INTO THE BLACK BOX: DESIGNING FOR TRANSPARENCY IN ARTIFICIAL INTELLIGENCE

Eric Stephen Vorm

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the School of Informatics and Computing,

Indiana University

November 2019

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Andrew Miller, PhD, Chair

_____

Davide Bolchini, PhD

December 13, 2018

_____

Khairi Reda, PhD

_____

Sasha Fedorikhin, PhD

DEDICATION

To those who boarded the bus and stepped through the door, but never returned.

ACKNOWLEDGEMENT

I would like to acknowledge the superb efforts of my advisor, Dr. Andrew Miller, who provided excellent advice, support, and perspective, always in the right amounts at the right times.

I would like to acknowledge the tremendous assistance of Dr. Chris Wickens, whose lifetime of research and scholarship inspired much of this research, and who personally aided me in the formation and execution of much of this dissertation.

I would like to acknowledge Dr. Jeff Morrison at the Office of Naval Research. His insights and lessons learned from decades of research in the field of human decision making enabled me to avoid many traps and pitfalls, and whose generous support enabled me to conduct much of this research.

I would like to acknowledge Dr. Michael Lillienthal, who stepped up and volunteered himself to be my personal mentor, and demonstrated to me what a leader and mentor really is and does.

I would like to acknowledge Dr. Dylan Schmorrow, who, against his will, agreed to select me into the Naval Aerospace Experimental Psychology program, and who never neglects to remind me of this fact.

Lastly, I would like to acknowledge my lovely wife, Jennifer, whose steadfast support truly enabled me to conquer the task of completing my PhD in three short years, and who has endured countless moves, adjustments, and deployments with profound grace, poise, and patience. Thanks, sweetie. I couldn't have done it without you.

Eric Stephen Vorm

INTO THE BLACK BOX: DESIGNING FOR TRANSPARENCY IN ARTIFICIAL INTELLIGENCE

The rapid infusion of artificial intelligence into everyday technologies means that consumers are likely to interact with intelligent systems that provide suggestions and recommendations on a daily basis in the very near future. While these technologies promise much, current issues in low transparency create high potential to confuse end-users, limiting the market viability of these technologies.

While efforts are underway to make machine learning models more transparent, HCI currently lacks an understanding of how these model-generated explanations should best translate into the practicalities of system design. To address this gap, my research took a pragmatic approach to improving system transparency for end-users.

Through a series of three studies, I investigated the need and value of transparency to end-users, and explored methods to improve system designs to accomplish greater transparency in intelligent systems offering recommendations.

My research resulted in a summarized taxonomy that outlines a variety of motivations for why users ask questions of intelligent systems; useful for considering the type and category of information users might appreciate when interacting with AI-based recommendations. I also developed a categorization of explanation types, known as explanation vectors, that is organized into groups that correspond to user knowledge goals. Explanation vectors provide system designers options for delivering explanations of system processes beyond those of basic explainability. I developed a detailed user typology, which is a four-factor categorization of the predominant attitudes and opinion

schemes of everyday users interacting with AI-based recommendations; useful to understand the range of user sentiment towards AI-based recommender features, and possibly useful for tailoring interface design by user type. Lastly, I developed and tested an evaluation method known as the System Transparency Evaluation Method (STEv), which allows for real-world systems and prototypes to be evaluated and improved through a low-cost query method.

Results from this dissertation offer concrete direction to interaction designers as to how these results might manifest in the design of interfaces that are more transparent to end users. These studies provide a framework and methodology that is complementary to existing HCI evaluation methods, and lay the groundwork upon which other research into improving system transparency might build.

Andrew Miller, PhD Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

Since the earliest examples of machine automation were first introduced, humans have had a love-hate relationship with smart machines. While the promises of increased efficiency and lowered workload have encouraged their growth in many industries, the "ironies of automation" have resulted in a shift from skilled manual laborer to unskilled automation supervisor, and this road from technology to teammates has not always been smooth (Bainbridge, 1983).

One special breed of automation that has historically featured conflicts between humans and systems are those that offer computer-generated recommendations. Systems that proactively offer recommendations are considered very high on a scale of autonomy, in part, because these recommendations are uncommanded (i.e., they do not require a user to prompt them). These recommendations can be helpful at improving efficiency in tasks that require detailed analysis and decision making by users. Because computers can process far greater amounts of data much quicker and more accurately than humans, they can quickly assess a number of options and present a user with a consolidated recommendation for what to do next. This greatly decreases the mental workload of a user, thus speeding up the process. Efficiency, however, sometimes comes at a cost.

When computer-generated recommendations are aligned with a user's expectations, they appear appropriate and helpful, and users subsequently feel comfortable accepting and acting on them. When recommendations do not align with user expectations, however, they can cause users to experience confusion, startle or surprise, which can lead to a range of incorrect or inappropriate interventions, from a minor delay in decision making, to a major miscalculation (Sarter & Woods, 1995).

Because the consequences of inappropriate reactions in many domains involving computer-generated recommendations are often harsh, recommender features have historically only been in high-risk expert domains such as aviation or nuclear process control. Recently, however, computer-generated recommendations have begun to appear in everyday technologies as well. Context-aware systems and recommender systems are common in today's digital marketplace. E-commerce sites like Amazon, and digital steaming sites like Spotify all feature recommender systems that track user behaviors and predict their likes and interests. The most successful recommender systems make use of a technique known as social information filtering or automated collaborative filtering (Herlocker, Konstan, & Riedl, 2000), which are algorithms that add an additional layer of prediction based on the opinions of users (i.e., user ratings) or by using some implicit measures (i.e., number of plays of a song, or number of times a song is skipped, etc). These types of services can be highly effective and very popular. For example, as of 2018, it is estimated that 75-80% of all that is consumed on Netflix results from recommendations (Amat, Chandrashekar, Jebara, & Basilico, 2018).

Today, computer-generated recommendations are beginning to make their way into more and more technologies, thanks largely to advances in artificial intelligence.

The Rise of AI

Since approximately 2010, artificial intelligence (AI) has rapidly grown in popularity as the result of two broad and unforeseen evolutionary developments: powerful graphical processing units (GPUs), and the availability of large labeled datasets. These two synergistically combined to create the "big data" paradigm, which has opened the door for a new generation of recommender systems. Like their earlier counterparts, these systems offer improved efficiency by processing large amounts of information and condensing it to a single "best guess" recommendation. Unlike their earlier counterparts, however, the true benefit of leveraging artificial intelligence comes from its ability to learn from data, thus eliminating the time consuming and challenging task of knowledge engineering and building predictive models by hand (Klein, 1994).

For example, computer vision algorithms today can help recommend decisions about what crops to plant and how best to rotate them to maximize yield and preserve soil quality (Jamuna & Karpagavalli, 2010; Snyder, 2018). Similar systems are being used by police departments to automatically identify criminal suspects using video feeds from traffic and other public closed-circuit cameras (Liptak, 2018). Lawyers who once had to manually sort through hundreds of thousands of court proceedings and related documents can today use artificial intelligence to automatically retrieve, tag, categorize and prioritize information in order to build their cases (Markoff, 2011). Doctors can identify and treat diseases better with the assistance of AI-based recommendations, which have themselves outperformed experts at identifying and diagnosing cancerous skin growths (Esteva et al., 2017), brain tumors (Yan, 2018), and breast cancer (Fingas, 2018). Figure 1 below

highlights recent AI accomplishments with significant scientific merit that have resulted in significant public attention.

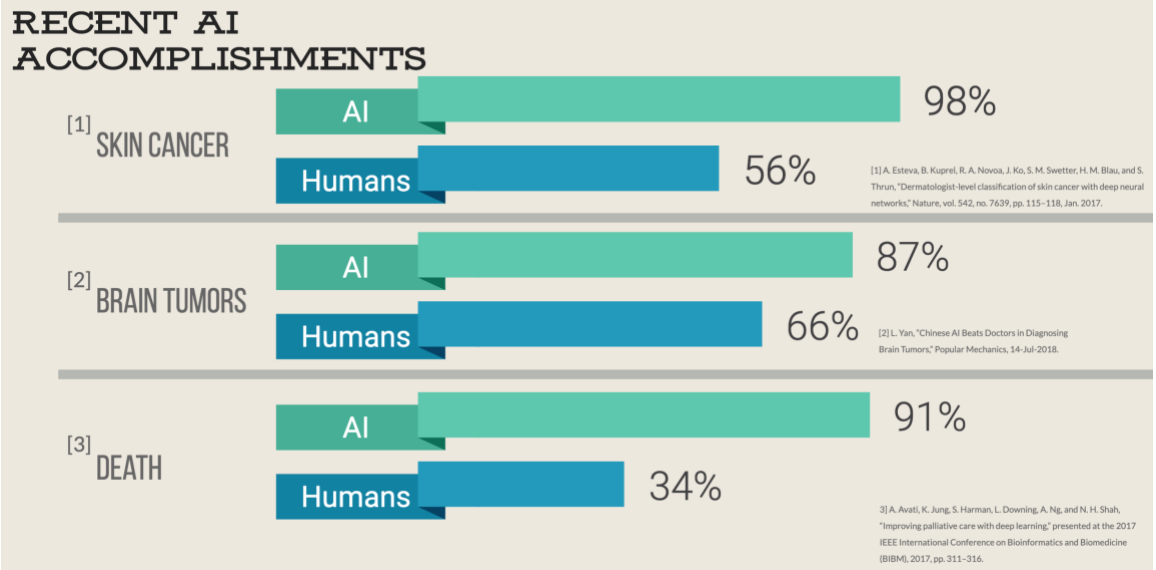

Figure 1: Recent AI accomplishments. AI and machine learning have made significant advancements in a number of fields in recent years. Perhaps the most notable and remarkable achievements have been in the field of physical medicine.

The impact of AI-based recommendations has even begun to affect how scientists determine what studies to run and what methods to use in their experiments (Waddell, 2018). Carnegie Mellon University, the world's leading program in artificial intelligence, has reported that it will be offering a master's degree in "Automated Science" starting in the Fall 2018 (Spice, 2018).

It is likely that modern businesses today view AI much as they viewed websites when the internet was first introduced to the mainstream: those who do not have it are sure to be left behind. And with the list of impressive AI accomplishments continuing to grow, businesses in the near future will almost certainly seek to leverage AI and its predictive capabilities more and more, meaning that AI-based recommendations are likely to become even more common in the very near future.

How AI Spells Problems for HCI

As mentioned earlier, computer-generated recommendations have been extensively studied in the systems engineering and human factors domains because of their potential to influence human perception and decision making, and the associated risks involved in those domains (Barg-Walkow & Rogers, 2016). Conflicts between humans and computer-generated recommendations, for instance, have led to numerous safety mishaps in the field of aviation. Investigations into the causes of these mishaps have led to many formal rules and standards that dictate how visual graphics and interface layouts must be designed in these applications. Virtually every aspect of the design of systems that offer recommendations in aviation, such as airborne conflict resolution aids (Trapsilawati, Wickens, Chen, & Xu, 2017), is strictly regulated to ensure that these designs are standardized and safe (FAA, 1980).

Unlike their industrial cousins, like autopilot systems that recommend emergency maneuvers, consumer-level recommender systems exist in low-risk consumer domains like e-commerce and media streaming, and so interface designs are far less regulated. Determining how much transparency to provide users about how recommendations are made is mostly made by practical considerations such as screen space, clutter, and perhaps user attitudes. Because there is not a large public demand to understand why a Lady Gaga song was recommended after a song by Cher, for example, information about how recommendations are made is usually left out of an interface's design; at best, it may be found in a separate section, such as frequently asked questions (FAQ).

Because of revolutions in computing power and predictive accuracy brought about by the rise of AI, there is a greater push to bring to bear systems that can not only

make accurate predictions, but that can also make recommendations for users. These recommendations are particularly valuable in domains whose decision space is inherently complex, and therefore time consuming for users. For example, a mutual fund manager may gain a competitive edge by speeding up decisions on which stocks to invest in. Similarly, an emergency room physician may be able to provide life-saving medicines to a patient suffering a low-grade stroke. In both examples, AI can examine hundreds of thousands of options, and arrive at an actionable conclusion in a fraction of the time it would take a human being. AI-based recommender systems represent a paradigm shift in the fields of decision support, and industries are currently moving to incorporate such systems in virtually every corner of the marketplace.

As recommender features shift from media streaming and e-commerce to areas with greater inherent risk, such as medicine, the importance of providing transparency to the end user of how these systems function, including how user data is collected, used, and shared, will become a greater concern. Decisions about where to invest money, what jobs to apply for, and where to live all carry serious consequences, and users will almost certainly demand to know more about the inner workings of AI-based recommendations before being willing to act on them.

The need for transparency in AI-based recommendations is also vital in order to detect and prevent algorithmic bias and unfair practices. Financial institutions using machine learning, for example, may predict who will default on a loan by analyzing not only a person's online behavior and spending habits, but also perhaps where they spend their free time by analyzing location data collected from their mobile devices. Human-resource departments using machine learning might predict employee performance and

attrition by analyzing not just resumes, but online profiles from social media in order to justify hiring decisions, raises, and promotions. Life insurance companies may predict the likelihood of premature death by combing through shopping and exercise data shared from mobile apps and social media, which could be used to justify a person's rates, coverage, and deductibles. Automobile insurance rates could be based on driver data, collected from smart sensors that track speed and reckless driving. Hospitals might predict when a patient might be discharged, or the prognosis of cancer, which could affect treatment decisions. Knowing not only the sources of data, but the way in which algorithms assign weights and prioritize data is critical to protecting against algorithmic bias and ensuring that decisions made on AI-based recommendations are fair.

Much research related to transparency in HCI has focused on low-risk contexts such as office automation (Cheverst et al., 2005) or music recommender systems (Herlocker et al., 2000), but little has been done to consider the unanticipated complications of AI-based recommendations in domains with greater risk, such as those examples mentioned above. Research related to transparency in the greater AI and computer science domains has largely focused on methods to make models more interpretable and explainable, but there are few studies that examine how improving transparency translates into the practicalities of system design. Because a central purpose of user-centered design research is to inform designers about the usability consequences of their design decisions (Dix, Finlay, Abowd, & Beale, 2004), HCI needs to consider the importance of providing transparency in AI-based recommendations in light of the rising risk of predictions and recommendations that are soon to be available in everyday technologies.

Contributions and Organization of this Dissertation

This dissertation aims to make progress toward understanding the importance, relevance, and limitations of providing transparency in AI-based recommendations to end-users. It is structured in the following manner: In Chapter two, I present a synthesis of the current knowledge on the use of predictions and recommendations in intelligent systems, and discuss design strategies for helping users understand these features. In Chapter three I introduce relevant background theories of human decision making, which serve as the theoretical foundation for my empirical studies. In Chapter four, I present the results of a user-centered design workshop focused on understanding what information users want when interacting with AI-based recommendations. From this workshop, I developed a taxonomy of user knowledge goals, which attempts to map the motivations behind why users ask intelligent systems questions when interacting with AI-based recommendations. This research also led to the development of a framework of explanation categories, which I call explanation vectors (Evs). Evs are different categories of information that can be used to provide transparency of different aspects of intelligent systems, and when aligned with appropriate knowledge goals, can be highly effective at improving user understanding. In Chapter five, I present findings from a follow-on study that sought to understand how users prioritize explanation vectors differently, which led to the development of a detailed user typology. This typology categorizes and describes the ways people demand information about AI-based recommendations to help them understand and ultimately trust their outputs. In Chapter six, I leverage the taxonomy of knowledge goals developed from Chapter four, and the theoretical background from Chapters two and three to develop a method of evaluating

and measuring the transparency of an interface featuring AI-based recommendations. This method, known as the System Transparency Evaluation (STEv) method, seeks to measure and improve the qualities of an interface associated with helping users resolve conflicts in perception and understanding in order to facilitate better decision making. Findings from a pilot study involving real-world intelligent systems are presented, and refinements to the STEv are suggested, as well as recommendations for further testing and development. In Chapter seven, the contributions and findings from these studies are summarized, and future research is discussed.

