Modeling User Information Needs to Enable Successful Human-Machine Teams: Designing Transparency for Autonomous Systems

Chapter · July 2020

DOI: 10.1007/978-3-030-50439-7_31

E.s. Vorm
Naval Air Warfare Center Aircraft Division

Modeling User Information Needs to Enable Successful Human-Machine Teams: Designing Transparency for Autonomous Systems

Eric S. Vorm
Laboratory for Autonomous Systems Research
US Naval Research Laboratory
Washington, DC
eric.vorm@nrl.navy.mil

Andrew D. Miller
Department of Human-Centered Computing
Indiana University Purdue University Indianapolis
Indianapolis, IN
andrewm@iupui.edu

***Abstract*—Intelligent autonomous systems are quickly becoming part of everyday life. Efforts to design systems whose behaviors are transparent and explainable to users are stymied by models that are increasingly complex and interdependent, and compounded by an ever-increasing scope in autonomy, allowing for more autonomous system decision making and actions than ever before. Previous efforts toward designing transparency in autonomous systems have focused largely on explanations of algorithms for the benefit of programmers and back-end debugging. Less emphasis has been applied to model the information needs of end-users, or to evaluate what features most impact end-user trust and influence positive user engagements in the context of human-machine teaming. This study investigated user information preferences and priorities directly by presenting users with an interaction scenario that depicted ambiguous, unexpected, and potentially unsafe system behaviors. We then elicited what features these users desired most from the system to resolve these interaction conflicts (i.e., what information is most necessary for users to trust the system and continue using it in our described scenario). Using factor analysis, we built detailed user typologies that arranged and prioritized user information needs and communication strategies. This typology can be adapted as a user model for autonomous system designs in order to guide design decisions. This mixed methods approach to modeling user interactions with complex socio-technical systems revealed design strategies which have the potential to increase user understanding of system behaviors, which may in turn improve user trust in complex autonomous systems.**

***Index Terms*—Human-Machine Teaming; Autonomous Systems; User-Centered Design; Interaction Design; Transparency; Artificial Intelligence**

I. Introduction

In this paper, we sought to capture the questions users ask when the autonomous systems they are interacting with behave unexpectedly or produce unexpected results. Through this modeling process, we argue that we can derive design strategies that better support user information needs (i.e., what information could an interface provide that would answer user questions in such a scenario). Our theoretical position is informed by a wealth of findings from both laboratory studies and real world mishap investigations [1,2,3,10,13] which outline numerous examples of how these off-nominal, unexpected system behaviors can have very serious consequences on human decision making. Thus, we argue that by improving how autonomous systems communicate in a way that aligns with user expectations and priorities, we can improve user trust, and strengthen the overall human-machine team environment. To accomplish our modeling task, we used a mixed method called Q methodology, a technique in which participants assess a bank of questions and rank them in a normalized distribution. These rankings can then be quantitatively analyzed using factor analysis in order to find commonalities amongst the factor groups. We ran a human-in-the-loop study with 110 participants in the US and UK, using the same interaction scenario. Our findings resulted in a detailed analysis of a range of user information needs and priorities. With this information we built a detailed user typology, which corresponds to the four distinct user types found in our data. In subsequent sections, we detail and describe each user type, discuss the ramifications and uses of building such a typology, and discuss how utilizing this approach to user interaction modeling can help guide the design of intelligent systems which are transparent and intelligible to their users. This work will ultimately help designers tailor intelligent systems based on the desires and priorities of users.

II. Background and Motivation

Generating explanations that are meaningful and relevant to lay users is complicated by a variety of factors, both psychological and technological. Models that are able to be explained and understood by humans are said to be ”intelligible” [5]. Intelligibility is a major component of the umbrella term of ”transparency,” which has lately come to refer to both the degree to which a system’s inner workings can be seen and understood by the user, as well as other factors such as fairness, accountability, and privacy. For this work, we focus on the intelligibility component of intelligent system transparency, particularly in the role it plays in helping end users understand and trust intelligent autonomous systems.

A good deal of research has been done towards developing methods to make models more intelligible [6-8]. Gregor and

Benbasat [1] presented a detailed review of explanation types, and identified a set of useful constructs used to generate explanations to users of early intelligent systems. These include: trace or line of reasoning (explaining why certain decisions were or were not made by reference to the underlying data and rule base), justification or support (linking ”deep” domain knowledge to portions of a procedure, such as providing a textbook reference or hyperlink to explore deeper), control or strategic (explaining the system’s behavior by providing its problem solving strategies and reasoning rules), and terminology (providing users with term definitions to aid in their comprehension).

These constructs have been used broadly to enhance user understanding and trust in intelligent systems with some success. Studies have found that users who consider systems to be “intelligible” tend to perform better on system tasks, demonstrate more appropriate trust (generally defined as knowing when and when not to use the system, depending on the circumstances), and often report higher levels of usability and satisfaction during interaction [3,9,10,11]. Exploring intelligibility in context-aware systems, Lim, et al. [7] examined how different explanation types (Why, Why Not, What If, and How To) had an impact on user trust, performance, and comprehension of system functions. Participants interacted with different context-aware systems and were shown basic input-output cycles, along with different reasoning traces in the form of the questions above. Participants were then measured on their understanding of how the system functioned. They found that the Why and Why-Not explanations improved participants’ task performance and understanding, as well as increased their trust in the system, while neither the How-To nor What-If explanations showed any improvements over no explanations. Interestingly, there were no significant differences between Why and Why-Not intelligibility types. These findings suggest that most users have a basic desire to understand system behaviors and inner workings, and that answering these user questions in the correct format plays a central role in determining the interaction outcome.

Despite these findings, however, much of the work in the realm of explainability and intelligibility has been done to produce what some might call ”back end” explanations, or explanations that would be considered intelligible by programmers and experts, but few others. This is not an indictment against these efforts - they are vital to the development of safe and effective algorithms, and methods to enable programmers to better validate models and detect potential unsafe deviations from operating parameters are critical. It is important to note, however, that the human user comprises at least 50% of the human-machine system, at least in principal, and so should receive commiserate levels of modelling and consideration towards system design. In reality, it could be argued that humans comprise a considerably higher portion of the variability in system performance since it is their interactions with systems that ultimately determine much of the system’s output. Yet a cursory review of the scientific literature on explainability and transparency reveals that user-centered studies focusing on the development of design strategies for autonomous systems are squarely in the minority, with most attention given to algorithm development, visualization techniques, or debugging efforts [8,23,27].

