Assessing the Value of Transparency in Recommender Systems: An End-User Perspective

Eric S. Vorm*
Andrew D. Miller†
Indiana University Purdue University Indianapolis
Indianapolis, IN

ABSTRACT

Recommender systems, especially those built on machine learning, are increasing in popularity, as well as complexity and scope. Systems that cannot explain their reasoning to end-users risk losing trust with users and failing to achieve acceptance. Users demand interfaces that afford them insights into internal workings, allowing them to build appropriate mental models and calibrated trust. Building interfaces that provide this level of transparency, however, is a significant design challenge, with many design features that compete, and little empirical research to guide implementation. We investigated how end-users of recommender systems value different categories of information to help in determining what to do with computer-generated recommendations in contexts involving high risk to themselves or others. Findings will inform future design of decision support in high-criticality contexts.

1 INTRODUCTION

New machines are embodied with increasing levels of authority and unprecedented scope. Decisions previously made by humans are increasingly being made by computers, often with little or no explanation, raising concerns over a plethora of social, legal, and ethical issues such as privacy, bias, and safety.

Transparency is often discussed in terms of back-end programming or troubleshooting. End-users, especially in the context of novice users interacting with recommender systems, are seldom studied. Yet recent developments in AI suggest that automated recommendations will be an increasingly common component in user's daily lives as technologies such as self-driving cars and IoT-enabled smart homes become commonplace. Developing methods to increase the transparency of computer-generated recommendations, as well as understanding user information needs as a means to increase trust and engagement with recommendations, is therefore crucial. Accomplishing transparent interface design is often complicated by a series of trade-offs that seek to balance and prioritize several competing design principals. Striking the appropriate balance between too much and not enough information is often more art than science, and is becoming more difficult with the cascading prevalence of data-driven paradigms such as machine learning [1].

Efforts towards improving the transparency of recommender systems commonly involve programming system-generated explanations that seek to justify the recommendation to users, often through the use of system logic [2]. Providing explanations and justifications of system behavior to users has proven to be a highly effective means to increase user acceptance and enhancing user attitudes towards recommender systems [3]. Studies have shown that providing explanations to users tends to increase trust [4], improves user comprehension [5], calibrates appropriate reliance on decision aids [6], and enables better detection and correction of system errors [7]. Generating explanations that users find both useful and satisfactory, however, can be a complicated task, and much research has been conducted to try to answer the question of "what makes a good explanation" [8].

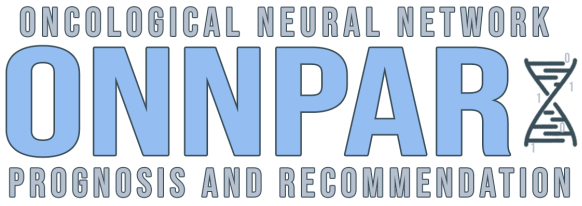

Figure 1: ONNPAR is a simulated clinical decision support system built on machine learning. It was used as the testbed for this study, serving the role of a highly-critical decision context.

While system-generated explanations represent the most *common* approach to transparency in recommender systems, in many cases simply providing users access to certain types of information can also improve transparency, and can dramatically improve user experience and the likelihood of further interaction [5]. In some contexts, affording users the opportunity to see into the system's dependencies, policies, limitations, or information about how the user is modeled and considered by the system can facilitate the same level of user understanding (and subsequent trust) as an explicit explanation [9].

Providing targeted information as a means of improving a user's mental model and trust (i.e., transparency) has two potential benefits over the building of explanation interfaces. First, it affords users an opportunity to use deductive reasoning to determine the merit and validity of system recommendations, which has been demonstrated to improve usability and user trust in many contexts. For instance, Swearingen and Sinha reported that recommender systems whose interfaces provided information that could help users understand the system were preferred over those that did not [10]. Research in cognitive agents has also demonstrated that providing users access to underlying system information, such as system dependencies or provenance of data, can greatly improve human-machine performance and reduce the likelihood of users acting on recommendations that are erroneous, known as "errors of commission" [11]. A second benefit of affording users the opportunity to see into the system in order to understand its processes is that it takes little to no additional programming. This is often because much of the information that

*esvorm@iu.edu
†andrewm@iupui.edu

could enhance user understanding of system functions and behaviors is already present in the system, but is often hidden from front-end interfaces in order to reduce clutter and streamline layouts.

This trade off between providing adequate information to communicate a system's intent and achieving a user-friendly interface design is a common challenge, often resolved through iterative design evaluations involving user testing. While research involving transparency in system design frequently focuses on behavioral outcomes, such as modeling the appropriateness of a user's interaction with a recommender system, little is known about what information is most efficacious to users in terms of improving mental models, resolving conflicts caused by unexpected or unanticipated system behaviors, or improving user trust and technology acceptance. Answering these questions requires an investigation into how user's subjectively value and prioritize different categories of information in an effort to resolve conflicts between expected and observed system behaviors, or in order to evaluate the validity or accuracy of a recommendation in order to determine whether to accept or reject it.