Unifying themes from this work are:

- The merging of AI and predictions/recommendations with everyday technologies will create new challenges for HCI User Experience (UX) and interaction design, requiring formative guidance to inform interface designs that support explanation-based reasoning and decision making.
- Recommendations require explanations in order to be actionable. Recommendations and explanations are social signals. Explaining is a social process, influenced by the context of use and individual differences.
- Explanations are prompted by conflicts between expected and observed outcomes, or mismatches in user mental models with a designer's conceptual model.
- Users seek information to resolve conflicts and refine mental models by asking questions. Questions are motivated by knowledge goals.
- Computer explanations should be formatted as social signals, and be aligned with knowledge goals in order to meet user expectations.

CHAPTER TWO: BACKGROUND AND MOTIVATION

In this chapter, I present a synthesis of the current knowledge on the use of predictions and recommendations in intelligent systems, review research from several fields of inquiry on the psychology of explanations and their role in human reasoning and decision making, and discuss the challenges and design implications in providing explanations of AI-based recommendations. Many volumes have been written about these topics in far greater detail than I can present in this dissertation. Wherever possible, however, I summarize these theories in order to expose their relevance towards improving system transparency in AI-based recommendations.

What Are Intelligent Systems?

An intelligent system is any system that can represent data, reason about it by examining patterns and relationships, and interpret that data to arrive at a desired output (Moore & Swartout, 1988). Intelligent systems are known by different names (e.g., recommender systems, collaborative filtering systems, ad placement systems, expert systems, context-aware systems, knowledge-based systems, and clinical decision support systems, to name a few). The term "intelligent systems" is often associated with work domains, such as business planning, but because of its broad definition, it can include virtually any system today.

A feature that is commonly distinctive of intelligent systems is prediction. A prediction is a probability of an action or outcome, typically derived from structural equations whose variables have been shown to have associative characteristics (Devore, 1995). Predictions in intelligent systems are used to inform a user about the likelihood of an event. A simple example of this is a mileage estimate shown on the dashboard of a

vehicle. This estimate predicts, based on previous fuel consumption and environmental conditions, when the vehicle will run out of fuel (SAE, 2016). While useful, perhaps the most powerful way that predictions can be manifested in intelligent systems is by using them to form recommendations.

Recommendations are a special kind of prediction. They are used to transform a prediction into something that is actionable. Recommendations in intelligent systems are used for a variety of purposes. Some types of recommendations are designed to predict user preferences. Netflix, for example, recommends shows to watch, which are predictions based, in part, on previous viewing history. Facebook's news feed is also a form of recommendation. The content shown to users is predicted to be what they are most interested in, based on a variety of algorithms (Eslami et al., 2015).

Recommendations can also be useful in speeding up human decision making. By assuming much of effort involved in assessing a situation and evaluating options, intelligent systems can output a recommended decision based on a high probability of success. These uses are designed to improve decision efficiency and mitigate possible errors, and are featured in domains that typically involve difficult decisions, such as medicine and strategic planning (Ricci, Rokach, & Shapira, 2015).

Recommendations & Explanations: A Give and Take Relationship

While recommendations offer enhanced capabilities to many systems, they often come with extra baggage. Because recommendations are based on an underlying assumption (a prediction), they are associative, in that they are derived from some other object. This means that in order to assess the quality or validity of some recommendation, its underlying associated information must also be provided.

Consider one person suggesting a restaurant to their friend. In doing so, they are predicting their friend will like the restaurant. That prediction is likely based on some knowledge about the friend, or may be an assumption based on how much the recommender enjoys the restaurant themselves. Either way, in order for the friend to act on the recommendation, unless done on blind faith, the friend will likely require additional information— a reason why, or some evidence that the prediction is likely to be accurate, etc. This additional information is called an explanation.

Recommendations and explanations go hand in hand, principally because in order for recommendations to be acted upon, an explanation, in some form, is almost always required. This reflects a characteristic that is axiomatic of successful recommendations: recommendations are transactional processes. In order to make a recommendation, knowledge of the recommended (i.e., the receiver) must be known, and in order for that recommendation to be actionable, knowledge of the recommender (i.e., the sender) must be known. Knowledge between both parties must be shared in order for a successful transaction to take place.

The extent to which this is true in humans is mostly governed by the complex social structures of human communication. Studies in cross-cultural social psychology suggest the recommendation-explanation relationship is strongly correlated across cultures around the world (Berry, Segall, & Kagitcibasi, 1997), suggesting it is likely a function of the underlying cognitive structures and reasoning strategies associated with human comprehension, problem solving, and decision making (Fiske & Taylor, 1991). While these findings are stable in people-to-people interactions, they are not so stable between humans and technology.

The Black Box Problem

Computer systems that do not provide an explanation of why or how a prediction or recommendation was made are said to be examples of 'black boxes.' Information goes in one side, and something comes out the other side, but the process from A to B is obscured.

There are a number of reasons why some computer systems do not provide explanations of their processes. The simplest reason is in order to declutter and streamline user interfaces. For example, design in domains such as aviation have historically focused on techniques that condense and synthesize multiple sources of data, and present them in ways that can convey meaning without taking up much screen real estate (K. Monk, Shively, Fern, & Rorie, 2015). Since space is limited in a cockpit, choosing what to show users is a tradeoff between having a functional display and having a well-informed user.

This is not always a bad thing, however. Cutting out the process and showing only the output in some cases can also improve usability. For example, internet search engines run multiple computations to determine not only what content to return, but also to establish its ranking. These computations are intentionally hidden to the user so they are not encumbered with extraneous information about how those search results were gathered, collated, sorted, and ranked. As a result, the user only sees the results, which to them appear near instantaneously.

Another reason that some processes such as algorithms may be kept hidden from users is to protect trade secrets. Many algorithms are proprietary elements that give companies a competitive advantage. Making them public could mean the loss of that advantage. Protecting algorithms is also done as a means of increasing security as well.

Sophisticated hackers often leverage the power of machine learning to detect and exploit weaknesses in systems, a tactic known as 'adversarial learning' (Chalasani, Jha, Sadagopan, & Wu, 2018). System developers and software engineers, therefore, are often forced to protect the information underlying system processes in order to protect both market interest and security.

Aside from these cases where black box models are somewhat justified, the importance of providing users with an explanation of how systems function, especially in relation to recommendations, is an important component of "good design," the lack of which can have a range of consequences, depending on the context of use.

System Transparency Lessons Learned from Industry

Early intelligent systems were developed for expert users in domains such as medicine and information systems (Buchannan & Shortliffe, 1984). The primary focus of these systems was to serve as adjuncts to the expert by augmenting, or in some cases replacing human decision makers for certain tasks. Very soon after their initial development, users requested that system processes such as recommendations be provided with an explanation. It seemed that potential users of these systems would not consider relying on them unless those systems could conform to patterns of communication commonly used within existing work structure. In other words, the question of "why" was just as important to these expert users as the "what."

As an example, a usability study investigating early medical decision support systems asked doctors to rank 15 desired capabilities and characteristics of the system. Somewhat surprisingly, participants ranked "never make an incorrect diagnosis" second to last (14 out of 15), while they ranked "explain their… decisions to physician users" as

the number one most important characteristic (Moore & Swartout, 1988). While these systems were designed to assist physicians in making diagnoses, users were far more interested in understanding the "hows" and "whys" of computer-aided diagnoses, without which, they could not trust the system, and hence rejected it outright.

Early attempts to provide explanations of system behaviors were nothing more than system rules expressed to the user, which because they were derived from computer language, bore little resemblance to human language at all. These rules also did little to affect the confidence or trust of users, who wanted not only to understand the causal relationship of input to output, but also wanted some form of justification for why recommendations were made (Moore & Swartout, 1988). This portended future challenges for expert and knowledge-based systems, and these challenges significantly limited the use of these systems in many domains for the next several decades (Hoffman, 2017) as designers struggled to program computers to communicate in ways that humans desired most.

Decades later, technology adoption is still hindered by the importance of designing systems that provide explanations to their users. For example, Sedasys, an automated anesthesia machine, was first introduced to a limited market in early 2013 (Frankel, 2015). It was able to manage the anesthesia-related aspects of routine surgeries like colonoscopies, with no human intervention. Surgeries with the Sedasys cost $150 to $200 for each procedure, compared to $2,000+ for a surgery with a human anesthesiologist. Only a year after it was introduced to the market, however, the technology came under intense objections surrounding the machine's lack of explanatory functions. These objections came not from patients, however, but from anesthesiologists

and nurse anesthetists who argued that such system explanations would be crucial to mitigating any unexpected event, and without such functions, they were ethically bound to refuse to adopt such a technology for patient care, despite its apparent benefits (Simonite, 2016). The single nail in the coffin for automated anesthesia was a letter of objection written by the American Society of Anesthesiologists, protesting the use of such a machine. Phillips, the maker of the Sedasys machine, recognized that its market for such a device had dried up, and so rather than work to improve the system's design, decided instead to abandon the project altogether.

Hard lessons such as these reinforce the importance of how good design can promote useful technologies, while poor design can hinder it. In both the early examples of intelligent systems, as well as the Sedasys example above, systems did poorly at providing explanations of their functions, and suffered as a result. This illustrated the principal challenge of "black box" systems, and gave rise to a dimension of usability research known as "transparency," which generally aims at improving users' understanding of system processes.

The metaphor of improving the transparency of the black box is useful inasmuch as it generally describes a desirable quality of a design (providing explanations of system functions), and its basic aim (in order to help users understand). Its usefulness is limited, however, by a range of complications in how the concept of transparency in computer systems is defined and operationalized. These factors complicate not only the act of defining transparency, but also complicate the process of trying to achieve it in design.

Transparency Is Not So Transparent

The origin of the Latin word transparency is dichotomous, consisting of the root trāns, which means “across” or “through,” and the branch pāreō, which means “to be seen.” Webster’s dictionary (Merriam-Webster, 2018) defines the term Transparent as:

Dictionary Definition of Transparent

trans•par•ent | \tran(t)s-ˈper-ənt \

1. a. Having the property of transmitting light without appreciable scattering so that bodies lying beyond are seen clearly: PELLUCID
   b. Allowing the passage of a specified form of radiation (such as X-rays or ultraviolet light)
   c. Fine or sheer enough to be seen through: DIAPHANOUS

2. a. Free from pretense or deceit: FRANK
   b. Easily detected or seen through: OBVIOUS
   c. Readily understood
   d. Characterized by visibility or accessibility of information especially concerning business practices

The most appropriate components of the dictionary definition that relate to non-physical entities are (1) readily understood, (2) visible, and (3) accessible. These three core components are ubiquitous in writings on transparency across a wide range of domains, including law (Walker, 2010), government (Muñoz & Bolivar, 2015), financial regulation (L. Zhang, Zhang, & Hao, 2018), software engineering (do Prado Leite & Cappelli, 2010), business practices (L. Zhang et al., 2018), and of course, artificial intelligence (Miller, 2017).

The word transparency, in any universal sense, however, cannot be described as a single absolute property. This is because transparency is defined by context. Just as a recommendation implies an underlying prediction, the term transparency implies something underlying that must be seen or understood. For example, in the context of

government, transparency often refers to greater degrees of openness and honesty as it relates to the process of lawmaking (Muñoz & Bolivar, 2015). In the context of financial regulation, transparency may refer to the degree to which regulators proactively disclose information, or the ease with which public citizens can request and receive public records (da Cruz, Tavares, Marques, Jorge, & de Sousa, 2015). Transparency in computer science and artificial intelligence frequently refers to how easily programmers can trace a model's reasoning (Owotoki & Mayer-Lindenberg, 2007), but may also refer to other aspects as well, such as how openly systems disclose how user data is collected and shared with other systems (Bernstein, Bakshy, Burke, & Karrer, 2013), or other aspects such as privacy (Murmann, 2018).

This reveals another characteristic of transparency: transparency is multi-dimensional. In the context of business, for example, transparency is not merely concerned with access to information, but also with other important aspects, such as the timeliness, appropriateness, and completeness of that information. In the context of information systems, transparency of computer systems depends in part on its intended users, tasks that users will perform, and both the physical, as well as the organizational and social environments in which the system is intended to be used.

Transparency is a relatively common term in the scientific literature, especially in social and political sciences, but becoming more so in computer sciences as well (do Prado Leite & Cappelli, 2010). Despite this, however, there are surprisingly few explicit definitions of the term, and what definitions do exist vary widely. This most likely reflects the extent to which transparency is operationalized according to certain criteria, such as the nature of the system or task involved. Because of this, no unified definition,

in computer science or elsewhere, has been developed. Nevertheless, it is worthwhile to consider the few formal definitions that have been provided through the computer science-related literature.

Unfortunately, merely using the term transparent, or using transparency as a design goal does not always have the same intended outcome. This is in part because the term itself is defined by the context in which it is being applied. For example, transparency in the HCI sense can refer to a quality of an interface known as observability (Preece, Sharp, & Rogers, 2015). Observability refers to what a user can infer about the current state of the system, typically through some form of display or user interface (Valverde, 2014). Unfortunately, however, this definition does little to aid in the comprehension of transparency in interaction design, because the term can be used to mean both to hide or to make something invisible to the user, and also to reveal and make something apparent.

HCI design guidelines sometimes use the term "transparency" to refer to processes that are kept hidden or happen behind the scenes, beyond user perception. This use of the term transparency is particularly common in domains such as web and interface design (Monk, 1985). For instance, in describing how a search engine returns search results, one could say that the search engine aggregates and rank orders potential results according to a number of variables in the search algorithm, but those variables are transparent to the user, i.e., the user only sees the ordered results, not the algorithmic steps that determined those results.

Other areas of design, such as Don Norman's classic 7 design principles, on the other hand, highlight the importance of making system processes visible to users as a

means of improving usability and user experience (Norman, 1988). Neilsen's 10 heuristics, another classic reference to improving system usability through design, also cites the importance of communicating system processes to users so that they remained informed of what is going on "under the hood" (Nielsen, 1994). Indeed, much research on transparency in the computer sciences is generally referring to the process of making visible something that was previously invisible to the user, and for good reason.

Commonly Used Definitions of Transparency

HCI textbooks refer to transparency as providing "the necessary knowledge within the environment… to support the user in building an appropriate mental model of what is going on" (Dix et al., 2004, p. 283), and "easy-to-understand and intuitive ways of interacting with the system" (Preece, 2015, p. 94). Literature from recommender systems refer to transparency as "exposing the reasoning and data behind a recommendation" (Herlocker et al., 2000, p. 241). Literature discussing intelligent agents describe transparency more broadly as "the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process" (Chen, 2014, p. 2). Studies in information systems have defined it as "…explaining to their human users both the knowledge they contain and the reasoning processes they go through" (Gregor, 1999, p. 498).

The intersection of these definitions, outlined in figure 2 below provides a fairly clear set of purposes for transparency. Themes common to a discussion of transparency in computer systems are seeing (as in visibly observing, but also accessing), understanding, predicting, instructing, and explaining.