This problem is not unique to today’s advanced autonomous systems. Early intelligent systems such as expert systems were only able to provide the most basic of explanations. These tended to be focused on verbalizing internal states, goals, and plans. These explanations were interpreted from a knowledge base, which limited their ability to answer questions from a static dataset and often bore little resemblance to human language [14]. As intelligent systems matured further, more sophisticated attempts to provide explainations emerged, and began to incorporate some degree of justification, offering not only the *what*, but also the *why* [15]. These systems offered explanations that were both understandable and satisfying, although only in limited scope. The newest generation of intelligent systems now strive to consider a variety of factors in their explanation capabilities, including the decision context, knowledge of the user, knowledge of the history of system performance, including reliability; modeling and knowledge of the goals of the user, and awareness of the domain [9].

Determining which of these features to include in an explanation, at what time, and in what format is still the subject of much investigation. Our own recent work has sought to explore methods of providing explanations by category (system parameters and logic, different qualities of data, how user personalization plays a role, providing some justification of why one option was recommended over another, and what other users have done in similar circumstances before) as a means of improving transparency and intelligibility of intelligent system recommendations [16]. While much work remains focused on developing methods to make models explain themselves to users, we argue that much of this model-based information is superfluous to most end-users, and that by prioritizing information that users care about most, system designs can achieve a higher net effect in terms of user comprehension, trust, and perceived usability. In many cases, these data already exist in the underlying system architecture, which means making them available to users through an interface is often relatively inexpensive and simple. Tracing decisions made by systems, especially in examples such as recommender systems, is also relatively easily accomplished. Most design decisions such as these, however, tend to be made out of pragmatic considerations. In other words, most designs of autonomous intelligent systems are driven by an effort to reduce clutter and streamline interface layout. In many cases, these decisions are made according to the priorities of the designers, rather than the end users, a phenomenon Cooper termed ”the inmates running the asylum” [12].

This means that in order to truly design an autonomous system interface that supports user trust and decision making, we must first determine what information users think is most important, especially in the context of highly complex autonomous, distributed systems that have the capability of choosing behaviors with little to no user input. Our purpose for this research is to assess the relative value or importance of various bits of information that would be potentially available in an interface with an autonomous system, and to quantify these priorities in a manner that supports future system designs.

To do this, we employed a method known as Q-methodology. Q-methodology is a mixed method, often referred to as the "scientific study of subjectivity" [17] and has been successfully used in previous HCI work [18] to elicit design feedback from stakeholders. Q-methodology asks users to sort statements or questions according to their preferences or priorities into a fixed matrix that represents a normalized distribution. Using a factor analytic approach, Q-methodology can identify patterns of subjectivity and thought in the data, which is used to identify groups, or clusters of people who share similar opinions and ways of thinking about a given issue. Interpretation and classification of these clusters is then made using traditional factor analysis techniques, effectively combining the strengths of both qualitative and quantitative research. The results from this method are data that is deep in texture and nuance that could otherwise be passed over by a purely statistical or survey sampling approach. We describe the steps of our methodology in the following section.

III. Methods

Since we are interested in understanding what features users find more or less valuable to help them understand and trust intelligent autonomous systems, we needed a way for them to consider and prioritize a large number of design elements. To do this, we created a bank of questions that users might ask of a system when it behaves unexpectedly or uncharacteristically. The motivation behind this decision was that by asking participants to identify what questions *they* would ask, we could more accurately model what information they considered vital to their decision making. This approach, we argued, would minimize issues common to survey or ethnographic methods (i.e., bias, response interpretation), while still allowing for detailed user priorities to be captured in a quantifiable and reliable manner.

These questions were deliberately developed to represent a variety of approaches to providing users critical information that could help them resolve conflicts with interactions in intelligent autonomous systems. In order to narrow down the potential list of conflicts, we chose to focus our study on interactions with intelligent recommender systems, or systems that provide recommendations to users (i.e., decision support algorithms). For example, in a basic recommender system scenario, if a user was presented with a restaurant recommendation which seemed out of place for their tastes, the user would probably want to know *how* that recommendation was made. To answer *how*, however, the user could be given a variety of information. For example, they may care to learn what data was used to create that recommendation, and in doing so better understand the recommendation. Or perhaps they might want to know whether or not the system actually has a model of themselves and their tastes, or whether the recommendation was made randomly. Because there are a variety of potential answers to "how was this recommendation made?" where some of those answers would be more valuable and satisfying than others to individual users, it is important that we try to model these so that our interface designs provide answers that are meaningful to the intended users of the system, rather than system architects. For some hypothetical examples of how system designs can be informed by these different user information priorities, see figures 3, 4 and 5 in section V.

In our approach, each question we developed is mapped to a potential design feature that could be achieved through an interface. Our purpose was to help determine which potential design features would most help users understand and trust intelligent autonomous systems, and which would be considered a nuisance or irrelevant.

A. Question Bank Development

Because the question bank we developed from an earlier project [16] was not specific to any one type of system, we first had to develop questions that would be most appropriate for interactions with our intended system in this project. To do this we started with Ram's taxonomy of question types as an initial starting point to ensure that we used a variety of question types [19]. Ram's taxonomy is useful because it describes a wide breadth of questioning strategies, and was developed explicitly to enhance the explainability of intelligent systems to end users. We refined these questions using Silveira et al's taxonomy of user's frequent doubts [20]. After iterative evaluation and consultation within the project team and with experts in intelligent system design in the US and UK, we arrived at an initial bank of 36 questions.

Figure 1. Our interactive testbed for this project was the Deep Securities and Accounting Management (DSAM) system. This system emulated an intelligent autonomous system that provides recommendations to its users.

Once the set of questions was developed, we presented participants with the Deep Securities and Accounting Management (D-SAM) system (figure 1). D-SAM is a research testbed, and was developed by reviewing recent submissions to the United States Patent and Trademark Office's Patent Full-Text and Image Database (PatFT). By exploring recent patent submissions, and combining these with our knowledge of intelligent autonomous systems research, we developed a near-future, plausibly relevant financial management system that embodies many of the most advanced efforts in intelligent systems today, and assumes their success in the near future. Our users were asked to interact with D-SAM, which resulted in a system-generated recommendation the user had to determine whether to accept or reject. This interaction deliberately introduced ambiguity and uncertainty into the scenario in the form of an unexpected or seemingly inappropriate recommendation. This ambiguity and unexpected system behavior is the most common combination found to result in significant user conflicts with intelligent systems [10], and thus served to create the need for users to seek additional information from the system in order to determine whether or not it could be trusted, or if its recommendation should be disregarded. This ambiguity and unexpected system behavior is one in which the concept of transparency is theoretically most critical, hence we used it to frame our study.

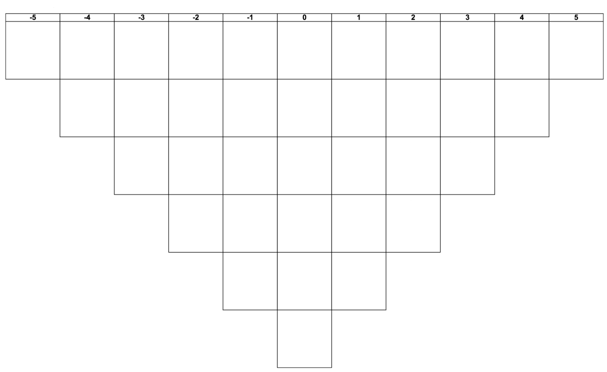

Figure 2. The forced distribution matrix. Cards are arranged from right (most important to me, +5) to left (least important to me, -5).

