Development of an Instrument for Measuring Users' Perception of Transparency in Recommender Systems

Marco Hellmann, Diana C. Hernandez-Bocanegra and Jürgen Ziegler

University of Duisburg-Essen, Forsthausweg 2, 47057 Duisburg, Germany

Abstract

Transparency is increasingly seen as a critical requirement for achieving the goal of human-centered AI systems in general and also, specifically, recommender systems (RS). However, defining and operationalizing the concept is still difficult, due to its multi-faceted nature. Currently, there are hardly any measurement instruments to adequately assess the perceived transparency of RS in user studies. Thus, we present the development of a measurement instrument that aims at capturing perceived transparency as a multidimensional construct. The results of our validation show that transparency can be distinguished with respect to input (what data does the system use?), functionality (how and why is an item recommended?), output (why and how well does an item fit one's preferences?), and interaction (what needs to be changed for a different prediction?). The study is intended as a first iteration in the development of a reliable and fully validated measurement tool for assessing transparency in RS.

Keywords

Recommender systems, transparency, explanations, user study

1. Introduction

The request for more transparency in intelligent systems has become steadily louder in recent years, formulated in academic research as well as in most public and corporate policies concerning the ethics of artificial intelligence [1, 2]. Although there is now broad agreement that transparency is of high relevance for developing human-centred AI systems, the concept is still elusive due to its multi-faceted nature and the different objectives it is intended to serve. The questions raised when asking for transparency include, for example, the system aspects that should be made transparent, or the riskiness of an AI function at an individual or societal level.

A need for greater transparency has also been noted for recommender systems (RS), a frequent, user-facing type of AI-driven technology, to better support users' in their decision-making and to avoid potentially negative consequences, e. g. users getting trapped in filter bubbles [3]. Various methods have been proposed to this end, ranging from disclosing the user profile on which a recommendation is based to providing explicit explanations. Still, the multi-facetedness of the concept makes it difficult to design effective transparent RS. A central question that must be solved to this end is how transparency of a RS can be measured and evaluated. While different aspects of the system, for example, the input data, the recommendation algorithm, or features of the recommended items may be exposed to the user, transparency as a user-centric quality can only be assessed by measuring users' perception and understanding of those system aspects that are relevant for their decision making and trust in the system [4].

Despite the acclaimed relevance of transparency in RS, the instruments available for measuring it from a user perspective are still very limited. Some instruments for assessing overall recommendation quality include a small number of items related to perceived transparency [5], but these measures still seem far from covering the multiple facets involved. To the best of our knowledge, there is no instrument focusing specifically on RS transparency. A further shortcoming of existing instruments is the lack of sufficiently considering the cognitive processes involved in users' understanding of recommendations and in their ability to influence the system according to their needs if such influence is possible.

In this paper, we describe steps towards a more holistic and cognitively grounded psychometric instrument for measuring perceived transparency in RS. We first explain the questionnaire development process that resulted in a validated set of items specifically focused on RS transparency. The candidate items for this development were chosen to reflect the different steps involved in cognitively processing the information provided about the recommendation process and its output. To further validate the instrument, we performed an analysis of the effects of perceived transparency as measured by our new instrument on factors related to trust in the RS and effectiveness of the recommendations. An influence of transparency on users' trust in the system and on the acceptance of the recommendations has been suggested in

Joint Proceedings of the ACM IUI Workshops 2022, March 2022, Helsinki, Finland

marco.hellmann@stud.uni-due.de (M. Hellmann); diana.hernandez-bocanegra@uni-due.de (D. C. Hernandez-Bocanegra); juergen.ziegler@uni-due.de (J. Ziegler)

prior research, e. g., in [5]. We analyzed these influences through structural equation modeling to show that the construct 'transparency' as measured by our instrument has in fact the assumed effects.

Our contribution is thus twofold: we provide a systematically derived and validated measurement instrument for transparency in RS, and we can show that the different transparency factors represented in the questionnaire have an impact on the effectiveness of recommendations and trust in the system, albeit to different degrees.

2. Related work

Users' perception of the transparency of a RS may be influenced by several factors. Providing explanations is one important aspect, and some studies have shown that transparency is positively influenced by the quality of the explanations given ([5], [6]) and that it is related to control over the system [7]. The effect of systematically varied levels and styles of explanation on perceived transparency has been studied and assessed via questionnaires (see e.g. [8], [9], [10]). Also, a positive influence of interaction possibilities as well as perceived control on the perceived transparency of the system was reported by [5]. Transparency perception seems to be enhanced both by the perceived quality of explanations and the perceived accuracy or quality of recommendations. In addition, the authors show a positive effect of transparency on trust and through trust an indirect effect on purchase intentions. According to [11], this can be related to evaluating the effectiveness of the RS. Moreover, studies suggest that perceived transparency promotes satisfaction with the system [12] [7].

The influence of personal factors on the perception of recommender systems has often been investigated in the light of the general decision-making behavior of users (see [13]). [9] showed that individuals with a rational decision-making style trusted the recommender system tested more and rated its efficiency and effectiveness higher. Furthermore, they showed that individuals with an intuitive decision-making style rate the quality of explanations better.

To date, however, few measurement tools exist to quantitatively assess the transparency of a RS as perceived by users. [6] surveyed perceived transparency using two items ("I understand why the system recommended the artworks it did"; "I understand what the system bases its recommendations on"), in the domain of art objects. [14] use a single item ("I did not understand why the items were recommended to me (reverse scale)"), for event recommendations. [8] proposed an item that explicitly refers to explanations: "Which explanation interfaces are considered to be transparent by the users?". [5] proposed an evaluation framework for RS, involving different domains an applications, and formulate the measurement of the construct transparency using only a single item ("I understood why the items were recommended to me"), this latter being a frequently used item for the evaluation of RS transparency.