To accomplish this, we used an approach known as Q-Methodology, commonly referred to as the systematic study of subjectivity [13]. To constrain our work and prevent over generalization of findings, we chose to investigate what information users value most when engaged with recommender systems in a highly critical decision scenario. We hypothesize that users involved in tasks that involve a high degree of personal risk or risk to others are more likely to critically interrogate computer-generated recommendations before accepting and acting upon them. This suggests that systems providing recommendations in highly critical decision contexts, such as medical, legal, financial, or automotive domains, amongst others, would benefit most from efforts to develop interfaces that enable users to quickly and accurately discern whether or not to trust those recommendations. Using the decision criticality framework as a guide, we developed a hypothetical recommender system named the Oncological Neural Network Prognosis and Recommendation (ONNPAR) System. ONNPAR was modeled after modern clinical decision support systems offering recommendations, and was designed to serve as the highly-critical decision scenario for our research.

2 METHODS

2.1 A brief introduction to Q-Methodology

Q-methodology is distinctly different from "R" methodology and has several distinctions that should be addressed. R-methodology samples respondents who are representative of a larger population, and measures them for the presence or expression of certain characteristics or traits. These measurements are made objectively, as the opinions of respondents is seen as potentially confounding and are therefore controlled. Using inferential statistics, findings are then abstracted to predict prevalence and generalize findings to a larger target population [50].

Q-methodology, on the other hand, invites participants to directly express their subjective opinions on a given topic by sorting statements (or questions) into a hierarchy that represents what is most or least important to them. Each participant's arrangement of statements or questions represents an individual person's point of view about a given topic, which ordinarily would not be of much value beyond understanding the points of view present in that particular group of individuals. Through the use of factor analysis, however, patterns of subjective opinion are uncovered, which reveal a structure of thoughts and beliefs surrounding a given topic and context. We can use these findings to understand or model a phenomenon, or in our example, infer the potential value of different design features through user input that is both qualitatively rich, yet statistically sound.

In Q-methodology, participants are given a bank of statements, each one on a separate card (or electronically using specialized software), and asked to rank order them in a forced distribution grid according to some measure of affinity or agreement, depending on the context of the study [13]. For our study, we employed q-methodology as a design-elicitation tool, similar to traditional iterative design strategies involving user evaluation of prototype designs. In this way, we provided participants with questions, each representing a design feature or suite of features that could be provided through a user interface (UI). We asked participants to sort these statements in a forced distribution, such as the one shown in figure 2, ranking them from most important to least important to them. Then, through the use of factor analysis, we analyzed the

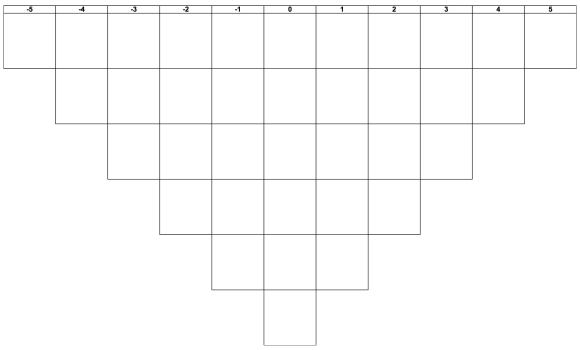

Figure 2: Example forced-sort matrix used for our study. Participants sorted all 36 questions into the array, ranking them according to personal value and significance in the context of information that could help them understand how the ONNPAR system works, and determine whether or not to accept or reject the computer-generated recommendation.

different ways that users value and prioritize these questions, thus inferring what design elements may add to or detract from an optimal user experience [15] and quantifying the potential value of different categories of information to users in the context of improving the transparency of recommender system interfaces.

2.2 Model Development

The first step for our study was to ensure that our approach was representative of the technical and theoretical issues related to transparency in recommender systems (i.e., ontological homogeneity). To accomplish this we used a combination of analytic and inductive techniques, combining findings from a systematic literature review with user input from a user-centered design workshop conducted for a previous project [16].

We also sought out the advice and guidance of subject matter experts (SMEs) to ensure that all technical and theoretical aspects of the concept of transparency in recommender systems had been addressed. We conducted informal interviews with a combination of academics who regularly conduct research in the fields of machine learning and intelligent systems, as well as applied researchers currently engaged in the development and design of recommender systems for industry. In total, nine SMEs were consulted and asked to review our preliminary categorization structure, and to offer suggestions for other technical or theoretical issues not already captured by our approach.

The result was a five-factor model of transparency in recommender systems. These categories consist of Data, Personal, System, Options, and Social. We briefly describe and discuss the relevance of these categories below.