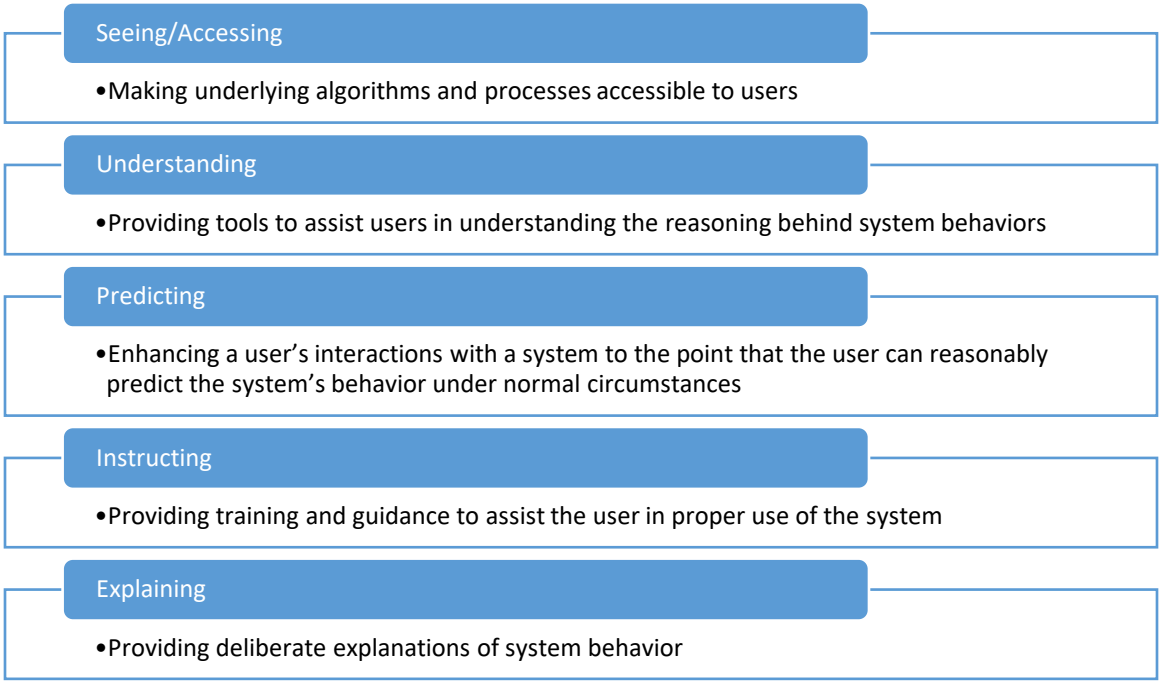

Figure 2: Common purposes of transparency. These are typically found applied in computer science and intelligent systems literature.

We can also understand much about the concept of transparency by observing the object(s) to which the concept often refers. For example, a common HCI design principle emphasizes the importance of making users aware of the current system state, e.g., "…when there is nothing in the state of the system that cannot be inferred from the display." (Dix, 2004, p. 612). Another way that the concept of transparency is discussed is as a function of good design that informs users of what the system can do for them, or making users aware of affordances, e.g., "…when it evokes an easy-to-understand system image in users" (Preece, 2015, p. 94).

Transparency is also concerned with aiding in the predicting of future state, or the consequences of an action, e.g., "…a description of the potential effects that taking a course of action will have on the pre-planned mission" (Pharmer, 2004, p. 3). This is tightly coupled with providing information about a system's intent or goal, e.g.,

transparency is "…the degree to which a system's action, or the intention of an action, is apparent to human operators and/or observers" (Ososky, 2014, p. 1). Discussions of this nature are particularly salient in systems with potentially greater degrees of autonomy, such as intelligent agents and human-robot interaction (Lyons, 2013).

Transparency can also refer to explanations, e.g., "…presenting textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction" (Ribeiro, 2016, p. 1). Often transparency is concerned with the processes involved in generating recommendations, e.g., "…exposing the reasoning and data behind a recommendation" (Herlocker, 2000, p. 241). The above descriptions give some clues as to the focus (i.e., the "what") of the concept of transparency in intelligent systems, which is conceptualized in figure 3 below.

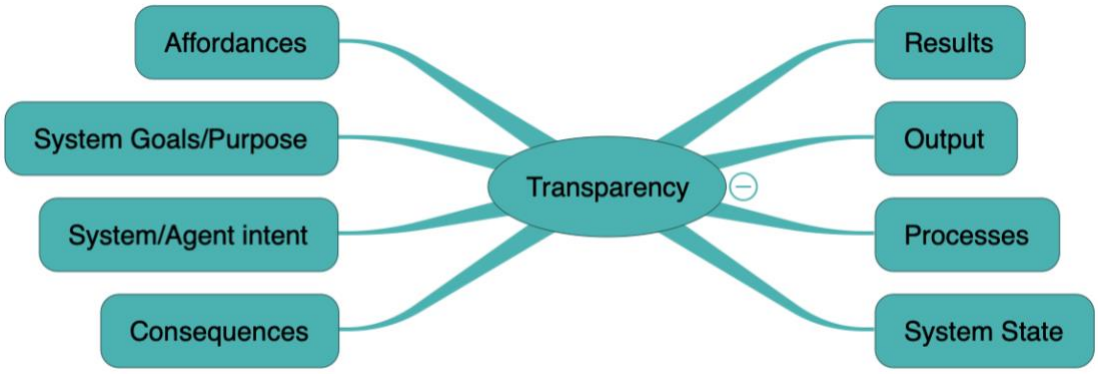

Figure 3: The focus of transparency. These are often associated in computer science literature.

Proposed Integrated Definition of System Transparency

Based on this conceptual model, a unified working definition of system transparency in intelligent systems might be:

> Transparency in intelligent systems refers to a measure of observability, accessibility, ease, and completeness of explanations of system functions and outputs.

As noted earlier, how transparency is defined is determined by a range of contextual associations, which limits the extent to which this definition will be useful. Nevertheless, for the purposes of orientating readers of this dissertation, this working definition is thus proposed.

CHAPTER THREE: THE ROLE OF EXPLANATIONS IN DECISION MAKING

"Explanation generation must be treated as an intrinsic part of any cooperative problem solving. One must explain to cooperate and cooperate to explain." *Patrick Brezillon*

Since the need for transparency in intelligent systems was first recognized, scientists have observed that because users tend to apply the same decision heuristics to their interactions with computers as they do with other people, they often expect or demand that computer systems explain their reasoning in the same way they would another person (Gregor & Benbasat, 1999). As discussed earlier, recommendations are a special kind of prediction, and in order for them to be actionable, they carry a unique transactional requirement- the need for an explanation.

Studies in communication have identified patterns or strategies that people use to explain their recommendations (Gedikli, Jannach, & Ge, 2014), which have yielded taxonomies of explanation strategies. A commonly cited example is that developed by Lombrozo (2012), which outlines three categories of explanation: (1) mechanistic explanations, which explain by describing the parts and processes involved; (2) teleological or functional explanations, which cite a person's or system's functions or goals; and (3) formal explanations, which address the kind or category of the recommender involved.

Less studied, however, is the process through which humans evaluate the goodness of an explanation in order to determine whether or not to act on it. Research suggests that because the act of communicating is a social process, and heavily relies on

the use of metaphors, symbols, and meta-cognition, social cognition theories are likely to be more informative than those from traditional cognitive psychology.

The internal process of evaluating the goodness of a recommendation is likely to start with a person evaluating the similarity between the recommender and themselves, often in terms of tastes, preferences, or motivations. The closer the recommender is to themselves likely plays a large role in helping to determine whether or not to act on a recommendation. This is easier to accomplish between two people than it is between a person and a computer, though research suggests that efforts to embody human characteristics in computers improve user perceptions and willingness to engage [89].

Prior history of recommendations likely play a role in determining whether to act on a recommendation as well. People often consider past interactions as a means of determining future actions (Ajzen, 1996). Therefore, it stands to reason that a recommender who consistently predicts a person's tastes is more likely to be accepted than one with a poor record of accuracy.

Lastly, a person seeking to determine whether a recommendation should be acted upon will likely inquire for more information in order to understand the basis for that recommendation (Sinha & Swearingen, 2002). This information is known as an explanation.

What Makes a Good Explanation?

Explanations are a unique form of communication (Lombrozo & Vasilyeva, 2017) and can come in a variety of forms. Some explanations explain why something exists in terms of its function (i.e., a car has an engine because it drives the car). Others explain things according to their relation to other things (i.e., because it has shelves, it must be a

bookshelf). Philosophers have been discussing and debating the form and function of explanations since antiquity, and have proposed many formal constraints on what constitutes an explanation in terms of its ability to formally explain something. These formal theories of explanation are quite dense, and are often ironically themselves rather unexplainable (or rather, difficult to understand). We will not review these models because we are interested in explanations from a purely functional point of view, that is, we are interested in knowing what makes a "good" explanation so that we can achieve the goals associated with delivering it (e.g., enhanced user understanding, willingness to engage, trust, etc.).

Since the 1980s, it has been understood that explanations generated by computers should be of the same format as explanations between people. Thus, reviewing the criteria for what makes a "good" explanation in human terms will inform much of what and how we create explanations from computers.

Good Explanations Are Satisfying

Technically, what makes something and an explanation is not defined as a property of text, or narrative statements, or other material forms. What makes something an explanation is defined by the interaction of the sender, the characteristics of the sender (their purpose, history, etc.) the message, the receiver, and the receiver's characteristics (goals, knowledge, beliefs, etc.) (Ahn & Bailenson, 1996). The content of an explanation, however, no matter its format, is judged not by how accurate or truthful or logical it is, but rather on whether it has an effect, or rather, whether it has explanatory value.

People ask questions because of a deep, powerful underlying quest to understand. Scientists theorize that all quests for knowledge are an effort to form theories of the

world in order to create structure (Keil, 2006). We call these drives knowledge goals. The quality and goodness of an answer to a question, therefore, could be said to be measured by how well that answer satisfies a person's reason for asking, that is, their knowledge goals. Hence, a good explanation, is one that is satisfying.

Because knowledge goals are contained within a person and are not overtly expressed along with every question, measuring an explanation on a scale of satisfaction is challenging. But while it may seem necessary to demand an objective measure of satisfaction (i.e., a user should be able to answer a question accurately if the explanation had satisfied their knowledge goals), assessing satisfaction on a subjective scale can be just as informative.

Consider again the motivation for asking questions. Each person has their own unique style and pattern of reasoning and knowledge structures, which greatly determine strategies for learning and understanding. Who is better equipped to determine whether the motivation behind a question is satisfied than the person themselves? For the purposes of this research, therefore, we adopt a phenomenological approach to satisfaction, and are concerned with a user's subjective opinion of whether or not an explanation was satisfying more so than an objective measure of whether that explanation has satisfied some formal rules of causality and logic.

Good Explanations Improve Understanding

One characteristic of a "good" explanation is captured in the HCI concept known as Learnability. Learnability is the degree to which features of an interactive system are easily learned and comprehended by novice users, allowing them to rapidly become experts (Dix et al., 2004). This accommodative process is what is behind the formation of

mental models- or cognitive representations of what the system is designed to do, and how it works (Streitz, 1988). Systems that provide explanations not only help the user assimilate knowledge, but also transform the user's knowledge as well. A good explanation, therefore, is one that not only satisfies a user's knowledge goal, but also should enhance their understanding of the situation.

Good Explanations Are Complete

Researchers have repeatedly concluded that people do not need a complete accounting of all causal elements and their associated effects in order to appreciate a sound explanation. Explanations about gravity, for instance, could refer back to events that occurred during the Big Bang, but children readily accept a simple description of gravity as an explanation for why a bouncing ball eventually stops (Miller, 2017). Most people express a preference not for complete logical proofs, but for simple, tractable explanations that fully answer their curiosity.

A characteristic of a "good" explanation that goes hand in hand with satisfaction is therefore completeness. In this case, the completeness of an explanation is not defined by some formal constraints that must satisfy all logical causal elements in order to be considered "complete." Rather, completeness refers to how comprehensive an explanation is in terms of satisfying knowledge goals. Many researchers have illustrated that explanations should allow for contrastive comparison: the "Why not" and the "What-if" of explanations. This is known as counterfactual reasoning, and it plays a significant role in how humans reason about causality. The counterfactual stance is that a particular event (X) would not have occurred if another certain event (Y) had not occurred first.

Examining relationships in this manner allows for casual determinations to be made, rather than merely establishing correlation (Byrne, 2017).

Explaining why a file was omitted from a search query could be accomplished by simply stating a rule, (i.e., it is not a writeable file). A more complete explanation, on the other hand, would allow for a deeper understanding of what caused the file to be omitted (i.e., the type of query chosen only considers writeable file types in its search results). The former explanation may lead a user to try to change the file type, thinking that is what caused the result, while the latter explanation informs the user that they have a choice in query type. Hence, the completeness of an explanation facilitates a causal level of reasoning.

Effects of System Transparency on User Engagement and Trust

Research has demonstrated that users who have a better understanding of how systems work tend to report higher trust in those processes, especially in systems that make recommendations to users (i.e., recommender systems; Herlocker et al., 2000; Kulesza, Burnett, Wong, & Stumpf, 2015; Sinha & Swearingen, 2002). By giving users explanations of system functions, users demonstrate more awareness (Mercado et al., 2016) and are better able to detect anomalies or system errors (Chen, Barnes, Selkowitz, & Stowers, 2016; Trapsilawati, Wickens, Qu, & Chen, 2016; Wright, Chen, Barnes, & Hancock, 2017). Users who are afforded logic rules that aid in their understand of how the system reasons also report higher levels of satisfaction (Cheverst et al., 2005), and making users aware of information underlying algorithms in social media has been shown to increase a greater sense of user control, and higher levels of engagement (Eslami et al., 2015).

As discussed earlier, the decision to make some processes visible to users is not made arbitrarily, and is sometimes driven by a range of competing priorities that range from high level concerns over intellectual property and trade secrets, to low level physical constraints such as available screen space. A practical motivation that often drives the decision to restrict or hide data from end users is that the processes underlying most computer systems are inherently difficult to understand by lay users, and the belief that not many users want to see every detail of each process. The decision of what to show users is often made on this assumption, which is occasionally backed by user testing (Lim & Dey, 2009). This tension between design pragmatics and keeping the user well informed is a source of significant challenge. In cases where providing users access to underlying system processes is more important, however, an additional challenge is how to provide a simple explanation of processes that are inherently complex.

Early generations of intelligent systems were simple rule-based systems that were programmed entirely by hand through a process known as knowledge engineering (Buchannan & Shortliffe, 1984). This had two primary consequences: 1. There were a finite amount of rules that could define a system's behavior and output, and 2. Those rules were entirely known to the programmer because they were developed by a human. If a user wanted to understand the process behind the output of a system, explaining that was fairly straightforward, even though these forms of explanation were unsatisfying and did little to help encourage use, as mentioned before (Ye & Johnson, 1995).

As the complexity of intelligent systems has grown, however, so too has the library of rules and conditions on which their logic operates.

How Users Perceive Algorithms

Because the concept of transparency in HCI most commonly refers to measures of observability in a system, one might suppose that making some things more observable would yield greater transparency. In some cases this holds true. Users need to know whether or not a GPS navigation system is considering time, or distance, or toll roads in their algorithm in order to decide whether or not to accept the recommended route.

Merely making something available to users, however, is unlikely to affect much towards improving their perception of transparency alone. This is because most lay users do not understand algorithms, as discussed earlier. And research suggests that a lack of knowledge about how algorithms work is only part of the problem. Many people are completely unaware of the presence of algorithms underlying system outputs.

It is estimated that 75-80% of all content consumed on Netflix comes from personalized recommendations, originating from algorithms considering not only user's preferences, but other personal data as well, including their device type, IP address, and even GPS location (Amat et al., 2018). The extent to which users are aware of these underlying functions, however, has been demonstrably low.

Studies have shown, for instance, that users of Facebook are often I to the fact that the content they receive (i.e., their news feed) is personalized for them based on algorithms that include data from connected sources outside of Facebook (Eslami et al., 2015). Other studies have found that users of social media and blogging platforms consistently underestimate the size of their audience, in part revealing a lack of appreciation for how search engine optimization algorithms affect distribution of online content (Bernstein et al., 2013; Viégas, 2006). A survey of 6,000 global consumers found

that only 33% thought they used or interacted with AI on a regular basis, while the actual percentage (based on reported demographics collected through the survey) was 77% (Pega, 2018). Similarly, a 2018 report estimates that less than two years since the introduction of smart speakers to the consumer market, nearly a quarter of the US population (43 million) own or have regular interaction with them (Edison, 2018). While these popular devices offer a range of conveniences, few users may realize or understand, however, that the technology behind those speakers is in fact using natural language processing algorithms trained and built using a combination of reinforcement and machine learning.

Awareness and knowledge of algorithms and their role in how system outputs are derived is important because these factors greatly affect user behavior and system interactions. Understanding contexts such as audience (i.e., who is listening or watching) and audience size greatly affects user behavior in online environments, mediating both the content that is shared, as well as how it is shared.

