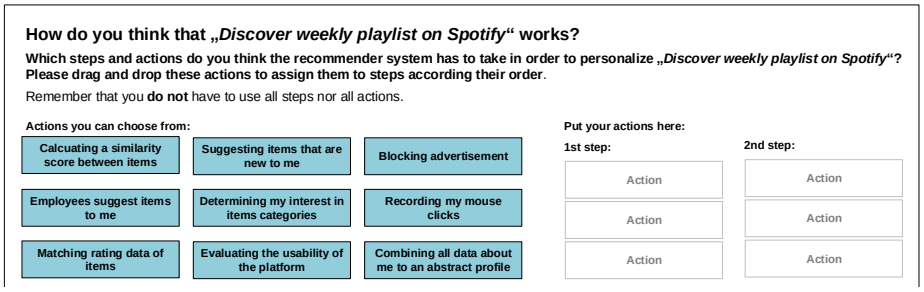


Fig. 1: Excerpt of the mental model task with shortened description and exemplified for *Discover weekly playlist on Spotify*. For reasons of space efficiency only nine actions and only two steps with three action slots are depicted here.

pants were debriefed and received 2.76 $ as compensation. We used Soscisurvey[2] as a survey platform in which we implemented the card sorting setting ourselves. On average participants took 13:39 minutes ($SD = 01{:}55$) to complete the study. This study was approved by the local ethics committee of the University of Duisburg-Essen. We included the complete lists of measures and items in the supplements. This section is organized according to the three parts of the user study.

4.1 Instruction

At the beginning, participants were presented with a definition of RS and the term "item", which we defined as all content subject to recommendations, whether it is a product on Amazon, or a person suggested as friend on Facebook. Participants chose a RS they encounter regularly. Eight options were provided: *Top pics for you on Netflix*, *Video recommendations on YouTube*, *Discover weekly playlist on Spotify*, *Recommendations of similar items on Amazon*, *Friend recommendations on Facebook*, *Trending hashtags for you on Twitter*, *Personalized feed on Instagram*, and *Daily news recommendations on Google News*.

Additionally, participants could opt for "*None of the above*", which resulted in an immediate end of this participant's session. If any of the eight options was chosen, participants were instructed to keep the chosen RS and its items in mind as point of reference for all subsequent questions. As auxiliary reminder, their chosen RS was also explicitly displayed in several texts throughout the survey.

4.2 Mental Model Task

Next, participants had to complete the card sorting task described in Section 3. They were briefed to use their RS chosen in the previous part as reference while sorting their cards. All 35 cards were displayed on the left and participants were

[2] https://www.soscisurvey.de

asked to sort as many of them as they deem appropriate via drag-and-drop in up to seven steps. The steps were displayed on the right (see Figure 1). The open and question mark cards were shown at the bottom of the card list. Action and distractor cards were presented in a randomized order.

After the task, participants were asked about the *degree of fidelity* that reflected how well participants were able to express their mental model. We measured this using two self-created items on a 5-point Likert scale (1 ("I strongly disagree") to 5 ("I strongly agree")). The items were: "*I was able to express my ideas through the arrangements of steps and actions very well*" and "*I feel very certain about the arrangement of steps and actions.*", (Cronbach's $\alpha = .725$).

4.3 Measures

We asked participants about their perception of RS and technical knowledge: on the one hand, through self-created items on technical or metaphorical perception of RS, and, on the other hand, through standardized scales for social presence, trusting beliefs, transparency, and other user-centric measures of RS.

Perception of the RS: To assess whether participants perceived the chosen RS as rather technical or metaphorical, we included a self-created semantic differential consisting of twelve pairs such as "*machinelike*" vs. "*humanlike*" (Cronbach's $\alpha = .809$). Items were assessed on a 5-point Likert scale.

We used the social presence scale from Gefen and Straub [17] consisting of 5 items (e.g. "*There is a sense of human contact in the system.*"). Furthermore, we assessed *trusting beliefs* using items from McKnight et al. [32]. Trusting beliefs consist of three dimensions: benevolence, integrity, and competence. For all of these scales, items were rated on a 7-point Likert scale.

We measured *transparency*, *control*, and *perceived usefulness* using the *ResQue* inventory [41], and added *recommendation quality* and *perceived system effectiveness* from [24]. All items were assessed on a 5-point Likert scale.

Technical Knowledge of RS: We assessed the prior *technical knowledge* of participants by using three self-created items, e.g. "*In the past I learned about how recommender systems work*" (Cronbach's $\alpha = .818$). Additionally, we specifically asked for the *confidence* in the capability of learning about RS through one item, ("*I would be capable of understanding the recommendation process, if someone would explain it to me.*"). All items were measured on a 5-point Likert scale.

4.4 Participants

In total, 170 participants were recruited through the UK-based crowd-working platform *Prolific*[3]. Participants' age ranged from 18 to 67 ($M = 31.42$, $SD = 11.64$). Regarding gender, 71 participants identified as male and 99 as female.