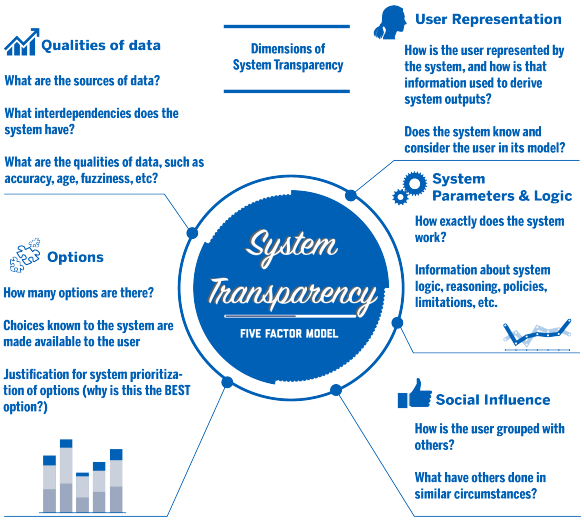

Figure 3: A five-factor model of system transparency. Each factor represents categories of information which can assist users in understanding and trust computer-generated recommendations.

System Parameters and Logic: Understanding the perspective of another in order to anticipate their actions or understand their intentions is the process known as building a mental model [17]. Information related to how a system works, including its policies, logic, and limitations, can help users build an appropriate mental model of the system. This is often critical, as many accidents, particularly in high-risk domains such as aviation, have resulted from users having an inappropriate or inaccurate mental model of system functionality [18]-[20].

Having knowledge of how a system functions can also help in determining when the system may be in error. Numerous studies have demonstrated that providing information about how the system processes information can improve the detection of system errors and faults[21]-[23], and can thereby lower so-called 'errors of commission' [24]. These studies indicate that providing users with information that assists their understanding of system functionality may be a viable way to improve the transparency of recommender systems.

Qualities of Data: In many instances, understanding the relationship of dependencies present in a system can provide meaningful insights into that system's functionality. A computer program may be functioning perfectly, but if the data on which it is operating is exceedingly noisy or corrupt, its outputs may still be incorrect or inappropriate. Numerous real-world examples from accidents such as the Space Shuttle Challenger and the Navy warship USS Vincennes serve as a testament to the importance of providing information on the quality and provenance of the underlying data to decision makers [25].

Efforts to make data-related information available to users of machine learning applications have been shown to result in higher user ratings of ease of understanding, meaningfulness, and convincingness [26]. Advances in visual analytic approaches have also improved the comprehensibility and intelligibility of data to users by presenting it in a manner that is more readily understood [27]. Different visualization techniques have also been demonstrated to improve user's understanding of cause and effect relationships between variables, even among users with little to no data analytical background (i.e., data novices, [28]).

Just as it is important to consider the source as well as the quality of information, so too must users be able to see into the system and understand the data on which it is operating. The current data-driven paradigm of machine learning, therefore, necessitates information that can help users answer questions about the qualities of the system's data. Affording users the ability to see this data may well improve the transparency of a system's interface from the user's perspective.

User Representation: The concept of personalization is central to the discussion of transparency in a variety of intelligent system domains such as context-aware and automated-collaborative filtering applications [4], [29]-[31]. Users often want to understand how they are modeled by a system, if at all, and to what extent system outputs are personalized for them. While commercial applications such as personalized targeted advertisement algorithms are an important component of this category, the importance of user representation extends well beyond the suitability of computer-generated recommendations like movies or music titles.

Future machine learning applications are expected to encompass a variety of domains that may very well necessitate extensive explanation of how users are represented by computer systems in order to achieve user buy-in and acceptance. For example, in the domain of personal financial trading, a machine learning algorithm may possess a model of risk that is very different from its user, and may perhaps prioritize one aspect of financial growth, such as diversification, over other aspects that the user may prioritize more, such as long-term stability. Understanding what a system knows about its user, and how that information is subsequently used to derive recommendations, is therefore of potential critical importance for applications to achieve acceptable levels of user trust, engagement, and technology acceptance.

Social Influence: The power of social media has been displayed in a variety of contexts over the past decade of its modern existence, and has become a powerful tool for marketers and influencers. As of August 2017, two thirds of Americans (67%) reported that they

received at least some of their news from social media [32]. Systems that group users according to online behavior in order to predict future interests and purchases, such as automated collaborative filtering algorithms, are abundant, and represent a foundational approach to modern marketing and sales [33]. In many cases, a user's understanding of how they are grouped by a system using social media information can provide meaningful insights into why a system output, such as a targeted advertisement, was generated. This is most important when conflicts arise between a user and an inappropriate system output. These are often the result of loose affiliations on social media with others who may hold radically opposing philosophical or political viewpoints, which some recommender systems incorrectly associate into their models. Providing users opportunities to see into a system and understand how they, the user, are categorized and represented in a social group, may improve user experience and trust, leading some users to remain more willing to interact with a system after such a conflict arises. There is also some evidence that some decision making may be socially-mediated as well.

Scientists have long studied the broad range that social influences can have on decision making and behavior. These can include various social biases [34] which can explain in limited cases how some people sometimes defer their decision making to a group or other individual, even when it would seem prudent not to do so [35]. Additionally, many people express the importance of social relationships in guiding and assisting in decision-making. In a 2017 Pew Research Poll, 74% of American respondents reported that their social circles played at least a small role in their decision making; 37% reported it played a significant role [36]. Systems that afford information that connects a user's system interaction with their social circles, may well improve user satisfaction and usability. For example, if we imagine a user attempting to determine whether or not to accept or reject a recommendation, in some contexts, social information, such as the prevalence of that recommendation to others in their social circle, or a ratio of accept/reject decisions from their friends or family, could prove to be valuable to some people, and could be used as a decision heuristic.