This highlights again the multi-dimensional nature of achieving transparency in intelligent systems. Simply making something available or visible to the user is only half the battle. That information must also be understandable in order for it to have a measurable effect.

Challenges with "Big Data"

"Intuition fails in high dimensions." *Richard E. Bellman*

Machine learning approaches such as deep learning and convolutional neural networks have revolutionized artificial intelligence, and have brought its usefulness and speed into the everyday spotlight. With these advances, however, have come a new

generation of learning models that are distinctly different from their earlier cousins in two ways: 1) these are built on a volume of data whose size and scope vastly exceeds human comprehension, and 2) in cases such as deep learning, these models are no longer built by hand, meaning that the rules they learn are not known to humans, even their programmers. Thus, the term "big data" is often used to refer to both (1) the volume of data, as well as (2) the distinctive technologies that have emerged from it— both of which have troublesome effects on system transparency.

The Paradox of Transparency

The term "big data" does little to express just how large and cumbersome data associated with machine learning applications can be. A single 24x24 pixel image, for example, may generate in excess of a petabyte (1,024 terabytes) of information, depending on the type of learning model used to learn its features (Goodfellow, Bengio, & Courville, 2016). Training data from billions or trillions of records, all from different sources, on the other hand, can easily generate data that are several billion or trillion exabytes (1,024 petabytes).

Generalizing in high dimensions (i.e., millions and billions of data points) is not intuitive, and reasoning at that level challenges human-level abilities to perceive relationships, a phenomenon known as the "curse of dimensionality" (Bellman, 1961). This is because our intuitions inherently come from our understanding of a three-dimensional world, which is dwarfed by the sheer scale of data involved in machine learning. What is significant to a model, therefore, may not translate well to what is significant to a person. Attempting to display data at this scale is completely impossible.

Even techniques such as visual analytics fail at these levels because the data cannot be broken down into sizable chunks (Liu, Wang, Liu, & Zhu, 2017).

Making all of that data and associations visible to the user, therefore, may not improve anything in terms of usability, and instead may lead to confusion, or to incorrect judgements about a model's accuracy or appropriateness (Domingos, 2012). There is also research to suggest that making this information available might actually hurt a system's perceived usability. For example, in many online and mobile contexts, users consistently express preferences for designs that do not involve lengthy paragraphs-long text, or interfaces that require extensive training to use (Gedikli et al., 2014; Glass, McGuinness, & Wolverton, 2008; Herlocker et al., 2000). Simply including more data or chains of algorithmic texts in interface design, therefore, would almost certainly fail to achieve any favorable characteristics of usability.

This raises a dilemma for interaction designers, who must somehow determine the right balance between providing enough information to support informed decision making and satisfy users, but not so much that produces confusion, clutter, or lowers perceptions of accuracy or trust. This dilemma has been termed the "transparency paradox" (Nissenbaum, 2011), and these tradeoffs and their design implications are becoming increasingly important and more complicated by the widespread introduction of AI to devices that permeate every sector of the digital landscape.

The Interpretability Problem

Deep learning has facilitated an impressive record of astonishing outcomes. Deep neural networks can predict, with uncanny accuracy, complicated things like who will develop diabetes, who will be readmitted to a hospital within the next three months, and

even who will be alive at the end of 6 months, and who will die (Goodfellow, Bengio, & Courville, 2016). The process of decoding the human genome took hundreds of dedicated scientists more than a decade. Today, because of deep learning, a person can have their personal genome sequenced in under 24 hours (Sivarajah, Kamal, Irani, & Weerakkody, 2017). These demonstrated successes, however, have come with unexpected consequences.

The principle axiom in deep learning is the belief that truth can be found from data (Theodoridis, 2015). So rather than building models on assumptions, the deep learning approach is to allow the machine to learn everything from the data. For example, rather than programming a model with the laws of aerodynamics into an autopilot application, the deep learning approach assumes that with enough data, a neural network will learn these laws on its own. And this approach is considered acceptable because deep learning is typically used in domains that have ground truth, which means that the process of how outputs are derived is considered less important to whether or not their outputs can be demonstrated to be accurate.

Efforts to explain computer processes to end users, especially those that involve machine learning approaches, are further complicated by the fact that while many of these processes are difficult for most lay users to understand, some are even beyond the comprehension of the programmers themselves. But because the results can be determined accurate, and those results can translate into tremendous gains, a kind of means-ends justification is adopted, resulting in an overall attitude that so long as the results are accurate, trust it and don’t ask too many questions.

Despite these challenges, much work has been done in recent years towards making models more interpretable. Model interpretability in this sense is a focus on methods that describe the inner workings of a system's outputs in ways that are accessible and amenable to human comprehension. Model interpretability is a rapidly growing field of interest in AI-related research, and has produced a large volume of methods. It is beyond the scope of this dissertation to review all approaches within this research space. Instead, in the next section I highlight two promising methods that most approaches fall into, and briefly review their strengths and weaknesses.

A Review of Approaches to Improving Model Interpretability

As discussed earlier, a fundamental problem of AI-based recommendations is the scale and size of their data structures. Thus, the challenge is to find ways to reduce the complexity of all these operations in ways that a) remain faithful to the model, and b) do not reduce or oversimplify the complexity to such an extent that system explanations become proxies for persuasions (Herman, 2017).

Much research has attempted to tackle this problem in recent years. A few of the most common approaches are summarized here.

Proxy Models

A proxy model is a model that behaves similarly to the original model, but in a way that is easier to explain (Gilpin et al., 2018). Proxy models typically begin by focusing on local as opposed to global explanations. A global explanation is one that explains the model's function writ large, whereas a local explanation explains a specific output. By focusing on a single instance, such as the classification of a single image, the

problem space is reduced, though still not to the extent which would be considered manageable by human comprehension standards.

Proxy models attempt to provide an approximation of what features contributed the greatest weight to a single classification in order to understand how they influenced the prediction. They accomplish this by taking one instance (i.e., one image, or one patient's worth of data, etc.) and permuting it, or replicating it with a few modifications. This permuted dataset is then fed back into the classifier, and the output is compared to the original output, along with a measure of similarity between the two datasets. By making small changes to the permuted data and then tracking the effects of those changes on the output, a simple model can be built in order to understand what features played the biggest role in determining the output.

Examples of proxy models can come in many forms. LIME, or local interpretable mode-agnostic explanations, is a popular example, and still considered one of the best proxy model explanation approaches to date (Ribeiro, Singh, & Guestrin, 2016).

Another graphical approach in the proxy model category are decision trees, which can map features or variables and their associated weights in a manner that is intuitive and easy for most users to understand.

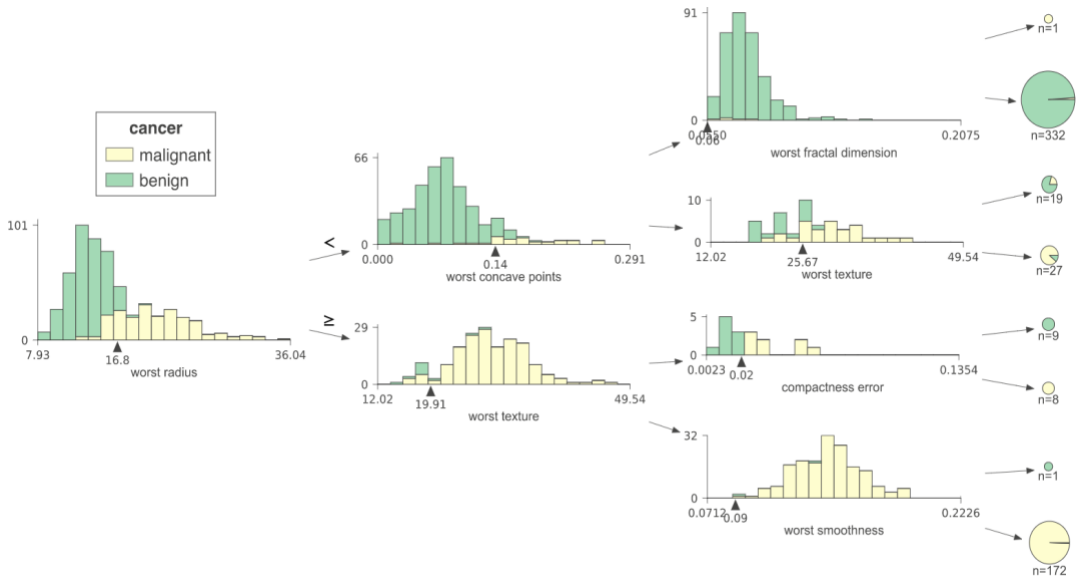

Figure 4: Example of a decision tree model. This figure explains features that contributed to a classifier's output. The above hypothetical example is a decision tree of a breast cancer classifier. Notice how it is possible to intuitively identify what variables played the largest role in the classifier's decision process. Decision trees, hence offer good transparency and visibility, but are limited to only processes with a moderate to low number of variables.

Although approaches such as decision trees, such as the example in figure 4 above, can produce explanations that closely approximate the original network, the generated outputs themselves can be quite large and somewhat difficult to interpret, meaning that they have limits in scalability and applicability to some applications. Despite this, proxy models represent perhaps the current best approach to providing local explanations of model outputs that are accessible to lay users.

Salience Mapping

Salience mapping is the process through which a network is repeatedly tested with various input components occluded in order to create a map that shows which parts of the data have influence on the network output. Various techniques exist within this category (for example, Layer-Wise Relevance Propagation, LRP (Bach et al., 2015); integrated gradients (Sundararajan, Taly, & Yan, 2017), and others). Salience mapping techniques

vary between focusing on activity, or highlighting areas where neurons are most active (and thus contribute the most to a model output), and sensitivity, or highlighting areas where changes would most affect the output.

The number of approaches to making models more interpretable is growing at a rate that challenges researchers to remain cognizant of. The above are two broad and general approaches into which much of the current interpretability research can be classified. These approaches are also amongst those that provide explanations that may benefit everyday users. Other approaches not covered here, while accurate and effective, produce explanations that are mostly only useful to programmers who want to verify model outputs and understand a model's learning in order to refine its processes. Hence, they are not very useful to end users, and do not likely translate well to practical design strategies for interfaces.

Summary of Chapter Three

To summarize this chapter: Recommendations are actionable predictions, but in order to be considered useful and trustworthy, in most cases they require some form of an explanation of their origination. Intelligent systems have suffered from poor usability related to explanation functions, which have limited their widespread acceptance, popularity, and success. Artificial intelligence, more specifically machine learning, is a new approach to building predictive models that is producing remarkably accurate results in a fraction of the time as other methods, and is thus bringing recommendation features to new technologies and market areas like never before. No longer relegated to expert domains, recommender functions are increasingly becoming part of everyday

technologies such as mobile phones and internet of things devices, meaning that everyday users are faced with determining whether or not accept or trust and use them.

Current approaches to providing explanations are complicated by the size of big data, which frequently exceeds human comprehension in terms of scope and complexity, and machine learning models are difficult or in some cases impossible to interpret. While much promising work has been done towards making machine learning models more interpretable, little effort has yet been made to bridge the gap between quantifiable model interpretability, and practical design considerations that produce usable and pleasurable results.

The next chapter will introduce the three empirical studies aimed at bridging this gap. In subsequent sections, I describe a model and framework that may be useful to help guide design decisions for improving system transparency in AI-based recommendations, and propose an evaluation technique for measuring and assessing system transparency in order to improve designs based on user feedback.

Introduction to Empirical Studies

As introduced in this chapter, models that are able to be explained and understood by humans are said to be "intelligible" (Lim & Dey, 2010). A good deal of research has been done towards developing methods to make models more intelligible to end users (Doshi-Velez & Kim, 2017; Lim, Dey, & Avrahami, 2009; Ribeiro et al., 2016). Gregor and Benbasat (1999) presented a detailed review of explanation types, and identified a set of useful constructs used to generate explanations. 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).

Lim and Day (2009) investigated the need for explanations further by examining what information users demand most from context-aware systems. Their findings provided a taxonomy of intelligibility types, which provides a list of different types of information that intelligent systems can offer to users in order to help them better understand system processes and outputs. Follow on studies suggest that users most want information about how the system works (i.e., system transparency) in situations when trust is likely to be low, or in situations with high degrees of uncertainty (Lim & Dey, 2010), and that demand for explanations was elevated most when system behaviors seemed inappropriate or out of place, such as if a user were given an unusual recommendation by the system.

These findings have been used broadly to create interface designs that improve user understanding and trust in intelligent systems. Studies have found that providing users with explanations of system processes tends to improve their performance on human-computer tasks, calibrates their trust (knowing when and when not to trust, depending on the circumstances), and improves usability and satisfaction during interaction (Glass et al., 2008; Herlocker et al., 2000; Krause, Perer, & Ng, 2016; Ososky, Sanders, Jentsch, Hancock, & Chen, 2014).

While these are certainly encouraging findings, however, they may be limited by narrowing the discussion of transparency to a single utilitarian dimension (i.e., how a

system works), rather than investigating a range of dimensions that might be relevant in a discussion of system transparency, such as how users and their data are known to the system, how users are grouped according to their likes or behaviors, and the role of other people's behaviors on user's willingness to engage with and trust system outputs.

Indeed, as introduced in Chapter three, transparency in intelligent systems is a multidimensional construct, and can actually be divided into three principal focus areas in the research domain: 1) Fairness, as in the process of uncovering, detecting, and mitigating algorithmic bias; 2) Accountability, as in the process of uncovering unethical practices such as collecting and selling user data, or illegally tracking users without their consent, etc.; 3) System transparency, as in the uncovering and providing explanations of the inner workings of system processes, especially as it relates to algorithmic decision making and the generation of recommendations to users. This is conceptualized in figure 5 below.

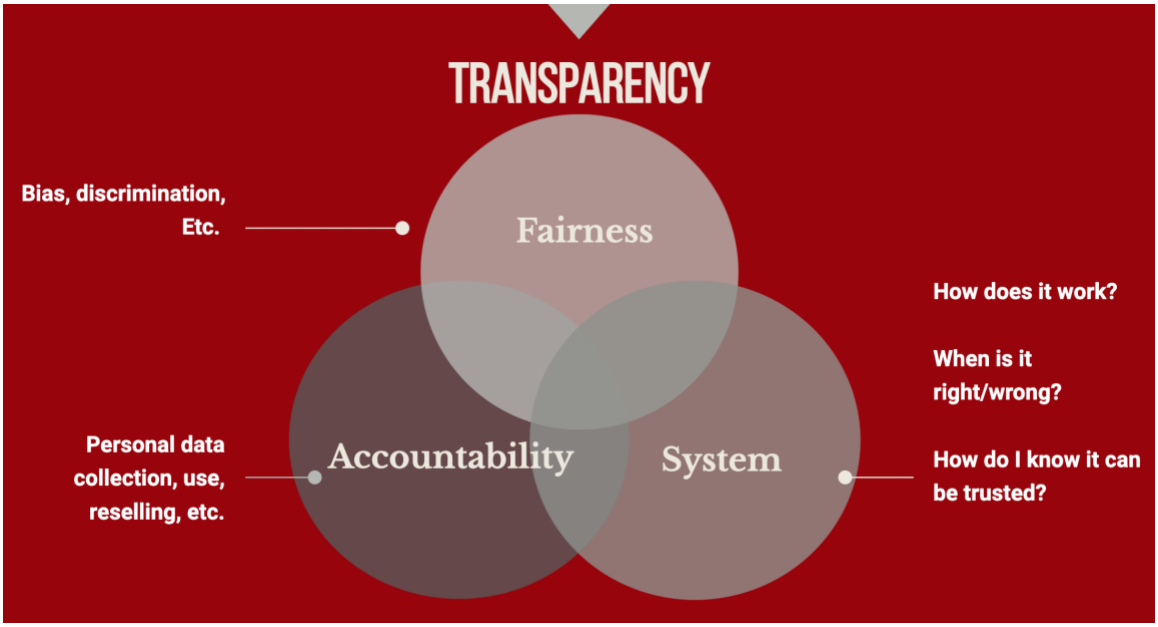

Figure 5: The multiple dimensions of transparency. Transparency's multiple dimensions require deliberate definitions, operationalization, and parsing to disentangle its many facets.