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

The sample was rather educated with 75 participants (44.1 %) holding a bachelor's degree, 55 participants (32.4 %) holding a high school diploma, and 26 (15.3 %) a master's degree. Six participants (3.5 %) held a PhD, while three participants (1.8 %) reported to hold less than a high school diploma. Five participants (2.9 %) indicated other degrees. Participants reported to have a low to moderate technical knowledge on RS ($M=1.87$, $SD=.99$).

Generally, participants were able to express their mental model through the task well: Descriptive analysis revealed a moderate degree of fidelity with a mean score of 3.18 ($SD=0.90$). Question mark cards were used very rarely (on average participants used $M=0.03$ ($SD=0.06$) of them). Only few open cards[4] were used: Participants created 63 cards themselves accounting for 2.32 % of all cards used. Most of them indicated similar ideas as existing action cards, e.g. "*Collecting other data such as gender*", or were specific to the RS, e.g. "*Monitoring what I watch*". 25.40 % of them were left blank or were unclear in their meaning.

Overall, participants used $M=15.74$ ($SD=8.20$) cards and $M=4.90$ ($SD=1.76$) steps to represent their mental model. When comparing action cards and all other cards, a t-test for paired samples revealed that action cards ($M=9.85$, $SD=4.15$) were used significantly more often than the others ($M=5.88$, $SD=4.80$), $t(169)=14.94$, $p=.001$. The proportion of actions to distractors was at 26.61 % ($SD=12.70$ %) on average, i.e. for each four action cards that were used in the mental model task, one was a distractor.

5 Results

We followed a *data-driven* approach to answer our RQs. Hierarchical clustering on participants' card sorts revealed three distinct user groups in our data. We conducted a descriptive analysis to compare the perceptions of RS of these groups. For the analyses we used SPSS 25 and R 4.0.2.

5.1 RQ1: Which Different Mental Models Do Users Hold across RS?

To determine clusters among the different mental models expressed, we first calculated dissimilarities between card sorts. While card sorts are commonly evaluated this way, we faced two specific challenges in our task setting: (1) The order of steps, the cards were sorted in, was relevant to us, which is not taken into account by typical dissimilarity measures (e.g. *Jaccard Index*). (2) Participants were free to use any number of steps (up to a maximum of 7) and any number of cards (up to a maximum of 35), which resulted in many missing values. To overcome these two challenges, we calculated the dissimilarity between participants as follows:

$$dis(p,u)=0.7*d(p,u)+0.3*q(p,u)$$

[4] Note that a qualitative in-depth analysis of open cards was not within the scope of this work.

Table 2: Overview of user groups descriptive statistics. SI values refer to the cluster cut within each group.

Group	N	No. of cards	No. of steps	Degree of fidelity	No. of clusters	SI
		$M(SD)$	$M(SD)$	$M(SD)$		
1	66	8.53 (2.56)	4.09 (1.84)	3.34 (0.88)	4	0.298
2	79	16.76 (3.39)	4.99 (1.42)	3.46 (0.85)	2	0.238
3	25	31.60 (3.20)	6.76 (0.52)	2.82 (0.99)	7	0.132

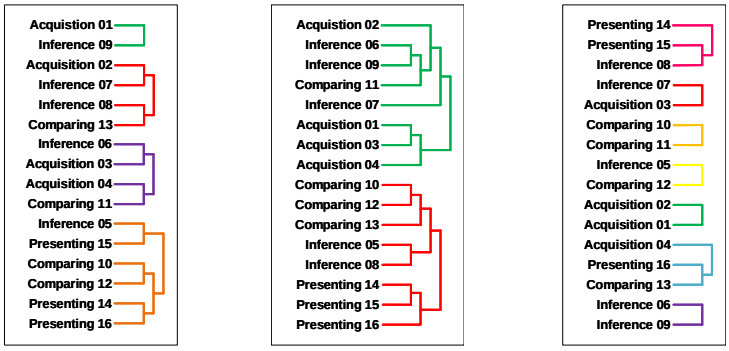
This dissimilarity calculation is based on two components. The first one ($d(p, u)$) determines the normalized *Manhattan distance* between any two participants. We interpret each participant's card sort as vector $p \in \mathbb{N}^c$, where c is equal to the number of available cards[5]. Each position of p, thus, corresponds to a specific card, while the value indicates the number of the step this participant assigned the card to. The Manhattan distance between these vectors accounts for challenge (1) as it considers the order of steps cards are sorted in. While this could be achieved with other similarity measures (e.g. *Euclidean distance*), the Manhattan distance treats coordinates as discrete, thus matching the discrete steps of our task design. This first component only includes cards that were used by both participants. Therefore, to account for (2), we add a second component ($q(p, u)$) as the difference of how many cards both participants used.

We acknowledged that both components should not equally contribute to the dissimilarity and deemed the step order as more important than the number of cards each participant used. Thus, we assigned different weights to each component and chose a factor of 0.7 for the first, and a factor of 0.3 for the second component. Detailed description of the formulas is included in the supplement.