Options: People often express a preference of choice over no choice in most decision-making contexts [37]. Accordingly, many systems strive to offer choices to users as a means of increasing engagement and satisfaction [38]. There are times, however, when providing multiple choices to a user may be undesirable.

For example, most navigation systems output at most three route choices to the user, and typically highlight the one recommended by the system. There may be, of course, several hundreds or even thousands more options available to the user, but displaying them all would unlikely benefit the user, and may in fact lead them to discard the technology due to its confusing and cluttered interface.

This "tyranny of choices" [39] is even more evident in light of the size and scope of many machine learning models, especially those involving deep learning. In these circumstances, it is practically infeasible to display every possible optional output to the user.

Common interface design strategies involve efforts that reduce choices in order to lessen cognitive load and improve the speed and efficiency of decision making [40]. Determining the trade-offs between interface aesthetics (i.e., clutter) and user preference for options is often a challenge for engineers and designers alike. Sometimes, these decisions are determined by external factors, such as corporate policy, or mandated safety requirements [41]. But in some contexts, users may want more options than they are often provided, or, at the very least, users may want to know whether or not other options exist before engaging in a decision. Closely related to this is the importance of providing some justification of why one option is deemed better than another.

Much has been written about the role that system explanations or justifications can have on a person's interaction with or sentiment towards intelligent systems [42], [43]. Users often demand some form of justification from a system to help them determine the merit of an output such as a recommendation [10]. There are a variety of sub-categories of this concept too, such as why one option is NOT the best, (known as counter-factual explanation).

The range of discussions over how precisely to engineer explanation systems in a format that is meaningful and understood by the user under different circumstances is the subject of much current discussion in the intelligent systems communities of practice, especially related to machine learning (for an exhaustive review, see [8]). Much of these are beyond the scope of this current paper, but for the purposes of this discussion, suffice it to say that the ability for systems to offer explanations of their outputs is central to the concept of transparency in recommender systems.

2.3 Concourse and Q-sort development

Having identified these five factors, we then created a bank of questions for our participants to sort. This bank is known in Q-methodology parlance as a 'concourse.' A goal of developing a concourse is to create as many statements as possible to ensure a comprehensive and saturated pool of opinions or sentiment from which to sample. We used Ram's taxonomy of question types as an initial starting point to ensure that we used a variety of question types [44]. This was then refined using Silveira et al's taxonomy of user's frequent doubts [45]. The initial concourse consisted of 71 questions. We then refined this concourse down to a reasonable bank of 36 questions through the use of five individuals who are subject-matter experts in recommender systems (either professors in Cognitive Psychology with experience with recommender systems, or programmers of recommender systems). Questions that appeared redundant were combined, and those that were deemed irrelevant or unrelated were discarded. Each of the five factors had a roughly equivalent number of representative questions.

This final bank of 36 questions was randomized and assigned numbers, then printed on 3x5 index cards. Each participant received their own deck consisting of 36 individual questions. Participants were given instructions for how to sort cards from most-to-least valuable or important to them. Participants were then shown a vignette on a computer screen or projector. The vignette described an interaction with ONNPAR, and ends with the user being given a recommendation which the user must determine whether or not to act on, or reject. Participants then sorted their cards, and recorded their arrangement on a form, along with two additional questions on a questionnaire: *In a few words, please explain WHY you chose your MOST/LEAST important question to ask."*

3 RESULTS

Our participant sample was comprised of *n*=22, 16 males, 6 females, aged 22-59, average age 33 years old. Expertise was evaluated by self-report. Participants were classified as novices if they had no knowledge of or personal use experience with recommender systems, and experts if they had participated in either the design or programming of recommender systems.

In the following sections we briefly describe the methodological analysis of q-methodology, and then present the findings from our ONNPAR study. We will describe interpretations and insights from each of the factor groups of our factor analysis in the discussion section.

3.1 Q-method Analysis Overview

The analysis of q-methodology is quite straightforward. Each question from the set is assigned a numerical value according to which column it was placed (-5 to +5 for our study). Each participant's arrangement of cards is then combined to create a by-person correlation matrix. This matrix describes the relationship of each participant's arrangement of questions with every other participant's arrangement (NOT the relationship between items within each participant). This matrix is then submitted for factor analysis, which produces factors onto which participants load based on their arrangements of questions. Two or more participants who load on the same factor, therefore, will have arranged their questions in a very similar manner, which represents similar reasoning styles or prioritization. These factors, or clusters of participants, are then analyzed by examining what questions were ranked highest and lowest by each group, as well as examining the similarities and differences between each factor group.

For simplicity's sake, we will henceforth refer to factors as factor groups, since in the context of q-methodology, factor analysis identifies groups of individuals. The term factor group is not to be confused with the five-factor model of transparency, used to guide our investigation.