Once presented with the interaction scenario, participants were then given a stack of 36 numbered index cards, each containing a different question. Each question on the card was meant to represent a different information seeking strategy. The concept behind this is that a user could potentially ask the system a wide variety of questions about all manners of different things. Our goal was to better understand what questions were more or less frequently prioritized as important to users, as a proxy of inferring their information priorities and preferences. Some example questions were ”How current is the data used in making this recommendation?”, ”Precisely what information about me does the system know?”, and ”What have other people like me done in response to this recommendation?” Participants were then given time to sort these 36 cards into the fixed distribution matrix described in figure 2 above. Participants were encouraged to consider each question as if it were something they would ask themselves, and then to determine which questions, if answered, would have the greatest impact on trust and their willingness to act on recommendations generated by the system.

Once cards were sorted, participants recorded their arrangements on a paper form, and answered two additional questions: ”Briefly describe why you chose this question as your most/least important question to ask.” Paper forms were then collected and prepared for analysis and interpretation.

Each column in the matrix in figure 2 is given a number value, corresponding to the degree of preference expressed by each participant- +5 for the rightmost column corresponding to “most important to me,” -5 for the leftmost column corresponding to “least important to me,” and everything in between. Each participant’s sort then represents a full arrangement of their preferences in a forced and normalized distribution. Once completed, each participant sort was arranged in a correlation matrix with each other participant sort. This matrix was then submitted to factor analysis.

Using principal components analysis (PCA) for factor extraction [21], we extracted 8 initial factors. We tested several possible solutions, ranging from two to eight factor groups, and ultimately settled on a four factor solution because together they explained the majority of variance (61%), and divided the majority of respondents into a relatively small number of groups that were distinct from one another, yet large enough to permit statistical analysis. We then used the VARIMAX method to obtain optimal rotation [22]. 11 of the participant’s arrangements were confounded because they loaded on more than one factor, and 18 participants failed to load on any of the four factors we extracted. This resulted in four distinct viewpoints of information priorities and preferences of the remaining 89 individuals.

B. Factor Interpretation

Once factor extraction and rotation was complete, we next set about analyzing how each factor group arranged their questions in order to intuit and interpret their reasoning and prioritization strategy. To accomplish this, we produced a weighted average of each participant’s arrangement of cards from within their factor group, and combined those arrangements into one exemplar composite arrangement per group, also known as a ”factor array.” We then compared each group factor array to one another in order to derive a statistical basis of comparison. By examining the placement of each question within each factor array and comparing those arrangements to each other factor array, we can begin to detect patterns, which can be used to infer how and why these clusters of individuals prioritize and value information differently.

To do this, we examined each factor array’s **distinguishing questions**. A distinguishing question is found when the participants in a factor group place a question in a significantly different position from all participants loading on other factors. For example, the highest ranked question from factor group two was ”What is the history of reliability for this system?” (composite score 5, $Z = 1.85$, $p < 0.01$). This question was placed significantly higher than any other factor group, thus partially defining factor group two. By examining distinguishing questions for each factor group, we began to uncover unique differences amongst the groups, and to describe how each group prioritized information differently.

Finally, in order to fully appreciate our findings, we examined participants’ qualitative feedback to contextualize and verify our analyses. This feedback was solicited from participants in the form of two questions which they answered in open comments on the data collection sheet, ”Briefly describe why you chose this question as the MOST/LEAST important question to you."

The result of this factor analysis is a detailed user typology that sorts participants into four cohesive, like-minded groups based on their shared priorities, reasoning strategies, and patterns of thought. In the following section we describe our findings from each of the four user typologies, and later discuss the potential implications of these findings in the greater design space of intelligent system transparency and intelligibility in section V.

IV. RESULTS

We identified four distinct user typologies for intelligent system transparency. In this section, we describe each group based on its quantitative features, and then provide an analytical interpretation of the characteristics associated with people in the group. A summary table of findings is available in table II, while a detailed table of all findings by question type is available in the Addendum.

A. Factor Group 1- "Interested & Independent"

Factor group one was defined by 24 participants and explained 14% of the total study variance with an eigenvalue of 20. 71% reported they had little to no working knowledge of intelligent systems. Roughly 60% of factor group one were less than 40 years old. Individuals in this group most want to know **"why was this recommendation the BEST option,"** indicating a desire for some sort of justification for why a recommendation was made, above and beyond a basic explanation (composite score 5, Z = 1.42, p <0.05). Individuals in this group also demonstrated an interest in some of the underlying components of how systems function, and would like to know **"What if I decline? How will that decision be used in future recommendations by this system?"** (composite score 4, Z = 1.29, p <0.01) and **"Can I influence the system? Will it consider my input?"** (composite score 3, Z = 1.06, p <0.01).

Individuals in factor group one were least interested in the opinions or behaviors of others when considering what to do with a computer-generated recommendation. They ranked questions like **"Is there anyone in my social network that has received a similar recommendation"** (composite score -5, Z = -2.1, p <0.01), **"How many other people have accepted or rejected this recommendation from this system"** (composite score -4, Z = -1.8, p <0.01), **"How similar am I to other people who have received this recommendation"** (composite score -4, Z = -1.58, p <0.01), and **"What have other people like me done in response to this recommendation"** (composite score -3, Z = -1.57, p <0.01) as their least important or valuable questions.

Because of their preference for deep system information, and their reluctance to place any priority on other users' behaviors or decisions, we named this factor group the "interested and independent" group. This descriptive name serves to differentiate factor group one from the other groups, as well as to identify a general information seeking strategy found amongst our data.

Factor Characteristics				
	Factor 1	Factor 2	Factor 3	Factor 4
No. of Defining Participants	24	16	24	17
Avg. Rel. Coef.	0.8	0.8	0.8	0.8
Composite Reliability	0.99	0.985	0.99	0.986
S.E. of Factor Z-scores	0.1	0.122	0.1	0.118
Eigenvalue	20	15.34	8.07	7.16
Explained Variance	14%	11%	12%	9%
Male [1]	67%	94%	67%	76%
Female	33%	6%	33%	24%
Experts	29%	75%	29%	41%
Novices	71%	25%	71%	59%
20-30 yrs old	46%	18.8%	58%	53%
30-40 yrs old	12.5%	43.8%	21%	29%
40-50 yrs old	29%	25%	8.5%	12%
50+ yrs old	12.5%	12.5%	12.5%	6%

Table I

CHARACTERISTICS OF FACTORS AFTER ROTATION. FACTORS DEFINE CLUSTERS OF PARTICIPANTS WHOSE ARRANGEMENT OF QUESTIONS WERE VERY SIMILAR, AND WERE MATHEMATICALLY CLUSTERED USING FACTOR ANALYSIS. WE REFER TO THESE AS "FACTOR GROUPS," OR BY THEIR GIVEN TYPOLOGICAL NAMES THROUGHOUT THE DURATION OF THE PAPER. FACTOR 1: INTERESTED & INDEPENDENT; FACTOR 2: CAUTIOUS & RELUCTANT; FACTOR 3: SOCIALLY INFLUENCED; FACTOR 4: EGOCENTRIC.