Consequently, we set out to formulate and validate a more comprehensive way to measure the perceived transparency of a RS, as described in the methods section. The procedure followed the typical procedure for developing psychometric measurement instruments (e.g. [15]):

(1) To operationalize a target construct, first a larger number of candidate items is formulated and compiled. Here, we draw on the basic structure of RS ([16], [17]) and typical user questions related to artificial intelligence algorithms [18]. Second, items were also derived from a qualitative preliminary study, to further analyze the uncertainties in users' mental models, which can be understood as the notion that users have about how a system or a certain type of systems work [19].

(2) We examined the factor structure of the transparency construct, which was formed as a reflective factor in the sense of classical test theory (see also [20]). We considered 4 factors that could group individual questionnaire items, and that might contribute to variances in perceived transparency, inspired on dimensions defined by [18]: Input ("what kind of data does the system learn from"), output ("what kind of output does the system give"), functionality ("how / why does the system make predictions") and interaction (what if / how to be that, "what would the system predict if this instance changes to..").

(3) The developed measurement instrument was validated. For this purpose, the framework model of [7] was used.

2.0.1. Mental models and stages of cognitive processing

Transparency is frequently discussed like an objective property of a system. A system becomes only transparent, however, if its users can understand the transparency-related information, such as explicit explanations, and evaluate it with respect to their goals. The degree of comprehension may depend on the mental model users have about how the system works [21], either based on preconceptions, previous experiences with similar systems, or on the interaction with and perception of the present system [22]. As discussed in [19], mental models that drift considerably from actual system functioning may result in broadening the "gulfs" described by [22]: 1) the gulf of execution, when the user's mental model is inaccurate in terms of how the system can be used to execute a task, 2) the gulf of evaluation, when the output (as consequence of a user's action) differs from what is expected, according to the user's mental model.

To bridge these gulfs, users must process the information provided by the system at different cognitive levels. The items of the proposed questionnaire were formulated to reflect the action levels according to [23]. According to their model, the quality of interaction with the system can be described through a cycle of evaluation and execution. For example, at first, the user may *perceive* the output of the system (e.g., the recommendations and explanations), then *interpret* the information gathered (e.g., how the system works), and thereby *evaluate* the state of the system (e.g., performance of the system and quality of the output). As a consequence, the user formulates goals aiming to achieve with the system or matches their goals with the evaluation of the system (e.g., get more accurate or diverse recommendations). The user then pursues an intention (e.g., improve recommendations), which is translated into planning actions (e.g., change input), which they finally execute. While this cognitive cycle is well-known in the HCI field, it has hardly been applied in the investigation of transparency for AI-based systems.

The authors in [23] assume that there are gaps between the users' goals and their knowledge about the system, and the extent to which system provides descriptions about its functioning (gulfs of execution and of evaluation, as mentioned beforehand). By taking actions to bridge those gaps, (making system functions to match goals, and making the output represent a "good conceptual model of the system that is easily perceived, interpreted and evaluated" [23]), system designers may contribute to minimize cognitive effort by users [23], and to decrease the discrepancy between the mental model of the system and its functioning, which may have an impact on the perception of transparency, as discussed by [19]. We argue then, that a more comprehensive instrument to measure perceived transparency is still needed, so that such impact can be evaluated not only on the basis of general perceived understanding ("I understood why recommended"), but also on the basis of the extent to which output and functionalities that reflect the conceptual model of the system are perceived, interpreted and evaluated by users.

3. Methods

To operationalize the construct of perceived transparency, we conducted the following steps, based on the typical procedure for developing measurement instruments (e.g., [15]): 1. Formulation and compilation of questionnaire items. 2. Examination of items quality and factor structure, based on an online study. 3. Validation of the measurement instrument. We describe each step below.

3.1. Formulation and compilation of questionnaire items

Here, we draw on the basic structure of RS ([16], [17]) and typical user questions to AI algorithms [18]. Candidate items were also chosen to cover different stages of the cognitive action cycle described in related work. Second, items were also derived from a qualitative pre-study, consisting of interviews with users to further analyze the uncertainties in users' mental models [19], in regard to different commercial RS, like Netflix, Spotify or Amazon.

A total of 6 interviews were conducted via video call, with voluntary participants. When selecting the interview partners, care was taken to represent in the sample different age groups and experience with Internet applications. Students and non-students from different age groups (20 to 50 years) were interviewed. Overall, previous exposure to recommender systems was equally strong among all participants. Only one interviewee had lower experience and one interviewee had slightly higher experience.

The aim of the interviews was to capture the experience, perception and evaluation as well as possible questions of users regarding the functionality or transparency of recommender systems. The subjects were asked to explain the functionality of RS from their perspective and to create a corresponding sketch. Following this, uncertainties and possible lack of transparency were discussed. Finally, prototypical explanations from [24] for increasing the perceived transparency were evaluated by the interview partners. The explanations refer differently to the input used, the functionality and the output. In addition, they use different visual forms of representation, e.g. star ratings, profile lines, text. In this way, uncertainties as well as wishes for more transparency by users could be identified. Each question encountered in interviews was directly transformed into one or more items.

A resulting set of 92 items was collected and discussed by the research team, where linguistic revision and elimination of redundancies were also performed. The discussions led to a reduction of the set to 34 items, which were used as input for the online validation described in the next section.