Several statistical packages are freely available to aid in the analysis of q-methodology studies. We used a version known as Ken-Q Analysis [46].

3.2 Factor Analysis

Once all sorts had been entered into our database, they were factor analyzed using the Ken-Q software. We used principal components analysis (PCA) because it has been shown to better account for random, specific, and common error variances [47]. The unrotated factor matrix was then analyzed to determine how many factors to retain for rotation. A significant factor loading at (P<0.01) is calculated using the equation $2.581\sqrt{n}$ where n = the number of questions in our set (36). Individuals with factor loadings of $\pm$ .48 were considered to have loaded on a factor and were arranged into a factor group.

For factor extraction, we used the common practice of evaluating only factors with an eigenvalue greater than one [13]. We also determined that only factors with three or more participants loading on them would be retained. These steps resulted in four factors, which were then submitted to rotation according to mathematical criteria (e.g., *varimax*). With this four-factor solution, all but one participant loaded clearly on at least one factor, resulting in four distinct viewpoints of information priorities and preferences of 21 individuals.

Factor Characteristics				
	Factor 1	**Factor 2**	**Factor 3**	**Factor 4**
No. of Defining Variables	8	5	5	3
Avg. Rel. Coef.	0.8	0.8	0.8	0.8
Composite Reliability	0.966	0.96	0.952	0.941
S.E. of Factor Z-scores	0.184	0.2	0.219	0.243

Table 1: Characteristics of factors after rotation.

3.3 Factor Interpretation

Once factor extraction and rotation was complete, we analyzed each factor group to interpret its meaning. This was first accomplished by producing a weighted average of each participant's arrangement of cards from within their factor group, and combining those arrangements into one exemplar composite arrangement, which serves as the model arrangement of questions for that factor group. Once these composite arrangements, or "factor arrays," have been developed for each factor group, they can then be analyzed for deeper interpretation. We next evaluated the questions that were ranked highest and lowest for each factor array. This provides an early indication of information priorities, and allows us to begin crafting a picture of how participants in each factor group tend to think about the value of each category of information.

3.4 Factor Groups

Here we will report the findings from the factor analysis. To do this we will describe each factor group's arrangements of the questions in terms of their highest- and lowest-ranked questions, as well as positive and negative distinguishing questions. Distinguishing questions are those where the absolute differences between factor z-scores are larger than the standard error of differences for a given pair of factors. All distinguishing questions are significant at ($p < .01$).

Factor Group One was defined by eight participants and explained 22% of the study variance with an eigenvalue of 6.7. Three of the factor loading participants were females, five were males, with an average age of 37.5 years old. Knowledge of recommender systems was split between five novices and three experts.

The highest ranked question of this factor group was "Why is this recommendation the best option?" (+5) The lowest ranked question of this factor group was "Is there anyone in my social network that has received a similar recommendation?" (-5) Other positive distinguishing questions for the factor one group were (in descending order): "What are all of the factors (or indicators) that were considered in this recommendation, and how are they weighted?" (4) "Precisely what information about *me* does the system know?" "What does the system think is *me* level of "acceptable risk?" (1) Negative distinguishing questions for Factor Group One were (in ascending order): "How much data was used to train this system?" (-4) "How many other people have received this recommendation from this system?" (-2) and "What does the system *think* I want to achieve?" (-1)

Factor Group Two was defined by five participants and explained 13% of the study variance with an eigenvalue of 2.8. All

of the factor loading participants were males, average age of 42 years old. All but one of this factor group were considered experts in recommender systems. The highest ranked question of this factor group was "What are all of the factors (or indicators) that were considered in this recommendation, and how are they weighted?" (+5) The lowest ranked question of this factor group was "Was this recommendation made specifically for ME (based on my profile/interests), or was it made based on something else (based on some other model, such as corporate profit, or my friend's interests, etc.)?" (-5) Positive distinguishing questions for the factor two group were (in descending order): "How is this data weighted or what data does the system prioritize?" (+4) "How much data was used to train this system?" (+2) "Is my data uniquely different from the data on which the system has been trained?" (1) Negative distinguishing questions for the factor two group were (in ascending order): "Is there anyone in my social network that has received a similar recommendation?" (-4) "What does the system think is MY level of "acceptable risk?" (-2) "What if I decline? How will that decision be used in future recommendations by this system?" (-1) "How is my information measured and weighted in this recommendation?" (-1)

Relative Rankings of Questions by Factor Group	
	Factor Group 1
Highest	Why is this recommendation the best option?
Lowest	Is there anyone in my social network that has received a similar recommendation?
	Factor Group 2
Highest	What are all of the factors (or indicators) that were considered in this recommendation, and how are they weighted?
Lowest	Was this recommendation made specifically for *me* (based on my profile/interests), or something else?
	Factor Group 3
Highest	Under what circumstances has this system been wrong in the past?
Lowest	What if I decline? How will that decision be used in future recommendations by this system?
	Factor Group 4
Highest	What is the history of the reliability of this system?
Lowest	What does the system *think* I want to achieve? (How does the system represent my priorities and goals?)