B. Factor Group 2- "Cautious & Reluctant"

Factor group two was defined by 16 participants and explained 11% of the study variance with an eigenvalue of 15.34. 94% were male, 64% were less than 40 years old, and 3/4 had extensive working knowledge of intelligent systems.

This group was exemplified by a deep concern over a system's past performance and reliability. For example, they most wanted to know **"What is the history of the reliability of this system?"** (Composite score 5, Z = 1.85, p<0.01), followed by **"Under what circumstances has this system been wrong in the past?"** (Composite score 4, Z = 1.4, p<0.01) and **"What data does the system depend on in order to work properly, and do we know if those dependencies are functioning properly?"** (Composite score 3, Z = 1.19, p<0.05). This group also appeared very interested in information that could help them gauge how the system considers uncertainty and risk, as exemplified by their high ranking of questions like **"How much uncertainty does the system have?"** (Composite score 3, Z = 1.12, p <0.01) and **"How does the system consider risk, and what is its level of acceptable risk?"** (Composite score 2, Z = 1, p <0.01).

Participants in factor group two were least interested in whether **"Is there anyone in my social network that has received a similar recommendation?"** (composite score -5, Z = -1.69, p <0.05). They also thought little of questions such as **"What does the system *think* I want to achieve? (How does the system represent my priorities and goals)"** (composite score -4, Z = -1.59, p <0.01), **"Can I influence the system by providing feedback? Will it listen and consider my input?"** (composite score -4, Z = -1.42, p <0.01), and **Was this recommendation made specifically for ME?"** (composite score -3, Z = -1.32, p <0.01).

[1] All participants identified as either male or female

Because the nature of questions prioritized by factor group two seemed to revolve around the kinds of information that could aid in validating that a system was operating normally, we named this group "Cautious and Reluctant." This group seemed to be the least willing group to interact with and perhaps most suspicious of intelligent autonomous systems, based on their information priorities. Thus this group would represent a particularly vulnerable user group whose needs would most need to be considered in the design of a system such as D-SAM.

C. Factor Group 3- "Socially Influenced"

Factor group three was defined by 24 participants and explained 12% of the study variance with an eigenvalue of 8.07. 67% were male, 79% were less than 40 years old, and 71% had little to no working knowledge of intelligent systems. Participants in this group most wanted to know **"Why is this recommendation the *best* option?"** (composite score 5, Z = 1.75, p <0.05) followed closely by **"What are the pros/cons associated with this option?"** (composite score 4, Z = 1.25, p <0.01). They also indicated an interested in learning what others have done by ranking **"What is the degree of satisfaction that others have expressed when taking this recommendation?"** (composite score 3, Z = 0.9, p <0.01), and **"How many other people have accepted or rejected this recommendation from this system? (What is the ratio of approve to disapprove?)"** (composite score 1, Z = 0.29, p <0.01) higher than any other factor group.

Participants in this group were least interested in knowing anything about the qualities of data used by the system. Questions like **"What is the signal-to-noise ratio of this data?"** (composite score -5, Z = -2.34, p <0.01), **"Can I see the data for myself?"** (composite score -4, Z = -2.22, p <0.01), **"How much data was used to train this system?"** (composite score -4, Z = -1.53, p <0.01), and **"Is the system working with solid data, or is the system inferring or making assumptions on 'fuzzy' information?"** (composite score -3, Z = -1.43, p <0.01) were all ranked lowest by this factor group.

Analyzing the priorities of this factor group revealed a pattern of preferences related to the behaviors and decisions of other users. While their highest rated questions revolved around understanding the recommendation itself, this group also highly ranked questions related to what other users have done. Relative to other groups, this group was the only one that considered this kind of information relevant or important. Given the majority of these participants were less than 40 years old, these findings could potentially indicate a user group with a posture towards intelligent systems that incorporates social components of usage, suggesting the increasing importance of utilizing features that provide this information in intelligent system designs.

D. Factor Group 4- "Egocentric"

Factor group four was defined by 17 participants and explained 9% of the study variance with an eigenvalue of 7.16. 76% were male, 82% were less than 40 years old, and expertise was almost evenly split between 59% who had little to no working knowledge of intelligent systems, and 41% who had extensive working knowledge of intelligent systems. Participants in this group appear most interested in understanding how recommendations relate to themselves, and others like them. Their top ranked question 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.)?"** (composite score 5, Z = 2.6, p <0.01), followed by **"Precisely what information about ME does the system know?"** (composite score 4, Z = 1.25, p <0.01), **"What have other people like ME done in response to this recommendation?"** (composite score 3, Z = 1.22, p <0.01), **"How many other people like ME have received this recommendation from this system?"** (composite score 3, Z = 1, p <0.01), and **"Is there anyone in my social network that has received a similar recommendation?"** (composite score 3, Z = 0.98, p <0.01).

Participants in this group appeared not to care much for details about other options, or how the system considers the concept of risk. They ranked **"What are the pros/cons associated with this option?"** (composite score -5, Z = -1.99, p <0.01), **"How does the system consider risk, and what is its level of acceptable risk?"** (composite score -4, Z = -1.63, p <0.01), **"Are there any other options not presented here?"** (composite score -4, Z = -1.42, p <0.01), **"How many other options are there?"** (composite score -3, Z = -1.21, p <0.01) and **"What does the system think is MY level of acceptable risk?"** (composite score -3, Z = -1.17, p <0.01) as least important to them.

Interpretations of this group's information priorities revealed a clear preference for self-referential information. Accordingly, we named this group the "Egocentric" group. While the egocentrics were the smallest of our four factor groups, their unambiguous preferences indicated a clear strategy in decision making. When faced with unusual or unexpected results from an intelligent system, at least in our recommendation scenario, these individuals consider themselves in the equation, and consider answers to these questions most important to help them understand and trust system outputs.