3.2. Online user study

We conducted a user study to examine item quality and factor structure, as described below.

Participants We recruited 171 participants (89 female, mean age 29 and range between 18 and 69) through the crowdsourcing platform Prolific. We restricted the task to workers in the U.S and the U.K., with an approval rate greater than 98%. Participants were rewarded with £1.15.

Time devoted to the survey (in minutes): M=13.2, SD= 7.33.

We applied a quality check to select participants with quality survey responses (we included attention checks in the survey, e.g. "This is an attention check. Please click here the option 'Disagree'". We discarded participants with at least 1 failed attention check, or those who did not finish the survey. Thus, the responses of 17 of the 192 initial Prolific respondents were discarded and not paid. 4 additional cases were removed due to suspicious response behavior, e.g. responding all questions with the same value within the same page. Thus, 171 cases were used for further analysis.

The target sample size was chosen to allow performing CFA analysis. [25], p. 389, recommend a minimum of n>50 or three times the number of indicators. [26], p. 102, recommend a minimum of n>100 or five times the number of indicators. Thus, given that we wanted to evaluate a set of 34 items, the sample size was set to a minimum of 170 participants.

Questionnaires We utilized the set of 34 items resulting of the formulation of items step described above. Additionally, aiming to further validate the final measurement instrument (4.3), we used items from [5] to evaluate perception of control (how much they think they can influence the system), interaction adequacy and interface adequacy, information sufficiency and recommendation accuracy. Furthermore we included items from [7] to evaluate the perception of system effectiveness (construct *perceived system effectiveness*, system is useful and helps the user to make better choices), and of trust in the system [27] (constructs *trusting beliefs* - subconstructs benevolence, integrity, and competence-, user considers the system to be honest and trusts its recommendations; and *trusting intentions*, user willing to share information and to follow advice). We used items described from [28, 29] for explanation quality, and from [30] to evaluate decision-making style. All items were measured with a 1-5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

Procedure Participants were asked to choose a service from five applications, for which they were required to have an active account: Amazon, Spotify, Netflix, Tripadvisor, and Booking. Participants were instructed to open the application, browse it at their own discretion. They were explicitly told to select an item that was relevant to them and which they would actually buy or consume. A real purchasing of items was explicitly not requested. Participants were asked to return to the survey after completing the task and to answer questions about the system they used.

Data analysis We performed an exploratory factor analysis (EFA) to further reduce the initial set of items and a Confirmatory Factor Analysis (CFA) to test internal reliability and convergent validity. Furthermore, we evaluated discriminant validity of the resultant set of items, in relation to other constructs of the subjective evaluation of RS, for example explanation quality, effectiveness and overall satisfaction, according to the frameworks defined by [7] and [5].

4. Results

4.1. Exploratory Factor Analysis (EFA)

The factor structure was exploratively examined, aiming to further reduce the set of items. A total of 5 EFAs with principal axis factor analysis and promax rotation were performed. First, items that did not have a unique principal loading or had a principal loading that was too low (<.40) were removed. In the first 4 EFAs, 11 items were removed based on this criterion. Subsequently, more stringent criteria were used (factor loadings <.50). The guideline values are based on [31]. Thus, 2 items were removed again. Subsequently, a 6-factorial structure resulted, with a total of 21 items and a variance resolution of 62.45%. Reliability of the factors fall in the range 'good' to 'very good' (.782 to .888), as defined by [32]. The internal consistency across all items is .867.

4.2. Confirmatory Factor Analysis (CFA)

Following the exploration of the factor structure, the result obtained was tested for internal reliability and convergent validity using confirmatory analysis. A first CFA was performed, resulting in 8 items with low factor loadings, which were eliminated from the set. Two factors were removed in the process because they did not load on a second-level overall transparency factor. A final CFA with 4 factors was performed (model fit X^2 = 86.997, df = 61, p = .016; X^2/df = 1.426; CFI = .975; TLI = .968; RMSEA = .050; SRMR = .047). Reliability across all items is equal to .884. This model comprises a final set of four factors and 13 items, which are reported in Table 1 along with factor loadings.

The four factors identified can be associated with the concepts *Input*, composed of 3 items, *Output*, also with 3 items, *Functionality* with 5 items, and *Interaction* with only 2 items. Although the initial item set comprised questions for all stages of the cognitive action cycle, after CFA, items related the perception level were only left for the factor *Functionality*, comprising questions about whether users are aware of transparency-related information if provided by the system (e.g.: "The system provided information about how well the recommendations match my preferences"). This factor covers mostly

Table 1
Test results of internal reliability and convergent validity of our proposed transparency questionnaire.

Factor	Items	Cronbach alpha	Factor loading
Input	It was clear to me what kind of data the system uses to generate recommendations.	0.842	0.817
	I understood what data was used by the system to infer my preferences.		0.901
	I understood which item characteristics were considered to generate recommendations.		0.712
Output	I understood why the items were recommended to me.	0.801	0.771
	I understood why the system determined that the recommended items would suit me.		0.794
	I can tell how well the recommendations match my preferences.		0.710
Functionality	The system provided information to understand why the items were recommended.	0.847	0.731
	The system provided information about how the quality of the items was determined.		0.705
	The system provided information about how my preferences were inferred.		0.736
	The system provided information about how well the recommendations match my preferences.		0.696
	I understood how the quality of the items was determined by the system.		0.760
Interaction	I know what actions to perform in the system so that it generates better recommendations	0.888	0.896
	I know what needs to be changed in order to get better recommendations		0.892

perception-related questions. The missing coverage of perception-related items in other factors is likely due to limitations of the systems used for the online study which do not, for example, provide access to the data on which recommendations are based, thus preventing users to become aware of input data . The factor *Output* comprises items related to the interpretation and evaluation stages. The factor *Interaction* has the smallest scope with 2 items and covers only the facets of action planning or action execution. This factor thus describes whether users know which actions they would have to perform if they wanted to receive other recommendations.