Table 2: Highest and lowest ranking questions of each factor group. Although this is only the most superficial analysis, distinguishing differences amongst groups begin to emerge as we analyze each group's prioritization and valuation of transparency information.

Factor Group Three was defined by five participants and explained 9% of the study variance with an eigenvalue of 1.9. Two of the factor loading participants were females, three were males, and an average age of 34 years old. All but one of the participants for this group were considered experts in recommender systems.

The highest ranked question of this factor group was "Under what circumstances has this system been wrong in the past?" (+5) The lowest ranked question of this factor group was "What if I decline? How will that decision be used in future recommendations by this system?" (-5) Other positive distinguishing questions for the factor three group were (in descending order): "What data does the system depend on in order to work properly, and do we know if those dependencies are functioning properly?" (+4) "Is my data uniquely different from the data on which the system has been trained?" (+3) "What have other people like me done in response to this recommendation?" (+2) Negative distinguishing questions for the factor three group were (in ascending order): "What is the system's level of confidence in this recommendation?" (-2) "Are there any other options not presented here?" (-2) "How much data was used to train this system?" (-1) "How does the system consider risk, and what is its level of "acceptable risk?" (-1)

Factor Group Four was defined by three participants and explained 8% of the study variance with an eigenvalue of 1.7. There were two males and one female, and an average age of 20 years old. Knowledge of recommender systems was split between two novices and one expert.

The highest ranked question of this factor group was "What is the history of the reliability of this system?" (+5) The lowest ranked question of this factor group was "What does the system THINK I want to achieve? (How does the system represent my priorities and goals?)" (-5) Positive distinguishing questions for the factor four group were (in descending order): "How many other people have accepted or rejected this recommendation from this system? (What is the ratio of approve to disapprove?)" (+4) "Is the system working with solid data, or is the system inferring or making assumptions on 'fuzzy' information?" (+3) "How many other people have received this recommendation from this system?" (+1) Negative distinguishing questions for the factor four group were (in ascending order): "Is my data uniquely different from the data on which the system has been trained?" (-3) "What are all of the factors (or indicators) that were considered in this recommendation, and how are they weighted?" (-2) "What have other people like me done in response to this recommendation?" (-1)

4 DISCUSSION

Findings from our factor analysis yielded several surprising insights. We begin with a discussion of questions that produced a high degree of either consensus or disagreement amongst factor groups, and then conclude with a discussion of each factor group.

4.1 Analysis of Consensus vs. Disagreement Findings

A common technique to examine these data is to explore questions that created either consensus or a large amount of disagreement in our sample. By examining the variance between all item rankings, we can explore what questions were generally agreed on (i.e., consensus), and what items produced large disagreement. For instance, all participants ranked "How clean or accurate is the data used in making this recommendation?" as either 0 or -1, indicating that this question was only moderately valuable to them in the context of a clinical decision support system. This is potentially valuable information for designers to consider, given that the fuzziness of data is sometimes displayed to users as a method of enhancing system transparency [48]. Given these findings, it may be useful to reconsider displaying information about the qualities of data to users in favor of other types of information deemed more useful or valuable.

Similarly, we can learn much from these data by evaluating questions that produced a great deal of disagreement between factor groups. For instance, the question "Was this recommendation made specifically for *ME* (based on my profile/interests), or was it made based on something else (based on some other model, such as corporate profit, or my friend's interests, etc.)?" had the largest variance, with factor groups one and three assigning it a positive value (4 and 3), and factor groups two and four assigning it a negative value (-5 and -4). Interestingly, factor group two assigned this question as the *least* valuable or important question of their q-set, while factor group one assigned this question as their *second most* valuable or important question.

Interpreting these findings can, at first glance, appear confounding to a designer looking for clear guidance. Clearly, some individuals would prefer to have information that could indicate how they, as a user, are modeled and considered (if at all) in system-generated recommendations as a means of improving their trust, while others clearly discount the value of this kind of information. These findings suggest that social influence information, such as what other users are doing in response to recommendations, may at times be valuable to some users in helping determine whether or not to accept or reject a recommendation.

Two other questions also produced wide disagreement across factor groups. "How many other people have accepted or rejected this recommendation from this system? (What is the ratio of approve to disapprove?)" and "Is there anyone in my social network that has received a similar recommendation?" were ranked near the poles by different factor groups. This indicates that the value of social media-related information in highly-critical contexts, while not important to some, is still considered valuable information by some users who may find it a valuable and important component to enhance their understanding and trust in system-generated recommendations.

Consensus Versus Disagreement	
Consensus	**Z-Score Variance**
Can I influence the system by providing feedback? Will it listen and consider my input?	0.024
How clean or accurate is the data used in making this recommendation?	0.029
How often is the system checked to make sure it is functioning as it was designed (i.e., for model accuracy)?	0.046
Disagreement	**Z-Score Variance**
How many other people have accepted or rejected this recommendation from this system? (What is the ratio of approve to disapprove?)	1.179
Is there anyone in my social network that has received a similar recommendation?	1.246
Was this recommendation made specifically for ME, or was it based on something else?	2.261

Table 3: Consensus questions are those which all participants agreed were of relevant importance, as indicated by a low Z-score variance in their arrangements. Disagreement questions are those which polarized opinion, as indicated by high Z-score variance in their arrangments.