V. Discussion

Thus far, we have demonstrated four distinct differences in user information preferences when interacting with unusual or unexpected recommendations from an intelligent financial planning system. These differences characterize different ways that users might seek to resolve conflicts with intelligent systems, especially when faced with unusual, unexpected, or ambiguous system behaviors. In our study we designed an interaction scenario where our participants were presented with a recommendation from a financial management system, and that recommendation seemed potentially unsafe or inadvisable enough that users would need additional information in order to determine whether or not to accept and act on

RELATIVE RANKINGS OF QUESTIONS FOR ALL FACTOR GROUPS

Interested & Independent	Interested & Independent	Cautious & Reluctant	Socially Influenced	Egocentric
Highest Ranked				
Why is this recommendation the best option?	5	2	5	1
Questions ranked higher in this array than any other factor array				
Can I influence the system by providing feedback? Will it listen and consider my input?	3	-4	-2	2
Can I see the data for myself?	2	0	-4	-3
What if I decline? How will that decision be used in future recommendations by this system?	0	-3	-3	-2
Questions ranked lower in this array than any other factor array				
How many other people have received this recommendation from this system?	-3	-2	0	3
What have other people like me done in response to this recommendation?	-3	-1	0	3
How similar am I to other people who have received this recommendation?	-4	-1	2	1
How many other people have accepted or rejected this recommendation from this system?	-4	-3	1	-1
Lowest Ranked				
Is there anyone in my social network that has received a similar recommendation?	-5	-5	1	3

Cautious & Reluctant	Cautious & Reluctant	Interested & Independent	Socially Influenced	Egocentric
Highest Ranked				
What is the history of the reliability of this system?	5	-1	1	-1
Questions ranked higher in this array than any other factor array				
Under what circumstances has this system been wrong in the past?	4	-2	0	0
What data does it depend on, and do we know if those dependencies are functioning properly?	3	1	0	2
How much uncertainty does the system have?	3	0	1	0
How does the system consider risk, and what is its level of "acceptable risk?"	2	-1	-1	-4
Questions ranked lower in this array than any other factor array				
Was this recommendation made specifically for *ME*, or on something else?	-3	2	4	5
Can I influence the system by providing feedback?	-4	3	-2	2
What does the system *think* I want to achieve?	-4	3	2	0
Lowest Ranked				
Is there anyone in my social network that has received a similar recommendation?	-5	-5	1	3

Socially Influenced	Socially Influenced	Interested & Independent	Cautious & Reluctant	Egocentric
Highest Ranked				
Why is this recommendation the best option?	5	5	2	1
Questions ranked higher in this array than any other factor array				
What are the pros/cons associated with this option?	4	2	-1	-5
Does the system know and understand my goals?	3	3	-1	1
How satisfied are others who took this recommendation?	3	-2	-2	-1
Questions ranked lower in this array than any other factor array				
Is it working with clean, or 'fuzzy' data?	-3	-2	-2	-2
How much data was used to train this system?	-4	-1	0	-2
Can I see the data for myself?	-4	2	0	-3
Lowest Ranked				
What is the signal to noise ratio of this data?	-5	-3	0	-2

Egocentric	Egocentric	Interested & Independent	Cautious & Reluctant	Socially Influenced
Highest Ranked				
Was this recommendation made specifically for *ME*, or on something else?	5	2	-3	4
Questions ranked higher in this array than any other factor array				
Precisely what information about me does the system know?	4	0	-2	0
What have other people like me done in response to this recommendation?	3	-3	-1	0
How many other people have received this recommendation from this system?	3	-3	-2	0
Is there anyone in my social network that has received a similar recommendation?	3	-5	-5	1
Questions ranked lower in this array than any other factor array				
How many other options are there?	-3	-1	0	-1
Are there any other options not presented here?	-4	0	1	0
How does the system consider risk, and what is its level of "acceptable risk?"	-4	-1	2	-1
Lowest Ranked				
What are the pros/cons associated with this option?	-5	2	-1	4

Table II

ALL FOUR FACTOR GROUPS, WITH THEIR IDENTIFYING QUESTIONS AND RELATIVE RANKINGS.

the recommendation, or to reject it. Our findings validate the argument that users use different reasoning strategies, and that those strategies can be quantified and described in sufficient detail to allow design recommendations to be created.

In this next section, we discuss the implications of these findings in terms of how they might be used to prioritize design elements, guide interface development, and structure communication strategies to promote effective human-machine teams. Our first step is to analyze questions that had near-universal consensus in our sample; that is, questions that nearly all users agreed were either very important, or very unimportant. These questions should be considered as most valuable in terms of design priorities, since they all had near full consensus in our sample. Next, we explore questions that produced the highest disagreement amongst all factor groups. These questions represent design elements that could please some users, while aggravating or confounding others. Since these questions were the source of much contention within our sample, we propose these as a starting point for interfaces. Next we analyze questions by category in order to extract valuable design insights and lessons learned, and show how our findings might be translated to design through illustrative interface mockups. We end with a discussion of the limitations of our study, and plans for future work.

A. Consensus Amongst Groups

While each factor group had identifying statements that distinguished it from others, there were some questions that all factor groups found either important or unimportant. These are known as consensus questions, or those that do not distinguish between ANY pair of factor groups.

All four of our factor groups thought the question **”What are all of the factors (or indicators) that were considered in this recommendation, and how are they weighted?”** as highly important (average score 3.75, Z score variance 0.06). That this question was the most agreed upon is not surprising, given that other studies have confirmed most individuals demand at least some degree of explanation and justification for system outputs in reference to automated recommendations [7]. Our participants also moderately valued **”What safeguards are there to protect me from getting an incorrect recommendation?”** (average score 1.5, Z score variance 0.031) across all factor groups. Despite the wide array of differences in information priorities and decision making heuristics we found amongst our participants, these two questions were agreed upon by all as having at least moderate importance for users of intelligent systems that make recommendations. These questions should therefore be considered highly valuable to answer through an interface, and those design elements should be considered a high priority in intelligent recommender systems such as the one described in our study.

On the other end of the spectrum, none of the factor groups found the questions **”Is my data uniquely different from the data on which the system has been trained?”** (average score -0.75, Z score variance 0.122), and **Is the system working with solid data, or is the system inferring or making assumptions on fuzzy information?”** (average score -2.25, Z score variance 0.109) as being very important or valuable to them. These questions are likely important to some people, such as programmers who may appreciate this granularity of information about the underlying data, but to end users they are unlikely to be very meaningful or to improve trust or acceptance. In contrast to the questions earlier, answering these questions through an interface would likely add either confusion or become an irritation to users of systems such as D-SAM. Examining what questions produced agreement from across all participants allowed us to quickly narrow down our potential design elements, illustrating a clear benefit of a mixed methods approach to user-centered design.