Hierarchical Clustering Hierarchical clustering can follow a *divisive* or an *agglomerative* clustering algorithm. Divisive clustering follows a top-down pattern, which starts with one cluster containing all items and divides them iteratively until each cluster contains only one single item. Agglomerative clustering takes the opposite approach and starts with each item as an own cluster and iteratively combines them until only one cluster remains [22].

We compared the *clustering coefficients* of divisive and agglomerative variants. This coefficient "describes the strength of the clustering structure" [22]. A coefficient closer to 1 indicates a stronger cluster structure and a better fit with the data. In our case, agglomerative clustering in tandem with the *Ward's* criterion [51, 35] resulted in the best performance with a *clustering coefficient* of .949. To determine the number of clusters that fits the data best, we then compared cuts of the hierarchy at 2–7 clusters. For this, we used the respective *average silhouette index* (SI) [43] which reflects the *cohesion* within clusters and *separa-*

[5] We ignored open and question mark cards, since they cannot be compared easily.



(a) Group 1. (b) Group 2. (c) Group 3.

Fig. 2: Dendrograms depicting clusters of how cards have been sorted for each identified user group. Clusters and actions within are ordered regarding the median of steps, they have been sorted in.

tion between clusters. The index ranges from −1 to 1. We found the highest SI of .237 when cutting at 3 clusters, and thus, 3 user groups.

Subsequently, we performed hierarchical clustering again. This second clustering was applied to cards within each of the 3 user groups and resulted in 2–7 clusters, depending on the group (Figure 2 and Table 2). Below we describe the mental models of each user group in detail.

Group 1: Users with a Parsimonious Concept-Based Mental Model This group used the lowest number of cards and steps (Table 2). Participants were convinced of their card sorts (degree of fidelity). Compared to the other groups, they expressed less prior knowledge in RS, but felt confident in understanding them. The group perceived RS as rather rational, planned, and machine-like (Table 3). The dendrogram of this group shows four major clusters (Figure 2a) and that this group held a rather *concept-based* mental model.

The first major cluster is small and pertains to elicitation and analysis of implicit and less tangible user data ("recording of mouse clicks" (card 01) and "analyzing content of items" (card 09)) which can be considered as a starting point of RS processes. The second major cluster refers to the *inference of a user model* comprising of several processes including processes of data acquisition, inference, and comparison, all regarding the user (card 02, 07, 08, 13).

Following this, the third major cluster (card 05, 10, 12, 14–16) represents the *processing of the user model*. Further user data, e.g. "mood" and "dwell times", are analyzed and recorded. Additionally, the user data is connected to item data. In contrast to this, the fourth major cluster (card 03, 04, 06, 11) focuses clearly on the *processing of items*. It includes different processes, i.e. inference, comparison, and presentation of items.

In sum, this group was parsimonious in their use of cards, i.e. they only used few actions and steps. Participants of this group focused on the concepts of *the user model*, *the user model processing*, and on the *items*. In each cluster,

Table 3: Overview of perception of RS for each user group.

Variables	$M(SD)$		
	Group 1	Group 2	Group 3
Technical knowledge of RS			
Knowledge of RS	1.56 (0.83)	1.97 (1.03)	2.20 (1.10)
Confidence	4.23 (0.74)	4.30 (0.65)	3.72 (0.84)
Perception of RS			
Technical/ metaphorical	2.74 (0.48)	2.65 (0.62)	3.01 (0.74)
Social presence	3.11 (1.44)	3.29 (1.58)	3.50 (1.62)
Transparency	3.76 (0.79)	4.09 (0.75)	3.88 (0.78)
Trusting beliefs (TB)	3.85 (1.22)	4.04 (1.29)	4.20 (1.38)
TB benevolence	3.39 (1.42)	3.44 (1.60)	3.95 (1.47)
TB integrity	3.61 (1.36)	3.66 (1.50)	4.18 (1.54)
TB competence	4.45 (1.48)	4.88 (1.34)	4.41 (1.57)

processes are mixed (e.g. acquisition and inferences processes regarding the user model, comparisons, inference, and presentation regarding the items).

Group 2: Users with a Feasible Procedural Mental Model This group could express their mental model through the task well. Their prior knowledge was higher than in group 1, but lower than in group 3 (see Table 3). Like group 1, they perceived RS as rational, planned, machine-like, but as more transparent. The card sorting task resulted in two major clusters (Figure 2b).

The first major cluster can be divided into two sub-clusters. The first one (card 02, 06, 07, 09, 11) pertains to the inference of a user model using *contextual* user data, such as the "age" or "mood" of the user. The second sub-cluster (card 01, 03, 04) pertains to the acquisition of *interaction* data that is dependent on the use of the RS, e.g. "mouse clicks", "dwell time".

The second major cluster consists of three sub-clusters that represent different processes of RS: comparison of items and users (cards 10, 12, and 13), inferences of the user's interest based on items (card 05, 08), and finally the presentation of recommendations (card 14–16).