5 CONCLUSION

We have illustrated our five-factor model of information categories that can be used to increase the transparency of recommender systems to end users. We developed a bank of 36 questions representing information gathering strategies that users could use to interrogate system-generated recommendations in an effort to understand its reasoning, and decide whether to accept or reject the recommendation. Using this bank of questions, participants sorted them according to those they found most valuable or useful in helping them determine whether to accept or reject a computer-generated recommendation. We analyzed how participants arranged these questions using a factor analytic technique. Our findings support other studies that find that transparency is a multi-dimensional construct, and achieving it is dependent on multiple variables, including to some extent the user's preferences for and valuation of certain categories of information. Our findings are intended to inform future interface design of recommender systems, as well as to broaden the discussion of the importance of building systems whose outputs and recommendations are easily understood by their users.

REFERENCES

[1] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," AirXiv, 2017.
[2] B. Buchannan and E. Shortliffe, Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, MA: Addison Wesley, 1984.
[3] L. R. Ye and P. E. Johnson, "The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice," MIS Quarterly, vol. 19, no. 2, p. 157, Jun. 1995.

[4] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," presented at the 2000 ACM conference, New York, New York, USA, 2000, pp. 241-250.

[5] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of Explanatory Debugging to Personalize Interactive Machine Learning," presented at the the 20th International Conference, New York, New York, USA, 2015, pp. 126-137.

[6] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," International Journal of Human-Computer Studies, vol. 58, no. 6, pp. 697-718, Jun. 2003.

[7] B. Lorenz, F. Di Nocera, and R. Parasuraman, "Display Integration Enhances Information Sampling and Decision Making in Automated Fault Management in a Simulated Spaceflight Micro-World," Proceedings of the Human Factors and Ergonomics Society 58th Annual Meeting, pp. 31-35, 2002.

[8] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," AirXiv, pp. 1-57, Jun. 2017.

[9] G. B. Duggan, S. Banbury, A. Howes, J. Patrick, and S. M. Waldron, "Too Much, Too Little, or Just Right: Designing Data Fusion for Situation Awareness," Proceedings of the Human Factors and Ergonomics Society 58th Annual Meeting, pp. 528-532, 2004.

[10] K. Swearingen and R. Sinha, "Beyond algorithms: An HCI perspective on recommender systems," ACM SIGIR 2001 Workshop on Recommender Systems, 2001.

[11] H. F. Neyedli, J. G. Hollands, and G. A. Jamieson, "Beyond Identity: Incorporating System Reliability Information Into an Automated Combat Identification System," Human Factors, vol. 53, no. 4, pp. 338-355, Jul. 2011.

[12] W. Stephenson, The study of behavior; Q-technique and its methodology. Chicago, IL: University of Chicago Press, 1953.

[13] S. R. Brown, "A primer on Q methodology," Operant Subjectivity. 16(3/4), 91-138. 1993

[14] S. Watts and P. Stenner, "Doing Q Methodology: theory, method and interpretation," Qualitative Research in Psychology, vol. 2, no. 1, pp. 67-91, Jan. 2005.

[15] K. O'Leary, J. O. Wobbrock, and E. A. Riskin, "Q-methodology as a research and design tool for HCI," presented at the CHI 2013, Paris, France, 2013, pp. 1941-1950.

[16] E. S. Vorm, "Assessing Demand for Transparency in Intelligent Systems Using Machine Learning," presented at the IEEE Innovations in Intelligent Systems and Applications INISTA, Thessonaliki, 2018, pp. 41-48.

[17] W. B. Rouse and N. M. Morris, "On looking into the black box: Prospects and limits in the search for mental models.," Psychological Bulletin, vol. 100, no. 3, pp. 349-363, 1986.

[18] N. B. Sarter and D. D. Woods, "How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control," Human Factors, vol. 37, no. 1, pp. 5-19, 1995.

[19] A. F. Zeller, "Accidents and Safety," in Systems Psychology, K. B. DeGreene, Ed. New York, NY, 1970, pp. 131-150.

[20] National Transportation Safety Board, "Loss of Control on Approach Colgan Air, Inc. Operating as Continental Connection Flight 3407 Bombardier DHC-8-400, N200WQ Clarence Center, New York February 12, 2009," National Transportation Safety Board, NTSB/AAR-10/01 PB2010-910401, Feb. 2010.

[21] G. G. Sadler, H. Battiste, N. Ho, L. C. Hoffmann, W. Johnson, R. Shively, J. B. Lyons, and D. Smith, "Effects of transparency on pilot trust and agreement in the autonomous constrained flight planner," presented at the 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), 2016, pp. 1-9.

[22] A. Sebok and C. D. Wickens, "Implementing Lumberjacks and Black Swans Into Model-Based Tools to Support Human-Automation Interaction," Human Factors, vol. 59, no. 2, pp. 189-203, Mar. 2017.