B. Disagreement

Just as we examined questions that all groups found equally important or unimportant to them, we also examined questions that produced the greatest disagreement between groups. These polarizing questions can help identify potential design elements that may be points of contention to some users. To analyze these questions in a way that is both detailed, yet practical, we arranged all questions into five categories, based on their similarity to one another. The first category was named parameters and logic, and describes aspects of system features that constrain its operations, such as how sensor data is used in deriving system outputs. The next category was named qualities of data, and describes features of relevance about data itself, such as its age, provenance, level of noise, etc. The next category was named user personalization, which describes how systems consider the user in deriving system outputs (this is especially relevant in recommender systems, such as the system we developed for this study). The next category was named justification of options, and describes how options are arranged, how they are prioritized by the system (again, this category of explanation is highly relevant to recommender systems that may generate several potential recommended options, but may only display one to the user). The final category was named social influence, which describes the behaviors and decisions of other users. This is a somewhat unique explanation strategy to recommender systems, commonly seen in music or movie recommendations (i.e., users who watched this show also enjoyed this other show). We will discuss how each factor group valued and prioritized these categories of questions in the sections below.

C. System Parameters & Logic

Questions explaining the inner workings of a system, including its reasoning, logic, policies, and limitations, were termed System Parameters & Logic. These questions produced a low degree of disagreement (average Z score variance 0.33) across all groups, with most questions averaging around the mean (score of 0). With the exception of the Cautious and Reluctant group (who were most interested in questions about reliability, uncertainty, and risk), all others found these questions to be of moderate to low importance, indicating them as medium to low priority design elements that are perhaps best delivered through menu options that can be accessed by those most interested. Designing explanations that provide information of this sort, therefore, is advisable, given that most participants, regardless of their factor group, ranked these questions moderately important.

D. Qualities of Data

Overall, questions pertaining to the qualities of data, such as age, noise and provenance generated moderate agreement between all factor groups (average Z score variance 0.419). Questions such as **How current is the data used in making this recommendation?”, ”How clean or accurate is the data used in making this recommendation?”,** and **How is this data weighted or what data does the system prioritize?** all averaged between 0-1 across all factor groups. It is important to note here that the forced distribution used for this experiment results in a mean score of 0. That these questions were all ranked around the mean indicates they are questions which the majority of stakeholders would like addressed in some form, plausibly in order to better understand and trust intelligent system recommendations.

Other questions related to the qualities of data, however, proved more divisive, and may be too much for some users to appreciate. As discussed in the section on Consensus, none of the factor groups found the questions **”Is my data uniquely different from the data on which the system has been trained?”** or **”Is the system working with solid data, or is the system inferring or making assumptions on fuzzy**

information?" very important to them, indicating a potential limit of the usefulness of displaying qualities of data as a means of improving intelligibility. While the Interested and Independent group demonstrated the most willingness and interest in these types of questions, none of the other factor groups was especially interested.

E. User Personalization

We termed questions aimed at helping users understand what of *their* data is known, and how that data is used to derive recommendations as User Personalization questions. This category generated a wider range of sentiment than questions about the qualities of data (average Z score 0.744), including the most divisive 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.)?"** On average, the Socially Influenced and Egocentric groups favored these types of questions more than the more analytical Interested and Independents, and Cautious and Reluctant. Examining user sentiment surrounding these questions helps perhaps to understand why variance was so high. For instance, people in the Cautious and Reluctant group commented things like *"I don't think 'me' is important... I need objective metrics!"*, whereas people in the Socially Influenced group expressed a different sentiment, *I want to know that the system has made the right choice for me and my lifestyle/preferences, and whether it has it really taken all my situations and personal feelings into consideration.* Yet, the recent increasing concern over potentially inappropriate collection and uses of personal data by social media and others, combined with the moderate rankings of many questions in our sample, such as **Does the system know and understand my goals?** (average score 1.5, Z score variance 0.51), and **Precisely what about me does the system know?** (average score 0.5, Z score variance 0.59), suggests new efforts should be made towards affording users information about and answers to these kinds of questions. Considering the strong prioritization of these questions by the Socially Influenced and Egocentrics, we strongly suggest designers consider making these affordances available wherever possible. To demonstrate one example of how some of these questions can be addressed in order to make systems and algorithms more transparent to users, we have provided a screen shot of a restaurant recommendation app currently under development, which will be used in follow on studies. Figures 3 and 5 demonstrate a few techniques which many users may find useful, and others, such as those aligned with the Socially Influenced or Egocentric groups, may soon demand.

1) Justification of Options: Closely related to explanations, justifications offer assertions about reasons for decisions or choices, examples, alternatives that are eliminated, or counterfactuals [23]. All factor groups in our study agreed that a justification of **Why this recommendation is the BEST option** is important and valuable to them (average score 3.25, Z score variance 0.66). Other questions related to justification

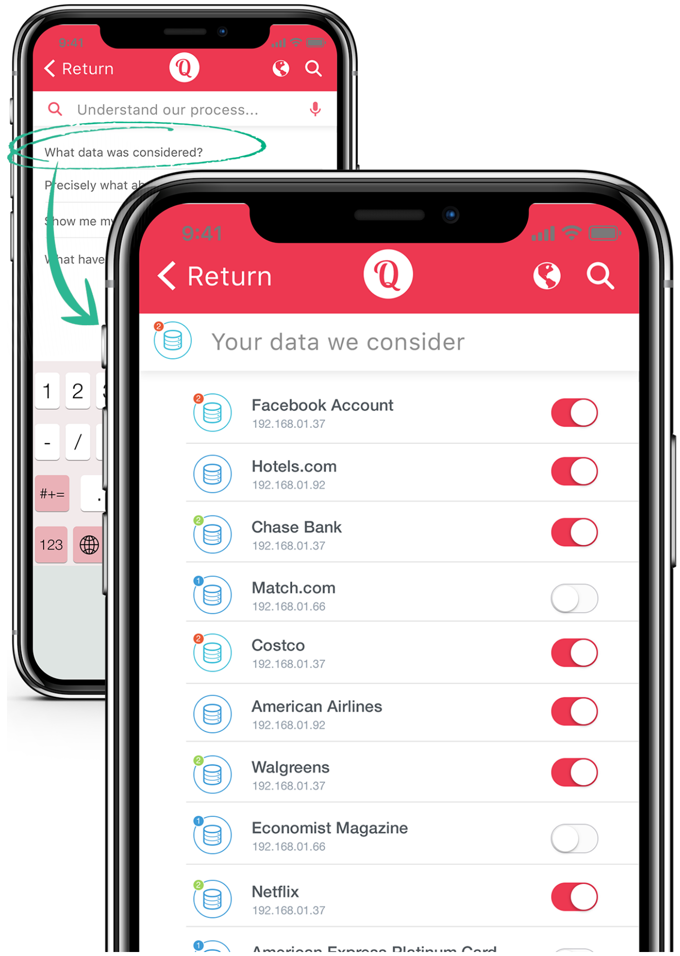

Figure 3. Knowing and controlling what personal data systems collect and use is a growing concern among users, and was a polarizing topic amongst participants in our sample. Enhancing the transparency of recommender algorithms based on user data is deeply significant to many, and should be afforded wherever possible.

of options were also agreed upon as *not* being valuable or useful to our factor groups, such as **Are there any other options not presented here** (average score -0.75, Z score variance 0.5), and **How many options are there?** (average score -1.25, Z score variance 0.25). These questions are likely too in depth for most stakeholders to appreciate, especially given that one of the principal reasons for leveraging decision support tools is to ease the burden of choice [24].