Much work on intelligibility has focused on this third aspect of system transparency. Research in this area has been based on findings that have demonstrated that providing users with explanations about the reasoning of intelligent systems can increase their understanding, trust, and willingness to use the system (i.e., technology acceptance) (Kulesza, Stumpf, Burnett, & Kwan, 2012). As the scope of intelligent systems grows to include artificial intelligence built on machine learning, however, and AI-based predictions and recommendations become more widely available to end users, the range of aspects related to transparency- fairness, accountability, privacy, etc., - grows.

This is partly because many of today's technologies such as mobile phones, wearable GPS-enabled computers, and wirelessly connected devices such as the internet of things are transforming everyday users into walking potential data mines. Users can provide businesses with lucrative details of themselves and their lives on a minute-by-minute basis with little to no awareness. Companies are easily able to capitalize on this opportunity and collect vast amounts of this data by tracking online behavior (i.e., what sites are visited, what products are purchased, what topics are searched), patterns of digital consumption (i.e., what TV shows and movies are streamed, what music is played), and in some cases even patterns of daily life (i.e., using mobile location data), all of which can be fed into machine learning algorithms and used to predict future behaviors such as purchases, and even life events (Allen, 2018).

This greatly expands the range of information that users may want to know from intelligent systems beyond that which has commonly been explored in other intelligibility-related research. This also suggests that designers of the next generation of

intelligent systems should consider the need for explanations that go beyond mere system functions. In order to do this, the range of questions that users may ask when interacting with AI-based recommendations, and their priority relative to contexts such as decision space, and use domain must first be assessed.

This is the motivation of my three empirical studies, 1) to first understand what questions users want answered by intelligent systems (and why); 2) what explanations are more or less useful than others, and under what circumstances might their importance shift; and 3) how these concepts can be combined and used to develop an assessment tool that can evaluate transparency from the perspective of the end user.

To answer (1), I conducted a user-centered workshop, which provided me with detailed information on the variety of questions users consider when interacting with AI-based recommendations. To answer (2), I used findings from the first study and conducted a study using Q-methodology to assess the range of predominant attitudes that affect what explanation types users find more or less important to them. Then, using findings from both of these studies, I developed an evaluation technique to answer (3). Chapters four, five, and six, outline in greater detail the approach, methods used, and results of these studies. The overall relevance and significance of these findings, their potential for application in design research, and areas of future research are discussed in Chapter seven.

CHAPTER FOUR: EVALUATING USER EXPECTATIONS OF SYSTEM TRANSPARENCY

Background and Motivation

Previous studies in intelligibility have revealed that some types of explanations are better than others at enhancing user understanding (Herlocker et al., 2000; Lim et al., 2009). These studies have attempted to map certain explanations to certain circumstances or moderating factors (such as situations involving higher risk, or systems whose outputs have a high degree of uncertainty) in order to target those explanation types as more useful and therefore a higher priority in those situations, thus providing meaningful design guidance.

Considering the context of the situation in which users may demand an explanation from an intelligent system, however, may not be the best way of determining what explanations are most effective or appropriate. Studies in the cognitive sciences have found, for example, that the quality and efficacy of an explanation is determined not by its accuracy or completeness alone, but rather by its relevance to the person asking about it (Lombrozo, 2012). For any explanation to be functional and effective, therefore, it must first address a user's motivation for asking.

One way to conceptualize this is to consider users' motivations for asking questions as knowledge goals, or the goals that fuel their need for an explanation. By observing the types of questions users ask, it may be possible to infer the motivations behind of those questions, and thus approximate a user's knowledge goal.

Considering user knowledge goals has several benefits that can help assess both the need for transparency in AI-based recommendations, as well as expand previous work

towards identifying what explanation types are most effective, and under what circumstances should they be prioritized.

First, considering a user's motivation for asking questions shifts the perspective from what is likely most interesting or important to programmers, to what is likely most interesting and important to end-users. This perspective shift in transparency from backend programmer to end user is not necessarily always a feature in design research, and there have been many examples of designs which fail to accommodate end users' needs for information as a direct result (Silveira, de Souza, & Barbosa, 2001).

Second, as discussed in the role of explanations in decision making section of Chapter two, effective explanations not only provide an answer to a question, but also add to a user's knowledge. This means that the act of providing an explanation is very similar to the act of teaching something. As has been repeatedly confirmed by research, considering the needs of others is an essential component of effective instruction (David, Krivine, & Simmons, 1993; Feltovich & Coulson, 2001; Jonassen & Hernandez-Serrano, 2002).

Introduction to Study One

This study was designed to answer the following research questions:
What questions do users ask when interacting with AI-based recommendations, and what are some motivations behind those questions.

In order to assess the need for transparency in AI-based recommendations and expand previous work towards identifying what explanation types are most effective, understanding and categorizing what questions users have, as well as beginning to understand their motivations is a necessary step in order that interaction designers may be

able to determine what information needs to be explained to end users, and how that information should be prioritized in the design space.

Approach

To accomplish this, I organized a user-centered design workshop to elicit and analyze the types of questions that users ask when interacting with intelligent systems. In this workshop, I qualitatively assessed user information needs in the context of interactions with AI-based recommendations, following closely the model provided by Lim and Dey's previous work in context-aware systems (Lim & Dey, 2009). My main contribution to Lim & Dey's taxonomy of intelligibility types is that rather than a survey-based investigation, I used a user-centered design workshop format, which allowed for more in-depth and detailed reporting of user interactions. Additionally, instead of focusing on existing technologies, I designed this workshop around a central theme of advanced AI-based recommender systems that are currently unavailable to the public. This focus on AI-based recommendations expanded the range of questions that users might ask beyond those related to how and why a system operates, to other elements of increasing importance in intelligent systems that offer AI-based recommendations, such as how users' personal data is collected and used by the system.

Development of Interactive Vignettes Using Design Fiction

To facilitate user interactions with these advanced AI-based recommender systems, I developed a series of interactive vignettes. These vignettes were developed through a combination of inspiration from personal experiences with the DARPA Explainable AI (XAI) program, by analyzing descriptions of AI-based startup ventures advertised on www.crunchbase.com, and by analyzing submissions to the United States

Patent and Trademark Office. The resulting analyses provided an ample basis for speculation of what future technological developments are likely to be enabled by artificial intelligence in the very near future. This approach to design research using speculative fiction to investigate future design needs and interactions is known as design fiction (Grand & Wiedmer, 2010).

Design fiction is an approach to design research that uses visions of the world that could be in order to anticipate and investigate future design needs and interactions (Sterling, 2009). Our vignettes, therefore, described intelligent systems that do not currently exist or are not currently available to the public, but are reasonable inasmuch as their existence can be inferred by current trends in AI growth and industry interest in embracing AI in their business practices.

As mentioned earlier, people usually require explanations in primarily two scenarios: when users perceive an anomaly (when expectations and outcomes do not align), and when a decision must be made (Landman, Groen, van Paassen, Bronkhorst, & Mulder, 2017). The descriptive vignettes were therefore designed to create situations that would create the ideal conditions for users to ask questions. Each vignette described an interaction between a user and an AI-based recommender system, which concluded with a system-generated recommendation that participants had to consider whether or not to act on, or ignore. In each scenario, however, the recommendation intentionally seemed out of place or ambiguous, suggesting that it could be incorrect, but might not necessarily be so. This ambiguity elicits users to ask questions related to transparency, e.g., how was this recommendation made, how can I tell if it is accurate or not, etc.? Therefore, in order for a user to ascertain whether or not the recommendation was made in error, users would

need additional information, which they were told they could obtain by asking the system questions.

We developed these vignettes to be broadly representative of a wide range of use cases: 1) a human resources intelligent agent that predicts success in the workplace, 2) a financial management system that provides recommendations based on machine learning, 3) a fabricated social network that displays ad content based on data learned from user web interaction, 4) a digital clinical assistant that recommends treatments to physicians, and 5) a personal intelligent agent that suggests movie, shopping, and restaurant choices. Each hypothetical system is described in brief detail below.

Human Resources Key Indicators of Talent (HR-KIT)

Human Resources Key Indicators of Talent (HR-KIT) is a human resources system that predicts optimal fit in the workplace using machine learning. HR-KIT parses text provided by candidate forms and resumes in order to evaluate professional backgrounds, level of educational, capability, level of interest, and goodness of fit.

Deep Securities and Accounting Management (D-SAM)

Deep Securities and Accounting Management (D-SAM) is a financial investment system designed to trade mutual funds that are predicted using deep learning. D-SAM

predicts future performance by evaluating hundreds of layers of variables fed from real-world financial data in real time.

Next Generation Social Media

In this vignette, users of a nondescript generic social network service experience an offensive and embarrassing out-of-place ad that plays out loud at their workplace for some unknown reason. The ad is reportedly curated for the user based on their social media and online web browsing history.

Oncological Neural Network Prognosis and Recommendation (ONNPAR)

Oncological Neural Network Prognosis and Recommendation (ONNPAR) is a clinical decision support system that can recommend treatments based on patient data. ONNPAR works on a machine learning platform of convoluted neural networks, and was trained on a large dataset of patient data and outcomes in order to derive its personalized predictions and recommendations.

Q-CONCIERGE (Q-Conc)

Q-Concierge (Q-Conc) is a system that recommends personal experiences like shopping and restaurants based on personal internet browsing history and social media.

The heart of Q-Conc is machine learning that has been trained on data from hundreds of thousands of users of several different personalized concierge systems.

Full interactive vignettes for each of the above systems are available in the appendix.

Methods

All participants (N=24) in this workshop were graduate students in HCI and Bioinformatics at Indiana University. Once introduced to the user-centered design workshop format, participants were given sticky note pads and provided writing utensils. They were then shown a descriptive vignette that was projected on a wall for the participants to read. After each vignette was read, participants were asked to imagine themselves as the main character in the vignette, and to think about what kinds of information they would want to know following each vignette in order to determine whether or not to accept and act on the provided system recommendation (i.e., in a scenario in which an HR system recommended an unexpected candidate for hiring, the participant would need to determine whether or not to go with the system's recommendation, or to discard it).

After each vignette, participants were asked to think of as many questions as they could, and to write them down, each one on a separate sticky note. Participants were encouraged to consider these systems as being capable of answering any question asked of them, and thus to ask the system questions that, if answered, would provide them with information that would affect their decisions and behavior following a computer-generated recommendation. Each question generated by each participant was written on a separate sticky note, which were then stuck on a large wall in no particular order.

Affinity Diagramming

Once all vignettes were presented, and all notes were posted on the wall, participants were asked to read through all sticky notes and collaboratively discuss how the notes might be combined or grouped. If any relationship between questions were identified, participants were encouraged to begin physically grouping notes, labeling the groups by drawing circles around them and naming them with dry-erase markers, in accordance with the affinity diagramming approach (Holtzblatt & Beyer, 2014).

Throughout this process, participants were encouraged to continue to add new questions as they discussed aspects and insights previously overlooked. These questions were then added to functional groups until all questions were physically arranged and labeled on the whiteboard.

Participants were not given instructions as to how to label these groups. This was done intentionally in order to allow for organic ideas to emerge. Accordingly, participants engaged in several versions of groupings and grouping schemes until they reached a consensus.

Semi-Structured Focus Group Exercise

Next, once all sticky notes had been arranged on the wall, participants were asked a series of open-ended questions in the format of a focus group to capture qualitative insights from participant comments. Participants were asked to explain why they asked the questions they asked, and invited to share their motivations for inquiry. Open discussions continued as the group considered potential motivations for individual questions.

Participants were also invited to infer what other people were trying to accomplish by asking their questions. This activity was insightful because it led to the uncovering of multiple potential goals underlying the same question. For example, a person asking "why did I receive this recommendation?" could be asking in order to understand the process behind the recommendation itself, and could also be asking in order to understand why this recommendation was presented to themselves (i.e., how does the recommendation relate to the user and their preferences?).

This focus group activity lasted approximately one hour until it was evident that all questions and insights had been captured. Once the workshop concluded, all questions were transposed from sticky notes, and notes taken from the focus group and think-aloud activities were entered into a spreadsheet for analysis.

A full list of all questions and notes taken during the workshop are available in the appendix.

Results

Results of this activity are broken into two sections. I first report on the analysis of questions users asked, and describe the process through which I developed a taxonomy of user knowledge goals. Following this, I describe a secondary analysis through which I grouped questions into functional categories which I call Explanation Vectors (Evs), and explain how the EV framework can be used to target explanations to meet the knowledge goals of users.

Coding Scheme

The first step in analysis was to analyze all questions recorded from the workshop. To do this, I used an open-coding technique borrowed from grounded theory (Corbin & Strauss, 2008). The coding scheme can be found below in table 1.

Having read through all collected questions, I developed three analytical categories into which I could sort questions: **focus**, or the focus towards which a question seemed to be directed (e.g., the question was focused on understanding how the system output was derived); **time**, determined by whether the question was looking to understand the past, present, or future; and **purpose**, or the apparent reason behind the question (i.e., in order to help the user determine whether or not the system is making an error, or to help the user improve their privacy). I discuss these codes and their descriptions in more detail below.

CODING SCHEME				
FOCUS				
INPUT	OUTPUT	PROCESS	USER	OTHER PEOPLE
TEMPORAL FRAME				
PAST		PRESENT		FUTURE
PURPOSE				
NORMAL USE	ERROR DETECTION	IMPROVED FUNCTION		PRIVACY CONCERNS

Table 1: Coding scheme used to code participant responses.

Focus

The first category was analyzing the apparent focus of the question, as in what was the object at which the question was directed? This category had six sub-categories: input, process, output, user, other users.

Input

Questions that appeared directed at understanding what data or actions led to system behaviors were coded as Input (IPT). Examples of questions related to the input include “Can I see the data?” “Where are all the sources of data?” and “How clean/accurate is the input data?”

Process

Questions that appeared directed at understanding the techniques and algorithms used by the system to deliver outputs were coded as Process (PCS). Examples of questions related to the process include “What criteria is used, and how is it weighted (i.e., what is the recipe)?” “How can I see what’s behind all of this?” “What kind of software is this running?” and “What is the model built on?” Other questions that appeared to be asking about elements related to the process, such as how risk is considered (e.g., “How is risk measured”), or how much uncertainty the system has were also coded with PCS.

Output

Questions that appeared directed at understanding the output, particularly in determining its validity or appropriateness were coded as Output (OPT). Examples of

questions related to the output include “What are the pros/cons?” “What are the odds you’re right?” “How accurate is that system?”

User

Questions that appeared directed at understanding how the user themselves, and their inputs (including their data) were known to the system were coded as User (USR). Examples of questions related to the output include “Does it know & understand my goals?” “Does it understand my limits?” “Does the system’s goal match my own?” and “What does the system have on me? (What personal data is the system aware of and considering?)” “Can I see user ratings from other people?” and “What part of my profile does the computer care about most?”

Other Users

Questions that appeared directed at understanding how other users have interacted with the system, perhaps in order to determine how they should interact, were coded as Other Users (OTH). Examples of questions related to what other users have done include “How common is this suggestion?” “Who else has taken this suggestion?” “How have others fared when this suggestion was accepted?”

Temporal Frame

The next category was based on the temporal focus of each question. This category was broken into three sub-categories, past, present, and future.

Some questions were focused on past operations (PST), such as “When was this thing checked for bugs?” and “Does the computer have a good track record?” Other questions were focused on current time (CRT), such as “How much time do I have to think it over?” and “Is this data any good?” Other questions were focused on the future

(FTR), such as “What will happen if I say yes?” and “How is my feedback incorporated or considered in future recommendations?”

Purpose

The final category attempted to categorize the purpose or reason behind asking each question. There were four sub-categories: normal use, improved function, error detection, and privacy concerns.

Normal Use

Some questions appeared to be aimed at enhancing a user’s understanding of the system in order to determine how it should normally work. These questions were coded as Normal (NML), and some examples are “What does this system do really well?” “Is this what the system was designed for?” and “What are the limits of this system?”