In sum, this group showed a procedural mental model that reflected our proposed procedure best (Section 3 and Table 1). Only the first major cluster represented a more nuanced understanding of the acquisition of user data which differed from our proposed procedure. Group 2 views the user model as a starting point that is characterized by contextual data, i.e. data that exist prior to the interaction with RS. Thus, they distinguish between contextual and interaction data. The second major cluster represented the last three steps of our proposed procedure accordingly.

This user group seemed to have the most structured comprehension of RS which is indicated by the rather high values for the degree of fidelity, confidence, and transparency.

Group 3: Users with an Extensive Social-Focused Mental Model This group used the highest number of cards and steps (Table 2). The confidence in their card sorts was the lowest, but they expressed a higher knowledge about RS[6]. Group 3 perceived the RS as more empathetic, spontaneous, and human-like. We found this tendency as well when examining the values for social presence and trusting beliefs, which were the highest in this group (Table 3). The dendrogram (Figure 2c) reveals seven major clusters of action cards.

The first major cluster mainly contains the presentation of recommendations, showing items that are new (card 14) and the user might like (card 05). Based on these presented items, additional, invisible data are considered (card 08). The second major cluster combines user information to an abstract profile (card 07), to which user data (e.g. "dwell time", card 03) are added. The third and fourth major cluster mostly pertain to comparison processes of the items (card 10-12) and determination of user interests (card 05).

The fifth cluster refers to the data acquisition of explicit user data (card 01, 02), while the sixth pertains to items in relation to users (e.g. user ratings of items (card 04), similarity between users (card 13), and presentation of what users liked in the past (card 16)). These processes were mostly assigned to step 3. The last cluster refers to inference processes on the "mood that the user is currently in" (card 06) and "analyzing content of items" (card 09).

This group used nearly all cards and steps, i.e. many distractor, open, and question mark cards were used. We conclude that this user group might have an extensive mental model consisting of many different processes that go beyond the recommendation process described in Section 3.

In sum, the mental model of group 3 appears rather unstructured. This is reflected by the high number of clusters (i.e. many small unrelated islands). Like group 1, participants of this group seem to follow a rather concept-based mental model. Yet, they distinctively assigned more human attributes and social presence to the system indicating a higher social focus of their mental models.

5.2 RQ2: How Do these Mental Models Relate to the Perception of RS?

Due to the *exploratory* nature of our approach, we analyzed our results descriptively[7]. First, we explored if we find differences for the RS choice in the measures. The descriptive data revealed that confidence intervals (CI) of means of each chosen RS largely overlap for all measures. This indicates that the results are independent of the particular RS a participant had in mind. Instead, we conclude that measured differences resulted from the particular mental model a participant held.

Then, we analyzed the group differences based on 95 % CI of mean differences and effect sizes (using Cohen's d with pooled standard deviation to account for different group sizes). To this end, we first performed a visual analysis of CI of

[6] However, the level of knowledge in all groups can be considered as low to moderate
[7] An overview of all descriptive data can be found in the supplement.

Table 4: Overview of descriptive analysis.

Group comparisons	Cohen's d	95% CI	Mean diff.	95% CI
Technical knowledge				
Group 1 vs. 3	−.66	[−1.13, −0.19]	.51	[0.10, 0.91]
Group 2 vs. 3	−.83	[−1.29, −0.37]	.58	[0.19, 0.98]
Technical/ metaphorical				
Group 1 vs. 3	.48	[0.01, 0.95]	−.27	[−0.60, 0.07]
Group 2 vs. 3	.55	[0.10, 1.01]	−.36	[−0.69, −0.03]
Transparency				
Group 1 vs. 2	.43	[0.10, 0.76]	−.33	[−0.64, −0.02]
Knowledge of RS				
Group 1 vs. 2	.43	[0.10, 0.77]	−.41	[−0.81, −0.02]
Group 1 vs. 3	.70	[0.23, 1.17]	−.64	[−1.19, −0.09]

group means for each measure. We only report results with moderate to large effect sizes and CI with little or no overlap (Table 4).

Regarding *confidence*, we found that group 3 was less confident than group 1 and 2. This indicates that users with a social-focused mental model were less confident in their capabilities to understand the RS. The analysis revealed the same pattern regarding *technical vs. metaphorical perception* suggesting that group 3 tended to view RS as more human-like than the other two groups. Concerning *transparency*, we found a difference between group 1 and 2 indicating that the procedural mental model might be associated with higher transparency perception. Finally, regarding *knowledge of RS*, we found that group 1 expressed lower knowledge than group 2 and 3.

The descriptive analysis suggests that the precision of the measures were low. Therefore, the results give first indications of relevant relationships between the structure of mental models and RS perceptions.

6 Discussion

This work extends the existing research body on the measurement of mental models through a novel card sorting setting. While it does not investigate single mental models in detail, as fully qualitative methods would, our approach allows for relevant analytical insights. We analyzed *the diversity of mental models* in a large sample. Thus, we envision our card sorting setting as beneficial in a second research stage, after a first general mental model was already revealed.