[23] J. Y. C. Chen, K. Procci, M. Boyce, J. L. Wright, A. Garcia, and M. J. Barnes, "Situation Awareness-Based Transparency," ARL-TR-6905, Apr. 2014.

[24] K. L. Mosier and L. J. Skitka, "Human Decision Makers and Automated Decision Aids: Made for Each Other?," in Automation and human performance Theory and applications, R. Parasuraman and M. Mouloua, Eds. NJ: Lawrence Erlbaum, 1996, pp. 201-220.

[25] C. W. Fisher and B. R. Kingma, "Criticality of data quality as exemplified in two disasters," Information and Management, vol. 39, pp. 109-116, 2001.

[26] J. Zhou, M. A. Khawaja, Z. Li, J. Sun, Y. Wang, and F. Chen, "Making machine learning useable by revealing internal states update - a transparent approach," International Journal of Computational Science and Engineering, vol. 13, no. 4, pp. 378-389, 2016.

[27] T. Muhlbacher, H. Piringer, S. Gratzl, M. Sedlmair, and M. Streit, "Opening the Black Box: Strategies for Increased User Involvement in Existing Algorithm Implementations," IEEE Trans. Visual. Comput. Graphics, vol. 20, no. 12, pp. 1643-1652.

[28] J. Bae, E. Ventocilla, M. Riveiro, T. Helldin, and G. Falkman, "Evaluating Multi-attributes on Cause and Effect Relationship Visualization," presented at the International Conference on Information Visualization Theory and Applications, 2017, pp. 64-74.

[29] V. Bellotti and K. Edwards, "Intelligibility and Accountability: Human Considerations in Context-Aware Systems," Human-Computer Interaction, vol. 16, no. 2, pp. 193-212, 2001.

[30] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," presented at the the 11th international conference, New York, New York, USA, 2009, p. 195.

[31] A. S. Clare, M. L. Cummings, and N. P. Repenning, "Influencing Trust for Human-Automation Collaborative Scheduling of Multiple Unmanned Vehicles," Human Factors, vol. 57, no. 7, pp. 1208-1218, Oct. 2015.

[32] E. Shearer and J. Gottfried, "News Use Across Social Media Platforms 2017," Pew Research Center, Sep. 2017.

[33] Adobe Inc., "Digital Intelligence Briefing: 2018 Digital Trends," Adobe Inc., Feb. 2018.

[34] A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," Science, vol. 185, no. 4157, pp. 1124-1131, Sep. 1974.

[35] S. Fiske and S. Taylor, Social Cognition. Reading, MA: Addison Wesley, 1991.

[36] J. B. Horrigan, "How People Approach Facts and Information," Pew Research Center, Aug. 2017.

[37] L. E. Blume and D. Easley, "Rationality," in The New Palgrave Dictionary of Economics, S. Durlauf and L. E. Blume, Eds. 2008.

[38] J. Preece, H. Sharp, and Y. Rogers, Interaction Design: Beyond Human Computer Interaction, 4 ed. Wiley, 2015, pp. 1-551.

[39] B. Schwartz, The paradox of choice: Why more is less. Harper Perennial, 2004.

[40] Rose, "Human-Centered Design Meets Cognitive Load Theory: Designing Interfaces that Help People Think," pp. 1-10, Oct. 2006.

[41] M. Zahabi, D. B. Kaber, and M. Swangnetr, "Usability and Safety in Electronic Medical Records Interface Design: A Review of Recent Literature and Guideline Formulation.," Human Factors, vol. 57, no. 5, pp. 805-834, Aug. 2015.

[42] S. Gregor and I. Benbasat, "Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice," MIS Quarterly, vol. 23, no. 4, p. 497, Dec. 1999.

[43] D. L. McGuinness, A. Glass, M. Wolverton, and P. P. Da Silva ExaCt, "A Categorization of Explanation Questions for Task Processing Systems.," presented at the AAAI Workshop on Explanation- Aware Computing ExaCt-, 2007.

[44] A. Ram, AQUA: Questions that Drive the Explanation Process. Lawrence Erlbaum, 1993.

[45] M. S. Silveira, C. S. de Souza, and S. D. J. Barbosa, "Semiotic engineering contributions for designing online help systems," presented at the the 19th annual international conference, New York, New York, USA, 2001, p. 31.

[46] S. Banasick, "Ken-Q Analysis."

[47] J. K. Ford, R. C. MacCallum, and M. Tait, "The Application of Exploratory Factor Analysis in Applied Psychology: A critical review and analysis," Personnel Psychology, vol. 39, no. 2, pp. 291-314, Jun. 1986.

[48] W. Yuji, "The Trust Value Calculating for Social Network Based on Machine Learning," presented at the 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2017, pp. 133-136.

[49] R. S. Amant and R. M. Young, "Interface Agents in Model World Environments," AI Magazine, vol. 22, no. 4, p. 95, Dec. 2001.

[50] J. Devore, Probability and Statistics for Engineering and the Sciences, Fourth. Brooks/Cole, 1995.