Improved Function

Other questions appeared motivated by a desire to modulate or improve the system’s functions, possibly by correcting its errors or updating the inputs. Those questions were coded as Improved Function (IMP), and some examples include “What if I don’t want what is presented? Can I change how the computer works?” “What links up with this recommendation?” and “How will my decision affect the system?”

Error Detection

Some questions seemed mostly focused on attempting to determine whether the system was operating as designed, or whether its outputs were incorrect. Those questions were coded as Error Detection (ED), and included examples such as “Is this data any good?” “What is the ratio of false positives?” “What is the system’s level of confidence?”

"What dependencies are used in these recommendations?" and "Are these subsystems measured for accuracy/fidelity, or is data from these systems considered infallible?"

Privacy Concerns

Lastly, questions that appeared motivated by concerns over privacy, such as how their data was collected, used, and potentially shared with other systems were coded as Privacy Concerns (PVY). Some examples include "Does this system get my personal data (credit cards, health records, etc.) or is it just data from when I use the system (likes on Facebook, etc)?" "Why wasn't I warned about my data being collected?" "Is there hidden information the computer isn't telling me about?" and "Does the computer know me?"

Development and Description of a Taxonomy of User Knowledge Goals

Once these codes were developed, the next step was to assign them to each question in order to develop a taxonomy. Ram's taxonomy of knowledge goals (Ram, 1993) served as an inspiration for this activity. His taxonomy, however, while extensive, did not capture the spirit of motivations behind the kinds of questions participants asked in our workshop. This was in part because Ram's taxonomy was developed to help computer systems interpret stories in order to infer the motivations of people in those stories (early work that has led to substantive developments in a variety of fields, including context-aware computing). Other existing taxonomies, such as Bloom's Taxonomy of Educational Objectives (Adams, 2015), or Zachary's Taxonomy of Decision Support techniques (Zachary, 1986) were considered, but they did not accommodate the unique variety of question types asked by our participants. Thus, we

decided instead to develop a simple taxonomy of knowledge goals for understanding intelligent system recommendations, which we propose here.

In applying codes to questions, I did not limit the number of codes applied to each question, and reciprocally, not every category of code was assigned to every question. For example, the question "Is the data accessible to me?" was assigned the Input code because it is focused on the data that is considered by the system. It was also assigned the Present code because it is focused on the current situation, and it was assigned all four of the Purpose sub-categories (normal use, error detection, improved function, and privacy concerns) because those are all possible motivations behind the question asked.

Having coded all questions, the next step was to analyze the groupings of focus, time, and purpose in order to ascertain possible motivations (i.e., knowledge goals) driving those questions. The result of this analysis was a taxonomy of five knowledge goals, which are outlined in table 2 below.

Categories of the Taxonomy

Functional

A primary motivation behind many questions is to understand the system's purpose and function in order to understand how to use it appropriately. This knowledge goal is termed Functional, and represents the primary components of what makes up a good working mental model of a system (Streitz, 1988). A good example of a question that might be motivated primarily by a functional knowledge goal is "what is the system's goal." This type of question does not require an in-depth explanation of every component in the input-output relationship, and would instead likely be satisfied by a high-level descriptive explanation of what the system is primarily designed to do (i.e.,

this system takes user data from social media and websites and uses it to make predictions about what kind of products you might be in the market to buy in the next six months.).

Structural

Structural knowledge goals are similar to Functional, but represent a motivation to develop a deeper understanding of system behaviors and functions. While functional knowledge goals may be aimed at helping users understand how to use the system as intended, structural knowledge goals are aimed at the detection and resolution of anomalies, or cases where system events are unexpected or unanticipated by the user. Questions that seek hints as to how inputs and outputs are connected, and what processes are involved fall within this category, as well as questions that seek to examine inputs and processes in finer granularity. A good example of a question that may be motivated by a structural knowledge goal is "How aware is the system of the physical operating environment?" This question requires a deeper level of explanation, with details that go beyond a simple high-level description of system purpose and function. This kind of question is also aimed at understanding a specific element of system function. For example, in the question above, a user may be trying to understand whether or not an autopilot system considers the terrain in its routing suggestions in order to determine whether its routes can be trusted while flying through dynamic mountainous regions.

Normative

Normative goals are distinct from other goals in this taxonomy because they are focused primarily on the actions and opinions of others, rather than on the system or its outputs. Because the scenarios used for this study involved recommendations, it is not

unexpected that participants in this study wanted to know how other people factored in to the equation. For example, one way to assess the validity of a recommendation that seems out of place could be to seek information about the experiences of others who have received a similar recommendation. If a large quantity of other people, for example, were to report that they received a recommendation for a restaurant that seemed odd, but then provided positive reviews after having visited the restaurant, then a user might conclude that, while the recommendation initially seemed out of place, it must be valid for it to have been received and acted upon by so many others. Normative goals, therefore, represent the social cognition aspects of human reasoning, and in addition to knowing the reviews of others, may also include information about how the user is grouped in order to provide greater awareness of potential for algorithmic bias.

Personalizing

Personalizing goals are those aimed gaining an understanding of how the user is known to and modeled by a system. Because recommendations are typically intended to be personalized for each individual, users may want to understand what details about themselves go into making those recommendations. For example, if a music recommender assumes a user's taste based on their demographics, then music recommendations may not appear accurate if, for instance, the user's taste do not align with those of their primary demographic. Because the efficacy of most recommendations depends on the accuracy of the data, which in this case is highly personalized and may include data that some consider private, questions that are motivated by personalizing knowledge goals may represent both a desire to improve system outputs, as well as protect their personal privacy.

Predictive

Predictive knowledge goals are those aimed at understanding how user inputs will affect future outputs. This is another component of a good working mental model, since a solid understanding of system functions allow users to accurately predict what the system will do next under most circumstances. In cases where this is not clear, however, such as in cases where continuous interactions with systems create dynamic outputs, then users may ask questions in order to understand and choose their interactions. A good example of questions that may be motivated by a predictive knowledge goal is "What will happen if I say yes?" and "How will my decision affect the system?"

Knowledge Goal	**Description**	**Example participant question from workshop**
Functional	Seeking to generally understand system purpose and function in order to use it as designed	What is the system's goal?
Structural	Gaining a deeper understanding, primarily in order to detect or resolve system errors	What data is considered in making this recommendation?
Predictive	Understanding the input-output relationship in order to develop a predictive model	What will happen if I say yes?
Normalizing	Seeking to understand what others have done in order to determine what to do next	How have others fared when accepting this recommendation?
Personalizing	Seeking to understand how the user is modeled and known to the system	Is this recommendation made specifically for me, or is it generic?

Table 2: Table of knowledge goals. This was developed from the user-centered design workshop

Application of the Taxonomy

Examining the potential underlying motivations of why users ask questions is a useful means of identifying and prioritizing appropriate and effective answers to those questions. Designs that assume too much about what information users want run the risk of leaving too many questions unanswered, and thus possibly losing user trust, or risk providing information that is considered a nuisance and does not satisfy users' curiosity.

The taxonomy of knowledge goals is an attempt to expand knowledge about the types of questions users may ask when interacting with intelligent systems that offer AI-based recommendations. This taxonomy represents the range of potential motivations behind questions participants asked when imagining interactions with intelligent systems that offer AI-based recommendations. Considering a wide range of motivations behind why users ask questions may lead to designs that offer explanations that better satisfy user needs, and thus are more successful and accepted. Participants in this study appeared to have a broad range of motivations fueling their questions that went beyond those typically investigated by other intelligibility and transparency-related research (i.e., providing explanations primarily about system processes). This reflects that complex nature of systems that offer recommendations based on machine learning, which should be considered when determining user interface design.

User concerns about transparency in AI-based recommendations encompass a broad range of issues in sociotechnical systems, from functional to social concerns. As artificial intelligence and big data machine learning approaches permeate everyday technologies and bring recommender features to devices such as mobile phones, the

internet of things, and self-driving cars, the need for user interfaces to anticipate and address potential user questions will become more important. Using the taxonomy of knowledge goals in the design process may assist designers of AI-based recommender systems in identifying areas of potential concern.

The Explanation Vector Framework

I have proposed a taxonomy of user knowledge goals, outlined in figure 6 below. This taxonomy attempts to account for a range of motivations underlying questions users might ask an intelligent system to answer. The next step in this study was to determine how best to answer these questions in order to satisfy user knowledge goals.

To accomplish this goal, I conducted a secondary functional analysis of the questions recorded from the user-centered workshop using the code scheme described above. Whereas the first analysis attempted to classify the motivations behind questions being asked, this analysis focused on classifying the functional areas each question was

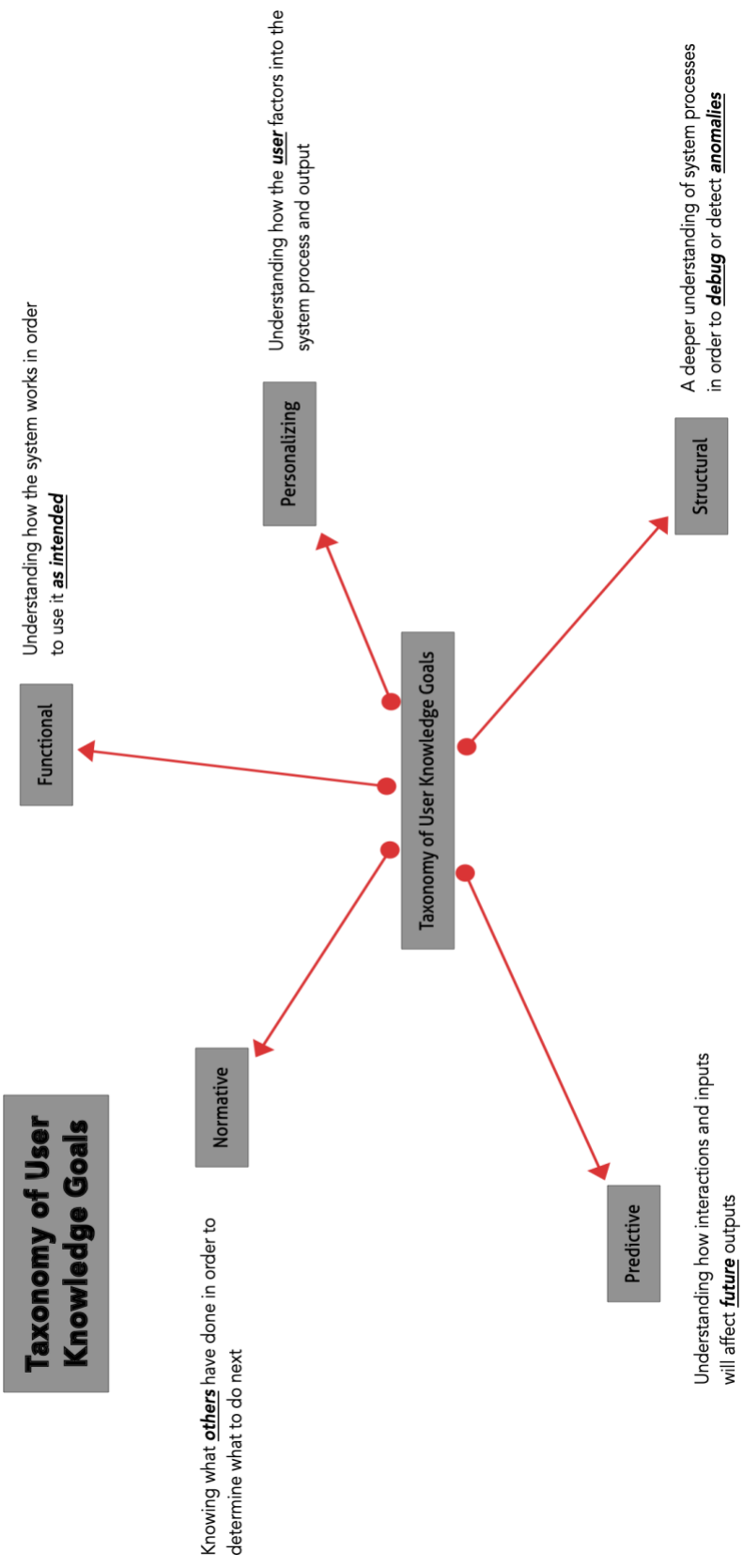

Figure 6: Overview of the taxonomy of user knowledge goals

aimed at understanding. For instance, questions such as “Can I see the data?” and “Is this data any good?” appear to be questions aimed at learning more about the data inputs of a system, whereas questions like “Does every user get the same recommendation” and “Does the system think I like this kind of thing?” appear more aimed at understanding what about a user the system knows, and how those data are used in deriving outputs. Using the same coding scheme as before, I examined each question and assigned codes according to the three analytical categories developed before, Focus, Time, and purpose. Examining the patterns of similarities amongst coded questions resulted in five categories of question types, which I call Explanation Vectors (EVs).

Description of Explanation Vectors

EVs are categories of information users seek in order to achieve knowledge goals related to understanding how system inputs map to system outputs. EVs can be thought of as categories of explanations, each of which associate with different aspects or dimensions of system transparency. Attempting to categorize different categories of explanations is a much-needed step towards developing systems that provide users the right information in a format that positively impacts their sense of usability and trust.

Current efforts to improve model interpretability/explainability are focused on explaining how machine learning algorithms work. System transparency, however, encompasses a wider range of features than just the inner workings of algorithms. For instance, many current systems that feature AI-based recommendations collect data from multiple sources, but do not reveal to the user what those sources are. Users seeking information of this nature would undoubtedly consider explanations pertaining to a model’s accuracy or level of confidence as insufficient for their questions related to from

where data is collected. Developers that fail to anticipate this broad range of potential user questions may experience barriers to acceptance, especially as AI-based recommendations are introduced to domains with higher inherent decision risk, such as personal financial management, or the healthcare market.

The explanation vector framework, therefore, attempts to identify this wide range of user concerns related to system transparency. The goal is to identify areas where interface design may provide explanations of system functions that directly address user knowledge goals.

Having coded all of the questions collected during the workshop and focus group activities, I identified five explanation vectors, which are described below.

System Parameters and Logic

A mental model is a person's mental representation of what something is, what it is for, and how it works (Rouse & Morris, 1986). Good mental models help users interpret, predict, and simulate system operations, and understand the system's limits. Users build mental models of systems through their experiences with them, which in turn determines subsequent interactions. Systems that restrict or hide information, therefore, can dramatically skew users understanding of those systems (Marwick & Boyd, 2011; Viégas, 2006), which in turn influences their use of and interaction with those systems.

Mental models can be broken into two broad types: Functional and Structural. Functional mental models are those that help a user interact with a system and understand how to use it. A person with a good functional mental model might be considered an ideal user of a system, or perhaps a "power user," someone who can operate and interact with the system in ways the system was meant to be used. Functional mental models,

however, do not include reasoning or understanding of what lies beneath the interactions, or what computations and processes make the system function. To understand this level, a user needs a structural mental model.

Much work in HCI is focused on assisting users develop functional models. A feature of "good design" is that users are guided and taught how to use a system naturally through the process of interaction, rather than requiring them to read lengthy user manuals (Dix et al., 2004). This kind of mental model is very important in order to achieve a benchmark of usability, and since most intelligent systems are designed for a mostly lay audience that is not expected to have a significant technical background, designing interfaces that build quality functional mental models is an appropriate goal in interaction design.

Functional mental models are useful and appropriate for 80-90% of user interactions, but are insufficient when system behaviors or recommendations do not align with user expectations. In these rare and unexpected cases, users can be left confused, wondering what happened, and uncertain about what to do next (Taleb, 2007). In these circumstances, users are motivated by the conflict between expected and observed outcome to find answers. In many cases, interfaces do not provide information that can help resolve these conflicts, for reasons discussed earlier.