One question: **What are the pros and cons associated with this option?** produced a very high amount of variance between groups (average score 0, Z score variance 1.56). Both Interested and Independents (composite score 2) and Socially Influenced (composite score 4) felt this question was important to them, while the Cautious and Reluctant (composite score -1) and Egocentrics (-5) did not. Since the Interested and Independents and Socially Influenced were not significantly aligned on any other questions, it is worth exploring why they should both see this question as one they would like answered through an interface. Understanding the reasoning behind these user priorities is an important component of this

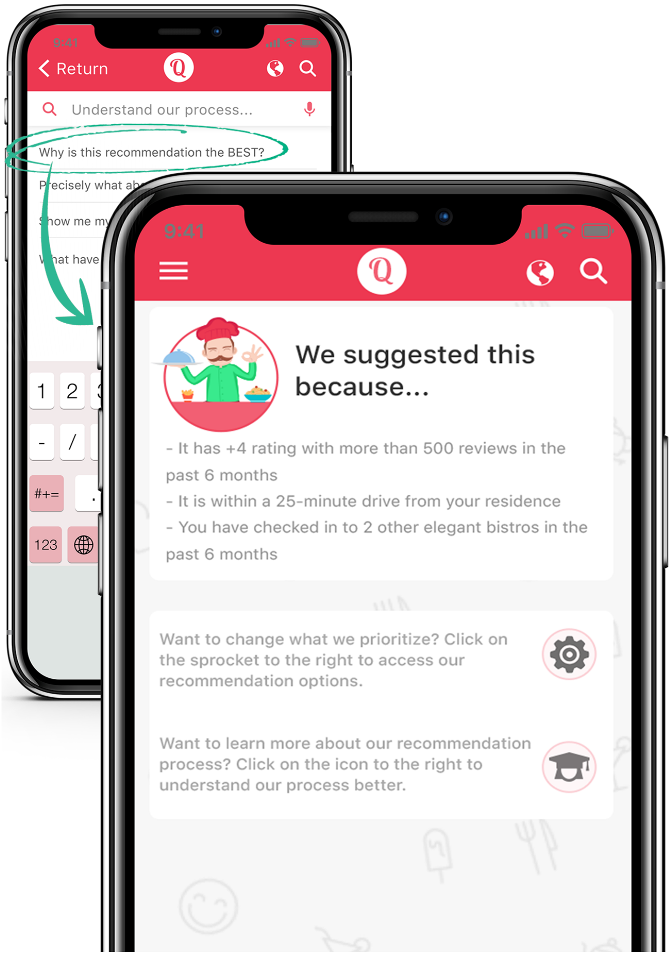

Figure 4. Explanation and justification of options was important to all groups in our study. Finer details in how best to justify options may require multifaceted approaches that include affording deeper analysis and education for those who need it.

research, and if we consider the above question in relation to what other questions these groups found valuable, we may better understand how designs can afford users answers that are meaningful to them.

In this case, while both Interested and Independents, and Socially Influenced want to know the pros and cons associated with a recommendation, precisely *how to answer that* is decidedly different. While the Socially Influenced are more likely to seek answers in the form of what other people report, such as user satisfaction metrics, Interested and Independents would prefer to understand what data was used and how it was weighted. Questions like the above are precisely those that motivate our research, since they have the potential to both confirm and confound user sentiment, depending on a variety of individual factors which are often difficult to measure.

To demonstrate how we perceive designers possibly addressing these challenges, we have provided figure 4, which demonstrates both a justification in plain English, as well as advanced controls which the user may use to re-prioritize how some algorithms work, and also access to deeper, more in-depth education about the system's inner workings for those like the Interested and Independent, who prefer this level of information.

2) *Social Influence:* We termed questions that pertained to the actions or opinions of others, or to how users are characterized and grouped with others as Social Influence questions. Questions in this category produced the greatest amount of disagreement between groups (average Z score variance 0.98), suggesting that as design elements they represent potentially polarizing options. Averaging all questions in this category, we see that the Egocentrics (average score 1.33) and Socially Influenced (average score 1.17) both consider this information valuable and useful to their decision making, while the Cautious and Reluctant (average score -2.33) and Interested and Independent (average score -3.5) clearly do not.

Socially-related information, such as how users are characterized and grouped into personas, and what other people like them have done in similar circumstances, is commonly used in current systems that offer recommendations, such as Netflix, Spotify, or Amazon (e.g., others who purchased this also bought XYZ). These features may improve decision making for some, like the Egocentrics, while they may be ignored by others, like the Interested and Independent. What is of potential interest, however, is how this type of information may soon be featured in other applications with greater scope.

There is considerable room for this kind of information to be considered useful, for instance, as crowd sourcing becomes a more common feature in several domains. There are already several notable examples, such as citizen science [25], personal wellness [26], and even app design [27] which make use of a community of distributed participants that collaborate to form something. These projects often feature consensus building activities that leverage the concept of "hive mind" or "wisdom of the crowd" to achieve common goals. While there are certainly limits to the use of crowd sourcing, especially in highly personalized domains such as clinical medicine or personal financial management, these approaches may very well become more commonplace as intelligent systems broaden and consume greater market presence in our everyday lives. Designers that choose to feature socially-related information into their products may well find those features appreciated and valued, especially as a younger technocentric generation assumes more of the user base.

F. Design Implications

Of the 36 questions in our sample bank, most were of value to one factor group or another, and (as we showed in the Consensus section above) very few were totally unimportant. For prospective designers of transparent intelligent systems, this presents something of a quandary. The most obvious solution–to present all data that could be relevant to someone–would result in impractical long lists of information that is not especially relevant to anyone.