4.3. Discriminant validity of measurement instrument

We determined discriminant validity of the instrument in relation to other constructs of the subjective evaluation of RS, for example explanation quality, effectiveness and overall satisfaction, according to the frameworks defined by [7] and [5]. Discriminant validity was assessed using inter-construct correlations (see results in table 2). We found that the squared correlations between pairs of constructs were all less than the value of average variances that are shown in the diagonal, representing "a level of appropriate discriminant validity" [5].

5. Structural Equation Model (SEM)

To explore the relation between the transparency factors assessed by the questionnaire and the effects of perceived transparency on recommendation effectiveness and trust in the system, as well as the impact of factors influencing transparency, we set up a Structural Equation Model (SEM). The model is based on hypotheses we derived from existing research that has shown the positive effect of higher overall transparency on the perception of the recommendations and the overall system. Furthermore, we assumed that transparency is influenced by system-related aspects (accuracy, interaction quality, and explanation quality) as well as by personal characteristics such as decision-making behavior) as described in the related work section. Some of these factors can also be expected to influence perceived control over the system, a construct that may mediate the impact of these factors on transparency perception. This led us to formulating the hypotheses shown in table 3.

In the following, the relationships of the factors in the structural equation model are presented (see fig. 1). Only significant paths with standardized path coefficients are shown. Indirect effects are only considered for the transparency factors relevant here. The final model is shown to have a very good fit: X^2 = 75.767, df = 57, p = .049; X^2/df = 1.329; CFI = .980; TLI = .965; RMSEA = .044; SRMR = .072. The model is thus adequate to describe the relationships in the data set.

Influences on perceived transparency of the system. Transparency with respect to interaction is rated higher when users are more likely to exhibit an intuitive decision-making style (0.186, p <.05) and users report higher perceived control (0.293, p <.001). The latter is increased by the quality of interaction (0.502, p <.001) and explanations (0.341, p <.001). Users thus know better how to influence recommendations when they have more opportunities to interact with the system, and can gather information about the system through explanations as well as through 'trial and error' (indirect: explanation quality →control →Transparency-interaction: 0.100, p < .01; interaction quality →control →Transparency-interaction: 0.147, p < .001).

Similar observations can be made for functionality. Again, transparency is rated higher when users are more likely to exhibit an intuitive decision-making style (0.141, p <.05) and users report higher perceived control (0.261,