In line with prior work of Norman [38], who observed the transfer of mental models from one system to another, we found that mental models exist *across systems*. Interestingly, we did not find any relationship between the referenced RS and the perception of RS, i.e. they were independent of another. In fact, differences in users' perceptions of RS were only dependent on users' mental model. We conclude that mental models appear to be more critical for the perception

of RS than the system itself. Hence, for contemporary user-centered design of RS, we suggest a shift from system-focused to mental-model-focused research.

In the following, we discuss the mental models of RS and their relation to the perception of RS. Furthermore, we address RQ3 (*Based on the identified mental models, which implications can be derived from them for the design of transparent intelligent systems?*) and discuss practical implications for the development of more user-friendly, trustworthy, and transparent user interfaces.

6.1 Seeing Is not Understanding

Many participants perceived the referenced RS (e.g. Youtube, Netflix, Spotify) as transparent. However, the transparency perception cannot be ascribed to a factual knowledge about the inner workings of these RS. Firstly, because these systems do not provide any sophisticated explanatory components and, secondly, because participants reported a low to moderate technical expertise of RS. We therefore attribute the transparency perceptions to participants' mental models, which are based on subjective explanations of how the RS work. These explanations, hence, merely form an *impression* of understanding that may not match the actual systems' functioning. In other words, "*seeing*" a system does not necessarily translate to *understanding* it [2]. We argue that such mental models, based on vague information of how the system works, may result in a gap between actual system behavior and users' expectations—a concept known as *gulf of evaluation* [39]. Such gulf was observed to result in false assumptions and erroneous behavior [34, 33]. Morris [33] found that social media users can misinterpret the opaque algorithms responsible for composing their news feed. In the case of Morris' observation, this led to the negative public misperception that new mothers post excessively about their newborns when in fact they do not. Muramatsu and Pratt [34] could show that false assumptions and erroneous mental models can be corrected through transparency.

Practical Implication: Dare to Provide Transparency to Users To avoid a false sense of understanding a system, typical straightforward explanatory components might be too shallow to provide "real" transparency in terms of an actual user comprehension. Thus, users' mental models need to be regarded, evaluated, and, if flawed, corrected by providing factually accurate insights into the system's inner working. While we note that such a correction could benefit from knowing the active user's mental model during runtime, it could also be based on a general elicitation of mental models prevalent in a user base. The presented study demonstrates how such elicitation could be performed. Yet, we acknowledge that further research in eliciting mental models and providing transparency of RS is necessary as intelligent systems become increasingly sophisticated.

Previous research has indicated users' interest in more algorithmic transparency, e.g. [15, 31]. Our study extends on that: It highlights that there is not only a user *interest*, but also that users feel *confidence* in their ability to understand intelligent systems when appropriate explanations are present. This is

especially interesting considering the low technical knowledge of our participants. We, thus, encourage developers to dare to provide sophisticated components of transparency, e.g. in form of explanations [7, 21] or visualizations [5, 28, 29].

6.2 Procedural vs. Concept-Based Mental Models

We could uncover three different mental models of RS that coexist in a large sample of RS users. We observed that these models exhibited different structures and perceptions of RS. Concept-based and procedural mental models were the most prevalent models that co-existed in our sample. An extensive and social-focused mental model was held by a minority of the participants.

The mental model of group 2 reflected the procedure, that our method was based on, best. Due to the opacity of RS, we cannot claim this procedure to be a *ground truth* of RS. Yet, it is based on established publications of researchers and practitioners in the field of RS and we deem it—to a certain degree—accurate. In this regard, group 2, interestingly, felt the highest degree of fidelity in expressing their mental model through the card sorting. Based on this, we assume that the mental model of this group was rather well-defined. Therefore, they perceived the highest transparency of RS. The well-defined mental model might also be the cause for the highest competence perception: The RS was perceived as reasonable leading to comprehension of the system and appreciation of its competence. As this group expressed low technical knowledge of RS, we conclude a close connection between a well-defined mental model, understanding the actual system functioning, the transparency and competence perception of the system.

Group 1 and 3 did not strongly adhere to a process-based mental model. It seems that they did not use the steps in a chronological, but in a concept-wise manner. Inspection of the clusters in the dendrogram of group 1 (Figure 2a) showed that many clusters consisted of actions from different chronological stages. The second cluster, for instance, comprised of four cards (02, 07, 08, 13) of which three cards belong to another chronological stage. Yet, they shared a conceptual focus: the user model. The most frequent and strict concepts in the mental models of group 1 and 3 were *item- vs. user-based recommending*.

Practical Implication: Increase Transparency through Procedural Explanations We conclude that there are several perspectives on a RS that users can adopt. Delivering different user interfaces to each of these groups might address this issue best. For users adhering to a procedural mental model, explanations that emphasize the chronology of the recommendation process can be useful. To prevent aforementioned false assumptions, we suggest great care that explanations reflect the actual recommendation process as closely as possible.