Providing information about system parameters and logic, (i.e., how a system works, including its policies, logic, and limitations) can help users build appropriate structural mental models of systems, and help navigate or explain unexpected events. This knowledge is not only useful and important towards building trust in systems, but is often critical for safety as well. Numerous accidents, particularly in high-risk domains

such as aviation, have resulted from user actions born from inappropriate or inaccurate mental models of system functionality (National Transportation Safety Board, 2010; Sarter & Woods, 1995; Zeller, 1970). Users who are not afforded information about the internal architectures and functions of intelligent systems, may be forced to operate on mental models that are inaccurate or inappropriate.

Accurate mental models also help in anomaly detection as well. Numerous studies have demonstrated that providing information about how systems process information can help improve the detection of system errors and faults (Chen et al., 2016; Sadler et al., 2016; Sebok & Wickens, 2017). This can be very important in cases when system recommendations may be made in error, but are assumed to be correct by the user, who thereby acts on the erroneous recommendation, with sometimes disastrous results (so-called 'errors of commission' (Mosier & Skitka, 1996).

Whether to help build trust in system recommendations, or to prevent potentially unsafe events, the system parameters and logic explanation vector seeks to assist users in building appropriate structural mental models of system functions.

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 serve as a testament to the importance of providing information on the quality and provenance of the underlying data to decision makers (Fisher & Kingma, 2001).

Providing information about the qualities of data, such as its provenance, fidelity, and age, can help users determine when the system is operating out of limits, and may help them determine when system recommendations should not be used.

Providing users information on the qualities of data in machine learning applications has been shown to improve user ratings of ease of understanding, meaningfulness, and the convincingness of system outputs (Zhou et al., 2016). Advances in visual analytic approaches have also been shown to improve the comprehensibility and intelligibility of data to users by presenting it in a manner that is more readily understood (Muhlbacher, Piringer, Gratzl, Sedlmair, & Streit, 2014), and 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; Bae, Ventocilla, Riveiro, Helldin, & Falkman, 2017).

User Personalization

Recommendations are about predicting user preferences. The most successful recommender systems learn user preferences automatically by observing user behaviors and interactions with the system, such as when a user presses a thumbs-up button on a song, or listens to a song all the way through without skipping.

There are dozens of behaviors that can be used to generate predictions of user taste and preference without users needing to express their preference directly. In low-risk domains such as music recommendations, this approach is justified on the assumption that most users would prefer not to have to manually train a system to know what types of music they like. These passive measures of personalization are employed as a convenience to the user. What data is collected, and how it is processed to make

predictions and recommendations, however, is typically not shared with the user (Amatriain, 2016).

When recommendations appear out of place and inappropriate, however, users may want to understand this information. Knowing how they are modeled by a system, if at all, and to what extent system outputs are personalized for them could help resolve conflicts that arise from unexpected or inappropriate results. How willing a user is to investigate and search for answers in situations like this is likely driven by a range of motivating factors, including personality. Many users may not consider it worth their time to investigate, and so this information is not considered critical. But as recommender features spread to other, more high-risk domains like personalized medicine or financial planning, the importance of understanding user personalization increases.

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, perhaps prioritizing one aspect of financial growth, such as diversification, over other aspects that the user may prioritize more, such as long-term stability. Expressing these preferences requires two-way communication between the system and the user (Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004), which includes a representation of how the user is modeled and "known" by the system, and how that information maps to output functions such as personalized recommendations.

Research has also shown that user interactions are moderated by how they perceive the system (Park & Blenkinsopp, 2016). Users who are unsure about what interactions are recorded and used for predictions and recommendations may therefore "tread lightly" and feel less willing to explore and use a system. Conversely, research has

demonstrated that users who are afforded an understanding of how their personal data is collected and used to make personalized recommendations demonstrate more active engagement and higher feelings of control (Eslami, et al., 2015).

Social Influence

The central tenet of social computing is that computer systems that provide socially-related information better support everyday functionality (Wang, Carley, Zeng, & Mao, 2007). Digital realms are therefore structured in patterns that mimic structures of social life. The ways users interact with computer systems is deeply informed by social signs and strategies, which affect how users perceive and shape expectations.

The power of social media has been displayed in a variety of contexts over the past decade of its modern existence. Its role in daily life has morphed beyond a simple photo sharing tool to become a powerful tool for marketers and influencers as well. User data from social media has become highly lucrative and commodified. Systems that group users according to online behavior in order to predict preferences are abundant, and represent a new standard in modern marketing and sales (Adobe Inc., 2018).

A user's understanding of how they are grouped by a system using social media information, (i.e., social influence) can provide meaningful insights into why a system output, such as a targeted advertisement, was generated, and can help users resolve conflicts that may arise between a user and an inappropriate system output.

Providing a user information about how they are categorized and grouped socially may also affect decision making 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 (Tversky & Kahneman, 1981), 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 (Fiske & Taylor, 1991). 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 (Horrigan, 2017).

Our study found that many questions users have about system functions or behaviors originate from a desire to compare themselves to others who are similar to them (what we have termed normative goals in our taxonomy of knowledge goals). For instance, many participants asked variations of this question: "How many others have received this recommendation, and what is the ratio of accept/reject?" The goal at the heart of this type of question is likely to help users determine their actions by establishing a reference to existing norms. This explanation vector is seen frequently in shopping recommendations (e.g., people who bought this item also bought these items), and may be an effective means at encouraging user understanding in other contexts as well, especially those particularly associated with the social dimension.

Justification of Options

People often express a preference of choice over no choice in most decision-making contexts (Blume & Easley, 2008). Accordingly, many systems strive to offer choices to users as a means of increasing engagement and satisfaction (Preece et al., 2015). There are times, however, when providing multiple choices to a user may be undesirable.

To use the GPS example from earlier, 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 be unlikely to benefit the user, and may in fact lead them to discard the technology due to its confusing and busy interface.

This 'tyranny of choices' (Schwartz, 2004) 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 infeasible to display every possible optional output to the user, as the numbers alone may range into the many billions, depending on the application domain.

Common interface design strategies involve efforts that reduce choices in order to lessen cognitive load and improve the speed and efficiency of decision making (Rose, 2006). Resolving conflicts between interface aesthetics (i.e., clutter) and user preference for options often involves some sort of tradeoff. Sometimes these decisions are determined by external factors, such as corporate policy, or mandated safety requirements (Zahabi, Kaber, & Swangnetr, 2015).

In some contexts, however, reducing options in order to declutter or simplify an interface may have more of a negative effect on users, especially when those options may include a user's subjective preference. A classic example of this would be a music recommendation that is provided, along with a list of associated songs the user may also like. In this case, providing a justification of options considered by the system gives users a sense of control, as well as potentially informing them of how an algorithm may be recommending music (in this example, if all of the songs were of similar genres or

featured strong lead guitars, the user might infer how the algorithm is suggesting music).

Another reason why showing users a range of options that are considered by a system is to highlight the importance of providing justification for why one option is chosen over another. As mentioned previously in Chapter two, much work has focused on the importance of providing system explanations that include a justification (Gregor & Benbasat, 1999; McGuinness, Glass, Wolverton, & Da Silva, 2007). Early generation intelligent systems largely failed to achieve widespread acceptance, in part, because their explanation functions could only provide explanations in the form of system rules, but could not justify why one action or decision was favored over another (Swartout & Moore, 1993). Users express preference for systems that can provide answers to their questions that help justify system behaviors, including recommendations (Swearingen & Sinha, 2001). Providing this information, therefore, may do well to increase a user's willingness to engage with system outputs, which may translate to gains in usability and acceptance.

Explanation Vector	Information involved	Results in
System parameters & logic	Internal architectures, policies, reasoning strategies, limitations	development and improvement of appropriate mental models
Qualities of Data	Sources of data; interdependencies; qualities of data such as provenance, age, fidelity	improve user's understanding of cause and effect relationships; determine when system is functioning out of limits
User personalization	How the user is modeled by the system, and how that information is used to derive system outputs	Moderates interaction behaviors; guides users to learn personalization features; promotes active engagement and a sense of control
Social Influence	How the user is categorized and grouped with others according to tastes and behaviors	Understand recommendation outputs; aid decision making using social info; assess relative norms
Justification of Options	Total range of options; justification for chosen options	inferences of algorithmic reasoning; development and improvement of appropriate mental models

Table 3: Taxonomy of Explanation Vectors and their associated purposes

Discussion

This study sought to answer what questions users ask when interacting with AI-based recommendations, and attempted to identify some potential motivations behind those questions. To accomplish this, I created five design fictions of future intelligent systems depicting user interactions that resulted in unexpected or surprising outcomes.

These scenarios were designed to elicit confusion and prompt users to investigate deeper in order to determine what happened and how to respond appropriately. Utilizing a user-centered design workshop format, I encouraged research participants to imagine themselves as the central character in these interactive vignettes, and then recorded each question participants asked in response to each vignette.

Using an open coding method, I categorized these questions according to the apparent motivations behind them, leading to the development of a taxonomy of user knowledge goals. This taxonomy represents the range of potential motivations behind questions participants asked when imagining interactions with intelligent systems that offer AI-based recommendations. This taxonomy may be used to consider explanatory interface techniques that are aligned with user knowledge goals.

A secondary analysis and coding activity considered what each question was attempting to answer. This coding activity led to a framework of explanation categories which I termed explanation vectors, detailed in table 3 above. Explanation vectors are categories that address user knowledge goals. This framework expands the domains of interpretability and explainability, and encourages designers to consider a broader set of explanations beyond those that focus solely on the inner workings of algorithms.

Limitations

This study has focused specifically with intelligent systems built on machine learning that aid in decision-making by offering recommendations to their users. These systems may afford users various other services, but only those that offer recommendations are relevant to this investigation.

The use of design fiction, especially the intentional design of each vignette to confuse or surprise the participants, may seem controversial, as these kinds of system responses may be considered anomalies. The systems chosen for this activity are based on actual systems currently in development. It is difficult to determine how they may function once fully developed. It is likely that many of the kinds of anomalous responses featured in these interactive vignettes may be completely eliminated by the time these systems become widely available to the public. Participant responses to these vignettes, therefore, may not reflect the kinds of questions that real-world users may ask under ordinary circumstances. Nevertheless, the use of startle and surprise in the vignettes was chosen because previous research has shown that these kinds of unusual, out of place, and unexpected system outputs are often the trigger of much larger mishap events. These circumstances, therefore, represent situations in which system transparency may be considered most critical. For this purpose, the vignettes were written the way they were.

The user-centered design workshop was conducted with participants that, while familiar with human-computer interaction and UX best practices, were not subject matter experts. Hence, participants of our study may not fully represent the entire range of user types. These findings, therefore, may not represent a complete taxonomy of knowledge goals, and additional explanation vectors may exist.

An additional dimension not assessed was how information is made available, whether on-demand or proactively delivered. Users will likely prefer answers to their questions in a manner that is quick and easily accessible, but that information may seem intrusive or obstructive if delivered proactively. In future studies, we plan to evaluate information demand on this dimension in order to determine what information is best

delivered proactively, and what information should be made available on a drill-down basis.

Finally, this workshop featured think-aloud, semi-structured discussion activities, meant to elicit spontaneous and unstructured user feedback. Formal design investigations in more controlled environments using real or simulated systems may produce different results.

Conclusions from Study One

I have investigated the kinds of questions users ask of intelligent systems that offer recommendations. Examining these questions has revealed a range of motivations behind user questions, as well as identified a range of explanation categories that should be considered by interaction designers looking to improve the transparency of their systems.

The utility of intelligent systems is evident, but adoption can be hindered when users cannot understand the system's reasoning. Users who interact with these systems will need explanations of its inner workings in order to establish and maintain sufficient and appropriate trust. Systems that do not explain themselves well are likely to encounter barriers to technology acceptance.

CHAPTER FIVE: DEVELOPMENT OF A DETAILED USER TYPOLOGY OF KNOWLEDGE GOALS

The findings from Study one represent a good first step towards identifying and understanding the kinds of information that end-users may seek in response to unanticipated or unusual recommendations. Perhaps most valuable of these findings is the recognition that users in my study wanted answers to questions that extend far beyond the kinds of information that is currently the focus of much of the research on interpretability or explainability. This suggests that as these technologies mature and become more main stream, interaction designers may have to manage a much wider range of information requirements than previously considered. This quickly leads to a question: in situations with limited screen real estate and multiple competing design elements, how do interaction designers prioritize that explanations are provided to users, and what questions must remain unanswered, an example of the transparency paradox discussed in Chapter two.

To answer this question, I designed a study that allowed me to examine patterns in user interactions with AI-based recommendations. These patterns, which were used to create a detailed user typology, can be used to identify user preferences for information, which can then be used to prioritize explanations.

This chapter is broken up into two main sections. First, I introduce the methodology used for this study, known as Q-methodology, and explain in detail the process through which I employed it for my research to develop a detailed user typology. Next I outline the study itself, and discuss the methods, results, and their limitations.

Background and Motivation for Study Two

A constant challenge for interaction designers is that they must wrestle with the practicalities of designing with limited screen real estate. Cluttered, busy visual environments, such as those found in many commercial airliners today, have long been the focus of much research because of their potential to confuse users. Because of this, designers and engineers must often work to reduce the number of graphical icons and menu items, and streamline their designs so as to produce an interface that is visually appealing and maximizes usability (Lidz, Pietroski, Halberda, & Hunter, 2011). Those same principals form the basis of much work in other interface designs, from web-based, to vehicle-based, to mobile and even wearable devices. The results of this drive to reduce clutter often means that designers must determine what elements or information is most critical, and therefore given priority, and what elements or information is not necessary, and therefore omitted from designs.

The work completed in study one suggests users may ask a broad range of questions that span multiple categories. Attempting to build an explanation engine that could potentially answer all of these questions is impractical, and would most likely result in reduced usability. Designers looking to utilize the findings from study one, therefore, would require guidance to make difficult decisions about what questions should be answered through interface designs to resolve these information priority conflicts.

The vignettes used in study one spanned a wide range of potential domains and decision contexts where AI-based recommendations may soon emerge. These results, while informative, do not provide detailed information about what questions might be

asked in specific decision contexts or domains. In other words, a user presented with an AI-based recommendation about their personal financial investments is likely to ask different questions than a user presented with an AI-based recommendation about what neighborhood they may want to look for a new home purchase. The goal of this study, therefore, is to see if there are patterns of user expectations that are domain-agnostic, which could help designers by providing a baseline with which to begin providing basic explainability.

Making the Case for a User Typology of Knowledge Goals

Research in human perception and decision making has confirmed that information is not homogeneous, and that some information is more influential than others (Mumaw, 2017; Mumaw, Roth, Vicente, & Burns, 2000; Parasuraman, Sheridan, & Wickens, 2000; Riveiro, Helldin, Falkman, & Lebram, 2014). Many factors play a role in influencing what information is considered valuable and important. One of the most challenging is the effect that individual differences have on perception and problem solving. Because a variety of individual issues including domain-specific knowledge (i.e., expertise), previous experiences, and attitudes towards technology are involved in the formation of a user's mental model of a system, strategies for understanding will be different (Liu & Stasko, 2010; Rouse & Morris, 1986; Streitz, 1988). Those strategies come in the form of questions a user might ask the system in order to satisfy their knowledge goals. Much work has been done to identify information-seeking patterns in human perception (Amirkhiabani & Lovegrove, 1996; Geng & Behrmann, 2005; Yantis & Jonides, 1990). To the extent that these patterns are stable and may generalize across

use cases, they could be useful in helping designers anticipate these interactions and therefore prioritize explanation design features to meet user needs.

The ability to personalize information based on the type of user would be particularly important where decisions must be made in uncertain or ambiguous situations. Explanations tailored for a person's cognitive and affective traits would be highly effective at helping determine the problems identified, the solutions considered, and the decisions that ultimately result. But tailoring this information to match each individual user's schemas or knowledge goals is a near impossible task, and would require extensive and intrusive testing for each user in order to provide this personalization.

One solution is to develop a user typology that is broadly representative of most users. Typologies are abstractions formed from the components of a given phenomenon (Kakar, 2016). By looking at generalizable systemic regularities within groups, typologies can enable descriptive understanding of those groups by analyzing them independently, and contrasting them with one another. Typologies can add to the insights commonly gained by traditional demographic analysis, and because they are formed on the basis of mathematical affinity, groupings and their associated traits can be quantified.