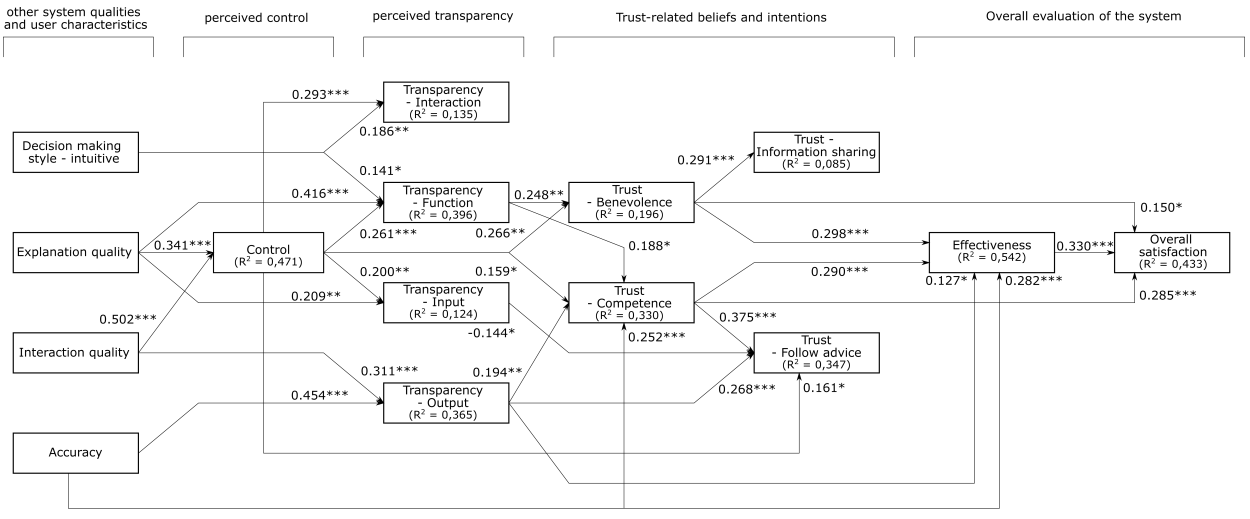
Table 2
Inter-construct correlation matrix. Average Variance Extracted (AVE) on the main diagonal; correlations below the diagonal; quadratic correlations above the diagonal. Target value for AVE $\geq$.5. $p<0.05^{*}$, $p<0.01^{**}$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1 Transp. - input	**0.662**	0.227	0.235	0.136	0.111	0.023	0.009	0.125	0.053	0.051	0.075	0.040	0.065	0.063	0.159	0.002	0.026	0.081	0.021
2 Transp. - output	0,476**	**0.577**	0.231	0.121	0.157	0.039	0.019	0.146	0.187	0.054	0.094	0.291	0.071	0.041	0.239	0.001	0.186	0.240	0.074
3 Transp. - function	0,485**	0,481**	**0.527**	0.155	0.246	0.021	0.061	0.341	0.153	0.022	0.114	0.094	0.147	0.153	0.183	0.021	0.127	0.168	0.094
4 Transp. - interaction	0,369**	0,348**	0,394**	**0.799**	0.119	0.000	0.055	0.048	0.072	0.000	0.056	0.030	0.060	0.106	0.070	0.008	0.064	0.035	0.038
5 Control	0,333**	0,396**	0,496**	0,345**	**0.775**	0.004	0.018	0.242	0.366	0.052	0.154	0.090	0.153	0.198	0.156	0.007	0.144	0.141	0.118
6 DM style - rational	0,153*	0,197**	0.146	0.016	0.061	**0.454**	0.041	0.062	0.004	0.073	0.018	0.032	0.017	0.017	0.026	0.004	0.036	0.058	0.030
7 DM style - intuitive	0.092	0.138	0,246**	0,234**	0.136	-0,203**	**0.502**	0.022	0.027	0.000	0.000	0.019	0.006	0.011	0.012	0.012	0.001	0.001	0.005
8 Explanation quality	0,353**	0,382**	0,584**	0,220**	0,492**	0,248**	0.148	**0.557**	0.091	0.080	0.265	0.151	0.112	0.100	0.230	0.030	0.199	0.177	0.171
9 Interaction adequacy	0,230**	0,432**	0,391**	0,269**	0,605**	0.064	0,163*	0,301**	**0.791**	0.082	0.065	0.048	0.116	0.151	0.101	0.020	0.147	0.118	0.084
10 Interface adequacy	0,226**	0,232**	0.147	0.008	0,228**	0,270**	-0.001	0,282**	0,286**	**0.618**	0.123	0.054	0.052	0.043	0.207	0.020	0.108	0.130	0.187
11 Info. sufficiency	0,273**	0,307**	0,337**	0,236**	0,393**	0.133	-0.001	0,515**	0,254**	0,350**	—	0.104	0.064	0.063	0.182	0.048	0.216	0.188	0.170
12 Recomm. accuracy	0,201**	0,539**	0,307**	0,174*	0,300**	0,180*	0.137	0,389**	0,220**	0,232**	0,323**	—	0.086	0.062	0.259	0.021	0.187	0.326	0.221
13 Trust - benevolence	0,254**	0,266**	0,384**	0,245**	0,391**	0.130	0.079	0,334**	0,341**	0,228**	0,252**	0,293**	**0.666**	0.661	0.366	0.095	0.162	0.332	0.282
14 Trust - integrity	0,250**	0,202**	0,391**	0,326**	0,445**	0.129	0.106	0,316**	0,388**	0,207**	0,251**	0,249**	0,813**	**0.476**	0.332	0.088	0.179	0.238	0.272
15 Trust - competence	0,399**	0,489**	0,428**	0,265**	0,395**	0,162*	0.111	0,480**	0,318**	0,455**	0,427**	0,509**	0,605**	0,576**	**0.608**	0.030	0.278	0.440	0.358
16 Trust - share info.	0.040	0.028	0.146	0.091	0.086	0.060	0.109	0,174*	0.141	0.140	0,219**	0.143	0,308**	0,297**	0,174*	—	0.064	0.062	0.078
17 Trust - follow advice	0,160*	0,431**	0,356**	0,253**	0,379**	0,189*	0.028	0,446**	0,384**	0,328**	0,465**	0,433**	0,402**	0,423**	0,527**	0,252**	—	0.213	0.269
18 Effectiveness	0,284**	0,490**	0,410**	0,187*	0,375**	0,241**	0.036	0,421**	0,344**	0,360**	0,434**	0,571**	0,576**	0,488**	0,663**	0,249**	0,461**	**0.545**	0.389
19 Overall satisfaction	0.145	0,272**	0,306**	0,194*	0,343**	0,174*	0.069	0,414**	0,289**	0,432**	0,412**	0,470**	0,531**	0,522**	0,598**	0,280**	0,519**	0,624**	—

Table 3
Overview of hypothesis addressed in SEM

Hypotheses	Reference	Relevant factor	Explanation
Factors influencing perceived transparency (X →perceived transparency)			
H-1.1	[6],[5]	Explanation quality	Comprehensibility and contribution of the explanations to the understanding of the system
H-1.2	[5]	Accuracy	Match between the items and the user's preferences
H-1.3	[5] (indirect effect)	Interaction quality	Possibilities of adaptation and feedback
H-1.4	[5]	Control	Possibilities of personalization
H-1.5	[10]	decision-making styles	Rational / Intuitive
Effects of perceived transparency (perceived transparency →Y).			
H-2.1	[3],[14]	Trust	Trusting beliefs and intentions
H-2.2	[11]	Effectiveness	Usefulness of the system
H-2.3	[14], [12]	Overall satisfaction	Satisfaction with the system

p <.001). The quality of interaction, promoting perceived control, has a positive effect on transparency concerning how the system works (indirect: interaction quality →control →Transparency-functionality: 0.131, p < .001). It both indirectly and directly (0.416, p <.001) increases the transparency of the functionality when users rate explanations positive (indirect: explanation quality →control →Transparency-functionality: 0.089, p < .01).