Our approach has uncovered a detailed view of the different manner in which users reason about systems, and can help designers better understand how some explanations can have

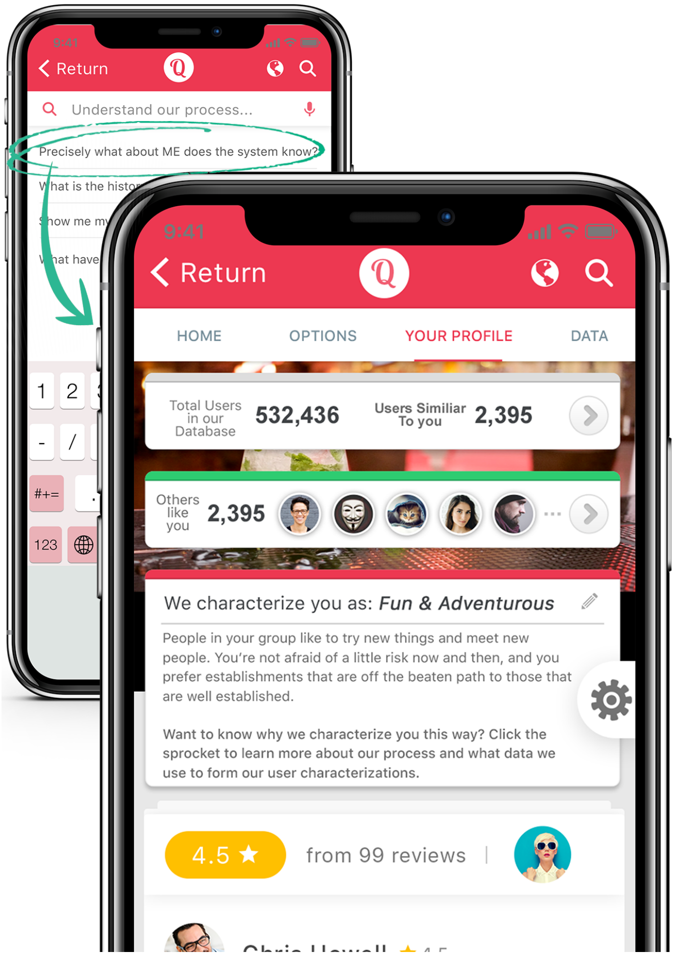

Figure 5. While socially-related information was a contentious topic amongst our factor groups, all participants expressed a desire to understand what data about them is known, and how it is used in deriving recommendations. Some may find this helpful, while others (like the Socially Influenced and Egocentric) find it critical.

a greater or lesser effect on user trust, engagement, and acceptance. For example, an explanation and justification of options (Figure 4) is most important to people like the Interested & Independent group, while users in the Socially Influenced group might respond well to social navigation cues, as shown in Figure 5 and the Cautious & Reluctant group would be more satisfied with a detailed description of the data that fed the model, and appreciate control over which data are used to make recommendations, as demonstrated in Figure 3. Using mixed methods approaches such as Q-methodology can add significant value to traditional user-centered investigations, and offer data that is qualitatively nuanced, while being quantitatively rigorous. An approach such as this could be successfully used early in the design cycle- and indeed we suggest and encourage the early involvement of users, but would also be appropriate in later stages as well. Designers can readily make use of data such as these to help resolve potential conflicts in priorities, and guide their communication strategies.

VI. Conclusion

We have explored potential design features for enhancing the intelligibility and transparency of intelligent systems to end users using a novel mixed methods approach. We have described a variety of reasoning and information seeking strategies of potential users of said systems, and have detailed them into a robust user typology. We have explored this typology in detail, and have compared and contrasted user preferences related to understanding system functions and behaviors to demonstrate how they can be used to guide design strategies. We have compared and contrasted potential design features in order to determine which may be more efficient and valuable to end users in the context of interactions with intelligent recommender systems. Our findings support and reinforce the argument that system transparency is a multi-dimensional construct requiring at least some consideration for user preference and individual differences in order to achieve the desired effect of improving trust, usability, and technology acceptance.

References

[1] 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.
[2] L. Sherry, M. Feary, P. Polson, and E. Palmer, "What's it doing now?: Taking the covers off autopilot behavior," presented at the 11th International Symposium on Aviation Psychology, 2001.
[3] 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.
[4] K. Swearingen and R. Sinha, "Beyond algorithms: An HCI perspective on recommender systems," ACM SIGIR 2001 Workshop on Recommenders Systems, 2001.
[5] B. Y. Lim and A. K. Dey, "Toolkit to support intelligibility in context-aware applications," presented at the the 12th ACM international conference, New York, New York, USA, 2010, pp. 13-11.
[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?," presented at the the 22nd ACM SIGKDD International Conference, New York, New York, USA, 2016, pp. 1135-1144.
[7] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not: explanations improve the intelligibility of context-aware intelligent systems," presented at the ACM CHI Conference on Human Factors in Computing Systems, New York, New York, USA, 2009, pp. 2119-2128.
[8] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," AirXiv, 2017.
[9] A. Glass, D. L. McGuinness, and M. Wolverton, "Toward establishing trust in adaptive agents," presented at the 13th International Conference on Intelligent User Interfaces, New York, New York, USA, 2008, p. 227.
[10] S. Ososky, T. Sanders, F. Jentsch, P. A. Hancock, and J. Y. C. Chen, "Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems," presented at the SPIE Defense + Security, 2014, vol. 9084.
[11] J. Krause, A. Perer, and K. Ng, "Interacting with Predictions," presented at the the 2016 CHI Conference, New York, New York, USA, 2016, pp. 5686-5697.
[12] A. Cooper, The inmates are running the asylum:Why high-tech products drive us crazy and how to restore the sanity. 2004.
[13] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," AirXiv, pp. 1-57, Jun. 2017.
[14] B. Buchannan and E. Shortliffe, Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, MA: Addison Wesley, 1984.
[15] W. R. Swartout and J. D. Moore, "Explanation in Second Generation Expert Systems," in Second Generation Expert Systems, no. 24, Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, pp. 543-585.

[16] E. S. Vorm and A. D. Miller, ”Assessing the Value of Transparency in Recommender Systems: An End-User Perspective,” presented at the RecSys 2018 Proceedings of the 12th ACM conference on Recommender Systems, Vancouver, Canada, 2018.
[17] S. R. Brown, ”A primer on Q methodology,” Operant Subjectivity.
[18] 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.
[19] A. Ram, AQUA: Questions that Drive the Explanation Process. Lawrence Erlbaum, 1993.
[20] 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.
[21] 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.
[22] J. Devore, Probability and Statistics for Engineering and the Sciences, Fourth. New York, NY: Brooks/Cole, 1995.
[23] O. Biran, C. C. I.-1. W. O. E. A. XAI, 2017, ”Explanation and justification in machine learning: A survey,” IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI). Melbourne, Australia, 2017
[24] A. Eriksson and N. A. Stanton, ”Takeover Time in Highly Automated Vehicles,” Human Factors, vol. 15, Jan. 2017.
[25] G. S. Thakur, K. Sparks, R. Li, R. N. Stewart, and M. L. Urban, ”Demonstrating PlanetSense,” presented at the the 24th ACM SIGSPATIAL International Conference, New York, New York, USA, 2016, pp. 1-4.
[26] E. Agapie, B. Chinh, L. R. Pina, D. Oviedo, M. C. Welsh, G. Hsieh, and S. Munson, ”Crowdsourcing Exercise Plans Aligned with Expert Guidelines and Everyday Constraints,” presented at the the 2018 CHI Conference, New York, New York, USA, 2018, pp. 1-13.
[27] T.-H. K. Huang, J. C. Chang, and J. P. Bigham, ”Evorus,” presented at the the 2018 CHI Conference, New York, New York, USA, 2018, pp. 1-13.