Users that adhered to a concept-based mental model perceived lower transparency. Hence, we suggest explaining the concepts more clearly to those users, i.e. practitioners could provide clear definitions and examples of explicit and implicit user data and explain their application in RS. Similarly, practitioners can stress clearly whether users or items are compared to generate recommendations. The latter was recently identified to cause confusion for users [36].

However, we acknowledge that treating each user group differently is not always possible, e.g. when no information on the active user is available. While our quantitative approach could be used to correlate mental models to user interaction data (e.g. mouse movements), thus forming a baseline for inferring the user's mental model during runtime, this demands further studies. Yet, in our study, we identified some procedural aspects in all user groups and are thus confident that a procedural perspective could be "imposed" on users with a more concept-wise mental model. Hence, we recommend considering procedural explanations in RS. Apart from matching most user expectations, our findings suggest that this form of explanation also results in a higher perceived transparency.

6.3 Technical vs. Humanized RS

While group 1 and 2 held a rather technical understanding of RS (rational, and machinelike), group 3 described them as neutral to metaphorical (empathetic, spontaneous, and humanlike). Thus, group 3 *humanized* the RS more than the other groups, i.e. they ascribed humanlike characteristics to a non-human agent. This humanization acts as a mechanism to combat uncertainty and situations in which a system seems unpredictable [14]. This effect might be at work here: Besides the more humanized mental model compared to the other groups, group 3 expressed low confidence in the ability to learn about the system.

Prior work in autonomous vehicles has indicated a link between humanization and more trust in the non-human agent [52]. Our study shows that this mechanism might also occur in intelligent systems: group 3 perceived higher levels of trusting beliefs. Furthermore, descriptive values indicate a higher social presence for group 3. We ascribe this also to the more metaphorical and humanized mental model of this group. In line with prior work [8, 27], this social presence may act as mediator between humanization and trusting beliefs in group 3.

Practical Implication: Educate Users and Create Social Presence Uncertain users might hold an unstructured mental model including metaphorical concepts. As a consequence, such user groups might perceive the system as unpredictable and tend to humanize it. From this, we derive two implications for practitioners: (1) There is a need to educate uncertain users, so that they do not need to develop metaphorical or humanized mental models. As a result the system could be perceived as more predictable and transparent. Yet, we also note that some desirable aspects may arise from a higher social presence of RS and thus, (2), suggest to include social aspects into a user interface. This could, for instance, be realized by adding elements that express metaphors or using a metaphorical language. We, however, note that this is speculative and emphasize the necessity of investigating these aspects in greater depth.

6.4 Limitations

We created the cards as carefully as possible and added open cards to formulate new actions. Still, some actions that participants created were redundant with

our pre-formulated cards. Therefore, we assume that some participants did not read all cards or did not fully understand them. Thus, we deem 35 cards as maximum in such settings and reconsider wording choices. Another limitation of our study concerns our task setting. While participants were able to express procedural mental models well, this did not necessarily apply to other forms of mental models (although participants managed to express them anyway, see Section 6.2). We conclude that the task design could be slightly adjusted to, for instance, express parallel actions or feedback loops. This could, for instance, be achieved through concept networks or flow diagrams. We also acknowledge that we have included only a small fraction of all existing RS in our study and that RS represent only one facet of the full range of intelligent systems. Future work might investigate mental models of additional RS and other intelligent systems.

7 Conclusions & Future Work

We introduced a method that enables us to identify mental models quantitatively and to examine their diversity in large samples and across platforms. It poses a substantial extension of prior research on mental models of intelligent systems which relied on qualitative studies with small samples.

We could reveal a relation between mental model structures and user perception of RS: Procedural mental models were positively related to transparency, implying that transparency can be increased through procedural explanations. Such type of explanations could also be imposed on users who hold a concept-based mental model. Additionally, uncertain users might hold social-focused mental models and perceive RS as more humanlike, which leads to ambivalent results: While social-focused mental models might positively relate to trust, they might lead users to be less confident and perceive a system as unpredictable.

Finally, this study highlights that mental models exist *across systems*, i.e. the perception of RS mainly depends on the mental models, and not on the particular system. We consequently emphasize the relevance of mental models for designing user-friendly intelligent systems and advocate a shift from system-focused to mental-model-focused research in that area.

Our method allows to identify mental model in statistically representative user studies, and thus, to make generalizable inferences about the mental models in a target audience and their relations to system perceptions. Moreover, we suggest an analysis of the relationship between user characteristics (e.g. personality traits such as need for cognition) and mental models of intelligent systems. Our method could be used to identify user groups that relate to certain personality profiles. This could contribute to measuring a user's mental model during runtime, enabling presentation of personalized transparency components, tailored towards their mental model and personality. This might be especially useful for system applications that require long-term relation between a user and a system.