Fit indices: X^2 = 75.767, df = 57, p = .049; X^2/df = 1.329; CFI = .980; TLI = .965; RMSEA = .044; SRMR = .072

Figure 1: Structural model. $p<0.05^{*}$, $p<0.01^{**}$, $p<0.001^{***}$

The input is perceived as more transparent the better users can interact with the system. Thus, here again, perceived control has a direct positive effect (0.200, p <.01). The quality of the interaction thus repeatedly has an indirect effect (indirect: interaction quality →control →Transparency-input: 0.100, p < .05). Similarly to what has already been shown with regard to functionality, the quality of the explanations also has a direct, positive effect on transparency of the input (0.209, p <.01) in addition to the indirect effect (explanation quality →control →Transparency-input: 0.068, p < .05).

The transparency of the output shows how well users can assess why a recommendation is made or should match the user's preferences. This is directly increased by the quality of the interaction with the system (0.311, p <.001), i.e. when possibilities are offered or used to indicate one's own preferences. On the other hand, there are no direct or indirect influences of the explanations. Instead, the accuracy of the recommendation has a positive influence on the transparency of the output (0.454, p <.001). Accordingly, the output is easier to understand if it is rated as suitable. Unsuitable recommendations would thus be more difficult for the user to comprehend.

As shown, transparency is positively influenced by the quality of explanations, accuracy of recommendations, opportunities for interaction, and perceived control. Hypotheses 1.1, 1.2, 1.3 and 1.4 can thus be considered confirmed. The influence of the decision-making style is limited to the intuitive style. Therefore, hypothesis 1.5 can only be partially confirmed.

Effects of perceived transparency of the system. No effects can be observed for transparency with regard to interaction. It is possible that effects exist on factors that were not surveyed in this study. For the other transparency factors, however, significant positive effects can be observed.

Transparency regarding the functionality has the strongest and most diverse effect. If users can understand the internal mechanisms, they trust the recommendation system more. Direct positive effects can be observed on benevolence (0.248, p <.01) and trust in the competence (0.188, p <.05) of the system. Indirectly, such transparency thus contributes to a better evaluation of the system's effectiveness (indirect: Transparency-functionality →Trust-benevolence →effectiveness: 0.074, p <.01; indirect: Transparency-functionality →Trust-competence →effectiveness: 0.055, p <.05). Via the increase in effectiveness, overall satisfaction with the system is also promoted (indirect: Transparency-functionality →Trust-benevolence →effectiveness →overall satisfaction: 0.024, p <.05). Via the increase in perceived benevolence, the willingness to share information about oneself is also increased (indirect: Transparency-functionality →Trust-benevolence →Trust-information sharing: 0.072, p <.05). Moreover, via trust in competence, the willingness to follow the advice of the recommendation system is increased (indirect: Transparency-functionality →Trust-competence →Trust-follow advice: 0.071, p <.05). Thus, it is clear that an understanding of the internal mechanisms of recommender systems leads to trusting beliefs and thus to trusting actions and a positive overall evaluation.

Transparency with regard to the input has a negative effect. If users can see which data is used, this has a negative effect on the willingness to follow the advice of the recommendation system in this model (-0.144, p <.05). Thus, this shows a certain counterbalance to a transparent functionality, possibly triggered by too much information or a general distrust regarding data privacy. This shows that transparency can also have negative consequences. However, these turn out to be comparatively small. Transparent output again has strong positive effects. If users can understand why the recommended item matches their preferences, this increases trust in the competence of the system (0.194, p <.01). Indirectly, transparency also promotes overall satisfaction via this increase in trust (indirect: Transparency-output →Trust-competence →overall satisfaction: 0.055, p <.05). Furthermore, the increase in transparency indirectly (indirect: Transparency-output →Trust-competence →effectiveness: 0.056, p <.05), but also directly (0.127, p <.05), contributes to a higher rating of the system's effectiveness. Indirectly, this in turn increases overall satisfaction with the system (indirect: Transparency-output →Trust-competence →effectiveness →overall satisfaction: 0.019, p <.05). Additionally, it increases the willingness to follow the advice of the recommendation system when users better understand the output (direct: 0.268, p <.001; indirect: Transparency-output →Trust-competence →Trust-follow advice: 0.073, p <.05).

As shown, the transparency factors have clear effects on trust in the system, evaluation of effectiveness and on the overall satisfaction. Therefore hypotheses 2.1, 2.2 and 2.3 can be considered confirmed. Thus, perceived transparency can also be viewed as a mediator of perceived control over the system, user characteristics, and other qualities of the system. The importance of the different factors of perceived transparency can be shown by the differentiated assessment.

6. Discussion

We aimed at developing a measurement tool that is specifically focused on capturing the transparency of RS as perceived by users. In an initial interview study, concerns and uncertainties in relation to RS transparency were identified, which are well in line with the general AI-related questions compiled by [18]. This indicates that the scheme developed by these authors can be a useful

starting point for developing measures also for specific systems such as RS, which address a wider range of users beyond more expert users as in the original work by [18].

Our confirmatory analyses confirmed our hypothesis that subjective perceived transparency can be characterized by the factors: *input*, *output*, *functionality* and *interaction*. Adequate reliability as well as convergent and divergent validity was demonstrated, which indicates that identified transparency factors can clearly be considered as independent, and they can be distinguished from each other and also from other factors of the subjective evaluation of RS (trust, effectiveness, etc.).