References

[1] Adomavicius, G., Tuzhilin, A.: context-aware recommender systems. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 191–226, Springer US, Boston, MA (2015)

[2] Ananny, M., Crawford, K.: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society **20**(3), 973–989 (2018)

[3] Beliakov, G., Calvo, T., James, S.: Aggregation functions for recommender systems: Recommender systems handbook. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 777–808, Springer US, Boston, MA (2015)

[4] Bussolon, S., Russi, B., Missier, F.D.: Online card sorting: As good as the paper version. In: Proc. of the 13th Eurpoean conference on Cognitive ergonomics: trust and control in complex socio-technical systems, ECCE '06, ACM, New York, NY, USA (2006)

[5] Cardoso, B., Brusilovsky, P., Verbert, K.: Intersectionexplorer: the flexibility of multiple perspectives. In: Proc. of the 4th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, pp. 16–19, IntRS 2017, CEUR Workshop Proceedings (2017)

[6] Castelo, N., Bos, M.W., Lehmann, D.R.: Task-dependent algorithm aversion. Journal of Marketing Research **56**(5), 809–825 (2019)

[7] Cheng, H.F., et al.: Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In: Proc. of the 2019 Conference on Human Factors in Computing Systems, pp. 559:1–559:12, CHI '19, ACM, New York, NY, USA (2019)

[8] Choi, J., Lee, H.J., Kim, Y.C.: The influence of social presence on evaluating personalized recommender systems. In: Pacific Asia Conference on Information Systems, p. 49, AISeL (2009)

[9] Conrad, L.Y., Tucker, V.M.: Making it tangible: hybrid card sorting within qualitative interviews. Journal of Documentation **75**(2), 397–416 (Mar 2019)

[10] Cooke, N.J.: Varieties of knowledge elicitation techniques. International Journal of Human-Computer Studies **41**(6), 801–849 (Dec 1994)

[11] Cramer, H., et al.: The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-Adapted Interaction **18**(5), 455 (Aug 2008)

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

[13] Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., Hussmann, H.: Bringing transparency design into practice. In: 23rd International Conference on Intelligent User Interfaces, pp. 211–223, IUI '18, ACM (2018)

[14] Epley, N., Waytz, A., Cacioppo, J.T.: On seeing human: A three-factor theory of anthropomorphism. Psychological Review **114**(4), 864–886 (2007)

[15] Eslami, M., Vaccaro, K., Lee, M.K., Elazari Bar On, A., Gilbert, E., Karahalios, K.: User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In: Proc. of the 2019 Conference on Human Factors in Computing Systems, p. 1–14, CHI '19, ACM, New York, NY, USA (2019)
[16] French, M., Hancock, J.: What's the folk theory? reasoning about cyber-social systems (2017), URL https://ssrn.com/abstract=2910571
[17] Gefen, D., Straub, D.W.: Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services. Omega **32**(6), 407–424 (Dec 2004)
[18] Gero, K.I., et al.: Mental Models of AI Agents in a Cooperative Game Setting. In: Proc. of the 2020 Conference on Human Factors in Computing Systems, pp. 1–12, ACM, Honolulu HI USA (Apr 2020)
[19] Ghori, M.F., Dehpanah, A., Gemmell, J., Qahri-Saremi, H., Mobasher, B.: Does the User Have A Theory of the Recommender?A Pilot Study. In: Proc. of Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS '19), p. 9, ACM, Copenhagen, DK (Sep 2019)
[20] Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining Collaborative Filtering Recommendations. In: Proc. of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 241–250, CSCW '00, ACM, New York, NY, USA (2000)
[21] Hernandez-Bocanegra, D.C., Donkers, T., Ziegler, J.: Effects of argumentative explanation types on the perception of review-based recommendations. In: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, p. 219–225, UMAP '20 Adjunct, ACM, New York, NY, USA (2020)
[22] Kaufman, L., Rousseeuw, P.J.: Finding groups in data: an introduction to cluster analysis. Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ (1990), URL https://cds.cern.ch/record/1254107
[23] Knijnenburg, B.P., Willemsen, M.C.: Evaluating recommender systems with user experience. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 309–352, Springer US, Boston, MA (2015)
[24] Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. User Modeling and User-Adapted Interaction **22**(4), 441–504 (2012)
[25] Koren, Y., Bell, R.: Advances in collaborative filtering. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 77–118, Springer US, Boston, MA (2015)
[26] Kulesza, T., Stumpf, S., Burnett, M., Kwan, I.: Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In: Proc. of the 2012 Conference on Human Factors in Computing Systems, pp. 1–10, CHI '12, ACM, Austin, Texas, USA (2012)
[27] Kunkel, J., Donkers, T., Michael, L., Barbu, C.M., Ziegler, J.: Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In: Proc. of the 2019 Conference on Human Factors in Computing Systems, pp. 1–12, CHI '19, ACM, New York, NY, USA (2019)

[28] Kunkel, J., Loepp, B., Ziegler, J.: A 3d item space visualization for presenting and manipulating user preferences in collaborative filtering. In: Proc. of the 22nd International Conference on Intelligent User Interfaces, pp. 3–15, IUI '17, ACM, New York, NY, USA (2017)
[29] Kunkel, J., Schwenger, C., Ziegler, J.: Newsviz: Depicting and controlling preference profiles using interactive treemaps in news recommender systems. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 126–135, UMAP '20, Association for Computing Machinery, New York, NY, USA (2020)
[30] Langan-Fox, J., Code, S., Langfield-Smith, K.: Team mental models: Techniques, methods, and analytic approaches. Human Factors **42**(2), 242–271 (2000)
[31] Lim, B.Y., Dey, A.K.: Assessing demand for intelligibility in context-aware applications. In: Proc. of the 11th International Conference on Ubiquitous Computing, p. 195–204, UbiComp '09, ACM, New York, NY, USA (2009)
[32] McKnight, D.H., Choudhury, V., Kacmar, C.: Developing and validating trust measures for e-commerce: An integrative typology. Information Systems Research **13**(3), 334–359 (2002)
[33] Morris, M.R.: Social networking site use by mothers of young children. In: Proc. of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1272–1282, CSCW '14, ACM, New York, NY, USA (2014)
[34] Muramatsu, J., Pratt, W.: Transparent queries: Investigation users' mental models of search engines. In: Proc. of the 24th Annual International Conference on Research and Development in Information Retrieval, p. 217–224, SIGIR '01, ACM, New York, NY, USA (2001)
[35] Murtagh, F., Legendre, P.: Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? Journal of Classification **31**(3), 274–295 (2014)
[36] Ngo, T., Kunkel, J., Ziegler, J.: Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study. In: Proc. of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 183–191, ACM, Genoa Italy (Jul 2020)
[37] Ning, X., Desrosiers, C., Karypis, G.: A comprehensive survey of neighborhood-based recommendation methods. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 37–76, Springer US, Boston, MA (2015)
[38] Norman, D.A.: Some Observations on Mental Models. In: Gentner, D., Stevens, A.L. (eds.) Mental Models, pp. 7–14, Psychology Press, New York, NY, USA (1983)
[39] Norman, D.A.: The design of everyday things. Basic Books, Inc., New York, NY, USA (1988), ISBN 978-0-465-06710-7
[40] Prahl, A., van Swol, L.: Understanding algorithm aversion: When is advice from automation discounted? Journal of Forecasting **36**(6), 691–702 (2017)
[41] Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proc. of the fifth ACM conference on Recommender systems - RecSys '11, p. 157, ACM, Chicago, Illinois, USA (2011)

[42] Ricci, F., Rokach, L., Shapira, B.: Recommender systems: Introduction and challenges. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 1–34, Springer US, Boston, MA (2015)
[43] Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987)
[44] Rugg, G., McGeorge, P.: The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. Expert Systems **14**(2), 80–93 (1997)
[45] Rumelhart, D.E., Norman, D.A.: Representation in Memory. No. 116 in CHIP report, University of California, San Diego (1983)
[46] Sparling, E.I., Sen, S.: Rating: how difficult is it? In: Proc. of the Fifth ACM Conference on Recommender Systems, pp. 149–156, RecSys '11, ACM, New York, NY, USA (2011)
[47] Torkamaan, H., Barbu, C.M., Ziegler, J.: How Can They Know That? A Study of Factors Affecting the Creepiness of Recommendations. In: Proc. of the 13th ACM Conference on Recommender Systems, pp. 423–427, RecSys '19, ACM, New York, NY, USA (2019)
[48] Tsai, C.H., Brusilovsky, P.: Beyond the ranked list: User-driven exploration and diversification of social recommendation. In: Proc. of the 23rd International Conference on Intelligent User Interfaces, pp. 239–250, IUI '18, ACM, New York, NY, USA (2018)
[49] Tsai, C.H., Brusilovsky, P.: Explaining recommendations in an interactive hybrid social recommender. In: Proc. of the 24th International Conference on Intelligent User Interfaces, pp. 391–396, IUI '19, ACM, New York, NY, USA (2019)
[50] Tullio, J., Dey, A.K., Chalecki, J., Fogarty, J.: How It Works: A Field Study of Non-Technical Users Interacting with an Intelligent System. In: Proc. of the 2007 Conference on Human Factors in Computing Systems, pp. 31–40, CHI '07, ACM, New York, NY, USA (2007)
[51] Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301), 236–244 (1963)
[52] Waytz, A., Heafner, J., Epley, N.: The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. Journal of Experimental Social Psychology **52**, 113–117 (May 2014)
[53] Xie, B., Zhou, J., Wang, H.: How Influential Are Mental Models on Interaction Performance? Exploring the Gap between Users' and Designers' Mental Models through a New Quantitative Method. Advances in Human-Computer Interaction **2017**, 1–14 (2017)
[54] Yang, R., Shin, E., Newman, M.W., Ackerman, M.S.: When fitness trackers don't 'fit': End-user difficulties in the assessment of personal tracking device accuracy. In: Proc. of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 623–634, UbiComp '15, ACM, New York, NY, USA (2015)
[55] Zhou, J., Chen, F.: 2d transparency space—bring domain users and machine learning experts together. In: Zhou, J., Chen, F. (eds.) Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent, pp. 3–19, Springer, Cham (2018)