The identified factors in our analysis reflect the basic components of RS as defined by [16], i.e., the *input* (what data does the system use?), the *functionality* (how and why is an item recommended?), and the *output* (why and how well does an item match one's own preferences?). Additionally, the factor interaction could be extracted. This factor is consistent with the category interaction (what if / how to be that, what has to be changed for a different prediction?) of the prototypical questions to AI, formulated by [18].

Furthermore, the final set of items can also be considered through the lens of the different interaction stages as defined by Norman [22]. In our examined context, for example, the stage *perception* relates to the presence of system functions that explicitly reveal information on how the recommendations were derived, e.g. through explanations. Items of the type "The system provided information about how…", grouped under the factor functionality, could be validated, indicating that making information about the recommendation process observable is a prerequisite for further cognitive processing. This indicates that the evaluation of perceived transparency should consider not only items related to users' interpretation (i.e. "user understands", as it has traditionally been evaluated in RS research), but also items related to the presence and perception of transparency-related system functions (e.g. "user notices that the system actually explains").

Once the user perceives a system output (e.g. the features of a recommendation or an explanation), the next stage is the *interpretation* of the system state, in which users use their knowledge to interpret the new system state [22]: in our context, to assess the recommendation inferred by the system. Our validated final set includes items which are related to the *interpretation* stage, and are of the type "I understood what data was used …", which can be grouped under the factor *input*), or "I understood how the system determined …", grouped under the factor *functionality* of our developed scale. This group of items is also consistent with the definition of perceived transparency by [5], which focuses on the perceived understanding of the inner processes of RS.

In a subsequent stage, users compare the interpreted system state to their own goal to decide about the next action, a stage defined by Norman [22] as *evaluation*. The item "I can tell how well the recommendations match my preferences" from our scale relates to this stage, by assessing explicitly the correspondence of the recommended items with one's own preferences. Items from the *interaction* group ("I know what needs to be changed in order to get better recommendations") can be associated with intent formation and the downstream path in the action cycle.

As discussed by [23], designers can contribute to close the gap between mental models (users' idea on how the system works [22]), and the actual system's functioning, by providing output and functionalities reflecting an adequate system's conceptual model, that can be "easily perceived, interpreted and evaluated" [23]. The above can in turn impact perceived transparency [19]. Consequently, our instrument can contribute to a more comprehensive assessment of subjective perceived transparency, by going beyond the one-dimensional construct addressing a general "why-recommended" understanding, and assessing instead, the extent to which output and functionalities reflecting the system's conceptual model are in fact perceived, interpreted and evaluated.

7. Conclusion and Outlook

The instrument developed can be seen as a first step towards assessing transparency in RS in a more comprehensive and cognitively meaningful manner. Overall, reliability and construct validity of the developed measurement instrument could be confirmed, identifying four transparency factors (input, output, functionality, interaction) and resulting in a 13 item questionnaire (see Table 1). The expected influence of system aspects and personal characteristics on the transparency factors could be demonstrated for the developed factors with the exception of transparency regarding interaction, which may be due to the limited interaction possibilities in the applications used by participants. Furthermore, we could show the impact of different transparency aspects on trust in the system and on the overall evaluation of the system.

The differentiated assessment of transparency makes it possible to elaborate the significance of individual aspects of transparency in more detail than it was possible with previous measurement instruments. Thus, it could be shown that transparency with respect to functioning and output is of greater importance for the dependent variables considered than transparency with respect to interaction and input.

The findings obtained here should be considered under the following limitations. Real systems were tested for this online study. On the one hand, this allowed us to obtain users' views with respect to applications they were

familiar with and that were fully functional. On the other hand, no controlled manipulation of influencing variables was possible. We also did not analyze the differences between the systems which would have required a larger sample, also addressing questions outside the scope of the present study. An effect of explanations could only be shown for the factors input and functionality, partly mediated by perceived control, which may also be due to the limited explanations provided by the systems used. In addition, only systems that were already known to the users were tested. Thus, a stronger expression of trust and overall more positive evaluation might be expected. In terms of social desirability or self-overestimation, perceived understanding might be valued higher than actual understanding would lead one to expect.

Follow-up research should be guided by the limitations mentioned here for further validation of the measurement instrument. The degree of perceived transparency should also be compared with actual, genuine understanding using parallel qualitative methods [6]. Furthermore, it is important to check to what extent the questionnaire is also able to evaluate systems that are unknown to the users. Assessing unfamiliar systems or specifically designed prototypes would provide the opportunity to systematically vary components of the recommender system (input, functionality, output), the quality of explanations, and/or the interaction possibilities [9]. Thus, the influence of these features on the transparency factors and likewise possible differences in their manifestation should be further explored.

Overall, a first validated version of a questionnaire to assess perceived transparency can be presented. The findings presented here also provide starting points for research into further elucidating the multi-faceted concept of transparency.

Acknowledgments

This work was funded by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

References

[1] N. Bostrom, E. Yudkowski, The Ethics of Artificial Intelligence, in: W. Ramsey, K. Frankish (Eds.), Cambridge Handbook of Artificial Intelligence, Cambridge University Press, 2014, pp. 316–334.

[2] U. S. a. H. S. C. (SHS), Recommendation on the Ethics of Artificial Intelligence, Technical Report, UNESCO, 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000379920.page=14.

[3] R. Sinha, K. Swearingen, The role of transparency in recommender systems, CHI EA '02 CHI '02 Extended Abstracts on Human Factors in Computing Systems (2002) 830–831.

[4] N. Tintarev, J. Masthoff, Explaining Recommendations: Design and Evaluation, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer, 2015, pp. 353–382. URL: https://doi.org/10.1007/978-1-4899-7637-6_10. doi:10.1007/978-1-4899-7637-6_10.

[5] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: Proceedings of the fifth ACM conference on Recommender systems - RecSys 11, 2011, pp. 157–164.

[6] H. Cramer, V. Evers, S. Ramlal, M. van Someren, L. Rutledge, N. Stash, L. Aroyo, B. Wielinga, The effects of transparency on trust in and acceptance of a content-based art recommender, User Model. User-Adap. Inter. 18 (2008) 455–496.

[7] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, in: User Modeling and User-Adapted Interaction, 2012, p. 441–504.

[8] F. Gedikli, D. Jannach, M. Ge, How should i explain? a comparison of different explanation types for recommender systems, International Journal of Human-Computer Studies 72 (2014) 367–382.

[9] D. C. Hernandez-Bocanegra, J. Ziegler, Explaining review-based recommendations: Effects of profile transparency, presentation style and user characteristics, Journal of Interactive Media 19 (2020) 181–200. doi:https://doi.org/10.1515/icom-2020-0021.

[10] D. C. Hernandez-Bocanegra, J. Ziegler, Effects of interactivity and presentation on review-based explanations for recommendations, in: Human-Computer Interaction – INTERACT 2021, Springer International Publishing, 2021, pp. 597–618.

[11] N. Tintarev, J. Masthoff, Evaluating the effectiveness of explanations for recommender systems, User Modeling and User-Adapted Interaction 22 (2012) 399–439.

[12] C.-H. Tsai, P. Brusilovsky, Explaining recommendations in an interactive hybrid social recommender, in: 24th International Conference on Intelligent User Interfaces (IUI 19), 2019, pp. 391–396.

[13] A. Jameson, M. C. Willemsen, A. Felfernig, M. de Gemmis, P. Lops, G. Semeraro, L. Chen, Human decision making and recommender systems, Recommender Systems Handbook (2015) 611–648.

[14] S. Dooms, T. D. Pessemier, L. Martens, A user-centric evaluation of recommender algorithms for an event recommendation system, in: Proceedings of the RecSys 2011: Workshop on Human Decision Making in Recommender Systems (Decisions RecSys 11) and User-Centric Evaluation of Recommender Systems and Their Interfaces - 2 (UCERSTI

2) affiliated with the 5th ACM Conference on Recommender Systems (RecSys 2011), 2011, p. 67–73.
[15] M. Bühner, Einführung in die Test- und Fragebogenkonstruktion, Pearson Studium, Aufl. München, 2011.
[16] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, Recommender Systems. An introduction, Cambridge University Press, 2011.
[17] J. Lu, Q. Zhang, G. quan Zhang, Recommender Systems. Advanced Developments, World Scientific Publishing, 2021.
[18] Q. V. Liao, D. Gruen, S. Miller, Questioning the ai: Informing design practices for explainable ai user experiences, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems 9042 (2020) 1–15. doi:https://doi.org/10.1145/3313831.3376590.
[19] T. Ngo, J. Kunkel, J. Ziegler, Exploring mental models for transparent and controllable recommender systems: A qualitative study, in: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization UMAP 20, 2020, pp. 183–191.
[20] D. Borsboom, G. J. Mellenbergh, J. van Heerden, The theoretical status of latent variables, Psychological Review 110 (2003) 203–219.
[21] J. Kunkel, T. Ngo, J. Ziegler, N. Krämer, Identifying Group-Specific Mental Models of Recommender Systems: A Novel Quantitative Approach, in: Human-Computer Interaction – INTERACT 2021, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2021, pp. 383–404. doi:10.1007/978-3-030-85610-6_23.
[22] D. A. Norman, Some Observations on Mental Models, In Mental Models, Dedre Gentner and Albert L. Stevens (Eds.). Psychology Press, New York, NY, USA, 1983.
[23] E. Hutchins, J. D. Hollan, D. A. Norman, Direct manipulation interfaces, Human-Computer Interaction 1 (1985) 311–338.
[24] Y. Zhang, X. Chen, Explainable recommendation: A survey and new perspectives, Foundations and Trends in Information Retrieval 14 (2020) 1–101.
[25] K. Backhaus, B. Erichson, R. Weiber, Multivariate Analysemethoden. Eine anwendungsorientierte Einführung, Berlin 13. Aufl, 2011.
[26] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, Multivariate data analysis. A global perspective, Boston 7. Aufl., 2010.
[27] D. H. McKnight, V. Choudhury, C. Kacmar, Developing and validating trust measures for e-commerce: An integrative typology, in: Information Systems Research, volume 13, 2002.
[28] P. Kouki, J. Schaffer, J. Pujara, J. O'Donovan, L. Getoor, Personalized explanations for hybrid recommender systems, in: Proceedings of 24th International Conference on Intelligent User Interfaces (IUI 19), ACM, 2019, p. 379–390.
[29] T. Donkers, T. Kleemann, J. Ziegler, Explaining recommendations by means of aspect-based transparent memories, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, p. 166–176.
[30] K. Hamilton, S.-I. Shih, S. Mohammed, The development and validation of the rational and intuitive decision styles scale, Journal of Personality Assessment 98 (2016) 523–535.
[31] A. G. Yong, S. Pearce, A beginner's guide to factor analysis: Focusing on exploratory factor analysis, Tutorials in Quantitative Methods for Psychology 9 (2013) 79–94.
[32] R. A. Peterson, A meta-analysis of cronbach's coefficient alpha, Journal of Consumer Research 21 (1994) 381–391.