To develop a detailed user typology, I utilized a factor analytic mixed method known as Q-methodology, which is introduced in detail in the following section.

Introduction to Q-Methodology

Q-methodology is a mixed method designed to identify and analyze patterns of sentiment and opinion in individuals. Q-method is often referred to as the *scientific study*

of subjectivity (Watts & Stenner, 2005), because it seeks to explicitly elicit and study people's subjective opinions and sentiment.

Q-methodology does this by using a factor analytic approach that allows for the identifying of patterns of subjectivity and thought in the data, which is used to identify factor 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 a traditional qualitative analysis, effectively combining the strengths of both qualitative and quantitative research, adding deep texture and nuance to the data that may otherwise be passed over by a purely mathematical approach.

Q-methodology is well-suited for studies interested in examining subjective opinions and values, and exceeds other survey-based methods in terms of both depth of analysis and mathematical rigor (Watts & Stenner, 2005). By examining not only how people rank items of interest as best/worst, but examining the tensions between those items, q-methodology enables a deep evaluation of shared opinions and points of view, as well as tradeoffs and priorities- all potentially important information for studies interested in how user's think about design features.

Q-methodology involves developing a bank of statements or questions, and then having participants rank order those statements according to how important they are to them. Q-method uses a forced distribution matrix as opposed to other kinds of scale rankings (i.e., Likert scale). Because this distribution is forced, participants must carefully consider each item's importance in relation to every other item, thus eliminating the possibility of over or under-inflating score bias, or a tendency for participants to rank all items around the mean (Watts & Stenner, 2005). And since all items are ordered in

relation to all other items, this also means that rankings using Q-method more accurately describe a person's priorities and sentiment, since this is how people must resolve competing priorities in the real world.

Once items are sorted, each arrangement (known as a q-sort) is combined with all other q-sorts to create a correlation matrix. This correlation matrix is then submitted to factor analysis, which mathematically groups participants together based on the positions of their cards. Participants grouped together, therefore, will have significantly similar arrangements of cards, indicating shared patterns of opinions and priorities.

The next step is to analyze these groups in order to derive their meaning. This can be done using traditional qualitative textual analysis (van Exel & de Graaf, 2005). Having identified and interpreted the different factor groups, further analyses can then be conducted to compare and contrast the groups on a variety of levels, according to the needs of the researcher.

The next section will outline the steps taken to develop the concourse and Q-set for this study, followed by the research design used for the Q sorting exercise.

Step One: Development of the Concourse

The first step in using Q-methodology is to develop a set of statements or questions that represents all of the possible viewpoints or opinions about a given issue or topic. This set of statements or questions is known as a concourse. The purpose of the concourse is to represent the entire range of opinions or sentiment surrounding a given issue in order that it is fully represented. A complete concourse will therefore represent a theoretical 360 degrees of perspective. In order to accomplish this goal, it is therefore critical that all aspects of a topic or issue are considered, including the person, their

environment, their role, and other contributing factors that may affect a person's point of view.

To develop the concourse for my study, I began with the questions recorded from the user-centered design workshop in study one. Recall from chapter three that study one involved a combination of semi-structured interviews, think aloud activities, and affinity diagramming, during which users developed a wide range of questions. After I analyzed all of these questions, I then augmented them with the help of subject matter experts in intelligent systems. These experts were asked to review the vignettes that I had developed, and then look over all of the questions. They were then invited to add additional material to cover any technical or theoretical areas our participants from study one did not address.

Additionally, I further surveyed scientific and periodical literature to ensure that all theoretical issues related to transparency in recommender systems were represented. For example, I used a series of lecture notes from Carnegie Mellon University's lecture series in AI to add questions from various ethical perspectives. I used newspaper articles from recent stories to add additional material pertaining to user privacy.

After these activities were complete, I had created a concourse of 81 questions.

Step Two: Development of the Q-Set

The next step in the Q-method process is to refine and distill this bank of statements or questions down to a smaller, representative sample that will be presented to participants for them to sort. This smaller bank of statements or questions is known as a q-set.

Because it is important that the Q-set retain its representativeness to the original concourse, choosing what statements to present to research participants is a critical decision. There are many methods that can be used to accomplish this task, and currently there are no specific rules, though best practices suggest that utilizing objective techniques that take into account measures of agreement obtain the most robust results.

For this study, I used an industry standard known as the Content Validity Ratio (CVR) method (Lawshe, 1975) to reduce the concourse down to a Q-set. The CVR uses measures of agreement to determine what items are retained, and what items are discarded, using a panel of subject matter experts to rate items into one of three categories: "essential," "useful, but not essential," or "not necessary."

Items that are ranked "essential" by a critical number of experts are retained, while items failing to achieve this critical number are discarded. The CVR is a linear transformation of a proportion of agreement between experts:

$$\frac{CVR = Ne - \left(\frac{N}{2}\right)}{\frac{N}{2}}$$

N_e is the number of panelists who rate an item "essential." N is the total number of panelists. CVR values range between −1 and +1 with -1 being perfect disagreement, and +1 being perfect agreement. Any CVR value above zero signals at least half of the panel members have rated an item essential. To account for random chance agreement, Lawshe (Lawshe, 1975) developed a table of critical values, or the minimum number of panelists needing to rank an item essential in order for the item to be retained, using a probability of Type I error at 0.05. Because the CVR critical values are based on the

normal approximation of binomial probabilities, panel sizes below 8 are required to achieve a CVR of 1 in order to retain items.

To utilize the CVR method and determine what questions would be retained from my original concourse, I arranged for a panel of 5 subject matter experts, with extensive AI experience in academia and industry. Panelists were given a table of questions from the concourse, and were ask to sort those according to the CVR method. Items that received a CVR of one were retained, while those with a CVR of less than one were discarded. While there is no predefined size limit to how many cards can be in a Q-set, industry standards suggest limiting the number to less than 60 (Watts & Stenner, 2005). After all panelists had concluded rating all of the statements, the result was a final bank of 36 questions for the Q-set.

The next step in this process was to randomize all of the questions and assign each question a number. Each question and its corresponding number was then printed on 3x5 index cards. Each deck, therefore, was comprised of 36 cards, each with its own individual question printed on it.

Step Three: Sorting

Once the Q-set has been developed, the next step in the process of Q-methodology is to have participants sort the cards and record their arrangements. This process typically involves introducing a topic to the participant. This can be done through a variety of ways, such as preparing summarized news articles, or simply by asking the participant to consider a topic in their mind.

Once the topic is introduced, the participant is invited to arrange their cards in a fixed distribution matrix from those that are most important or relevant to them, to those

that are least important or relevant to them. The exact wording of this arrangement is dependent on the purpose of the research study. For instance, participants may be invited to sort cards according to statements that most represent their opinion, or characteristics that are most like themselves, etc.

The distribution of cards is intended to create a normal or quasi-normal distribution of cards, such that the cards on the outermost tails of the distribution represent the extremes of the participant's opinions, and the cards in the utmost center represent the average. Each card is assigned a value based on the column in which they are placed. In the study that I ran, I used a matrix that utilized 11 columns, from -5 to +5, including 0. So, cards that are placed in the +3 row are all given the value of +3, and cards that are placed in the -4 row are all given the value of -4, etc.

Once the participant has completed sorting all of the cards into the matrix, their arrangement is recorded, and the statements and associated values are transferred to a table for analysis and interpretation.

Step Four: Factor Analysis

The next step in the process is to analyze the arrangement of cards. Once the cards and their associated values are recorded, each participant's q-sort is 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). Participants with high correlations between them, therefore, will have arranged their cards in a very similar manner. These correlations are useful in initially identifying groups of similar opinions, but the real purpose of the correlation matrix is to prepare data for factor analysis.

Factor analysis is a statistical process that analyzes groups of data and produces factors onto which variables load based on their affinity or correlation. In Q-method, factor analysis identifies participants who load on factors, which identifies groupings or families of individuals who have arranged their cards in very similar manners, which translates to their attitudes or opinions being statistically very similar.

Because the factors in Q-method are made up of participants rather than individual variables (such as is the case in traditional R factor analysis), it may be easier to consider factors as groups of individuals. For this reason, I will refer to factors as factor groups henceforth.

The next step in the process is to interpret each factor group in order to understand their attitudes and shared opinions. To accomplish this, all sorts within each factor group are averaged together in order to create a composite of their sorted cards. This composite sort can then be examined in detail in order to interpret the factor group and uncover their attitudes and opinions they share in common. Once this is accomplished, this data can be used to qualitatively understand and investigate the subject or issue being studied.

The analysis process, from creating correlation matrices to conducting factor analyses, can all be accomplished using a variety of statistical packages that are freely available. For this study, I used a purpose-built Q-method statistical software package known as Ken-Q, developed by Shawn Banasick (Banasick, 2018).

The following section outlines in detail study two, which used Q-methodology to investigate patterns of user interaction with AI-based recommendations in order to build a detailed user typology.

Introduction to Study Two

Recall from earlier that study one explored what questions users might ask of intelligent systems that offer AI-based recommendations. The questions collected represent a wide range of questions users might ask when interacting with AI-based recommendations, but in order for those questions to be useful as design guidance for future interface design, they must somehow be prioritized.

As discussed in the introduction of this chapter, one innovative method of prioritizing explanations is to consider the user and their knowledge goals, and use that information to provide explanations that target those goals. Systems that are able to assess the user and provide individualized explanations tailored to their user type are more likely to achieve the right balance between transparency and usability.

This study was therefore designed to develop a detailed user typology of knowledge goals. This study sought to answer the following research question:

Generally, what are the range of these information seeking strategies, across a broad representative spectrum of decision contexts related to intelligent systems?

Since it is possible to map design features to potential user knowledge goals to help resolve conflicts or confusion, exploring and identifying the different range of user understanding strategies can help to prioritize which explanation vectors are more useful and should therefore be prioritized in interface designs.

This study investigated the effect of user perceptions and information priorities on the value and effectiveness of explanation vectors when interacting with intelligent systems using the factor analytic approach of Q-Methodology, described earlier.

Methods

35 participants from the UK, and 75 participants from the US participated in this study for no compensation. 89 were male, 21 female, average age was 29 years old. 71% of our participants reported little or no expertise of intelligent systems or artificial intelligence. Expertise was established by self-report, and measured by working knowledge of one or more commonly used programming languages (Python, MATLAB, Keras, Caffe, TensorFlow, Java, Scala, C/C++, Flask, Torch/Lua, Javascript, HTML5, CSS3, R) and experience designing or programming one or more type of intelligent system application (recommender, context-aware, clinical decision support, tactical decision support, natural language processing, visual classification, other machine learning).

Participants were each given their own deck, and provided an example of the sorting diagram (shown in figure 7 below) with instructions for how to sort cards from most valuable or important, to least valuable or important to themselves. Once instructions had concluded, a vignette was displayed on a computer screen or projector. The same vignettes used in study one were used for study two. Each vignette described an interaction that results in a system-generated recommendation which seemed inappropriate or unexpected, thus prompting users to ask questions in order to understand why. Full vignettes are available in the appendix.

For each vignette, participants were asked to imagine themselves as the central character in each scenario. Participants were given instructions to "Sort the questions according to which ones you would MOST want to ask the system in order to feel

comfortable using this output." Participants sorted their cards in this manner for each vignette, resulting in a total of five sorts per participant.

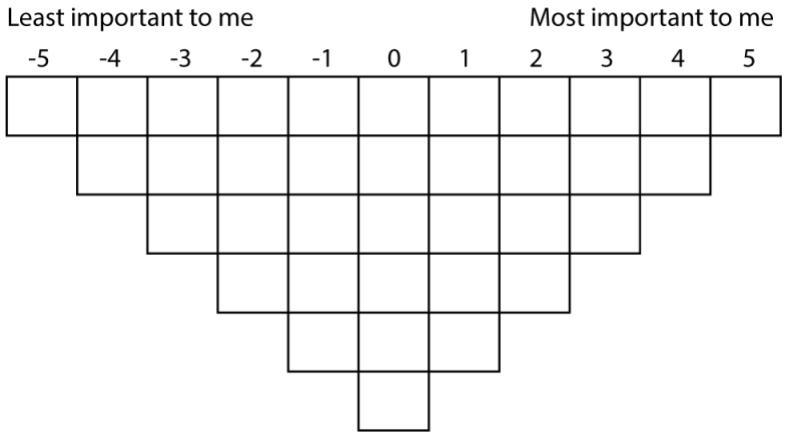

Figure 7: Sorting matrix used for the Q-methodology study. 36 questions were sorted for each scenario into this matrix, ordering questions from those most important to the participant (+5, extreme right), and those least important to the participant (-5, extreme left).

Once participants had completed sorting their cards, they answered two additional questions on a questionnaire: "In a few words, please explain WHY you chose your MOST/LEAST important question to ask." Participants wrote their answers on the provided form, which were then collected and prepared for analysis.

Results

Factor Analysis

Data was recorded on special sheets of paper for each sort. Those sheets were then collected and data was inputted into an excel spreadsheet and prepared for analysis. Once prepared, all sorts were combined into a correlation matrix using the Ken-Q software (Banasick, 2018). This matrix was then submitted to factor analysis using the principal components analysis (PCA) method for factor extraction (Ford, MacCallum, & Tait, 1986). Eight initial factors were extracted using PCA.

Several possible solutions were tested, ranging from two to eight factor groups, by examining each factor's eigenvalue and total amount of explained variance. A four-factor solution was ultimately chosen 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. Using a scree test below in figure 8 supports this decision.

Using the VARIMAX method to obtain orthogonal rotation of the factors (Devore, 1995), the four factors were rotated. 11 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 81 individuals.

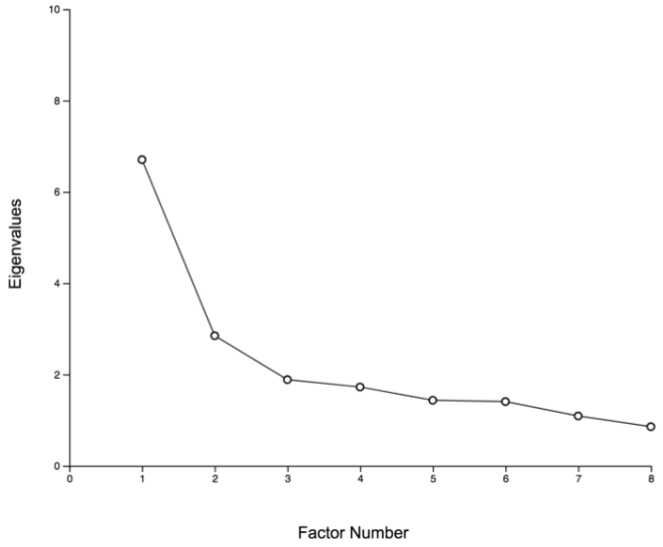

Figure 8: Scree plot of initial factor analysis before extraction and rotation

Factor Interpretation

In order to interpret the viewpoints of each factor group, I produced a weighted average of each participant's arrangement of questions, then combined each individual's

arrangements into one exemplar composite arrangement per factor group. This composite arrangement, or "factor array," was developed for each factor group, then analyzed by examining the relative placement of each question in relation to each other question.

The results of this analysis are provided below for each factor group. After a description and interpretation of their sorting strategies, I provide a between-groups quantitative comparison analysis.

Factor Group 1: Interested & Independent

Factor group one was defined by 24 participants and explained 14% of the study variance with an eigenvalue of 20. 71% reported they had little to no working knowledge of intelligent systems. Roughly 60% were less than 40 years old.

An analysis of the top distinguishing questions of this group reveals what seems to be an intellectual curiosity, and a posture of openness towards developing a deeper knowledge of system functions. For example, the highest rated question was “Why is this recommendation the BEST option,” indicating a desire for a justification of recommendations beyond a simple 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, $Z = 1.06$, $p < 0.01$). Full details on Factor Group 1’s composite sort can be found below in figure 9.

Responses to the open-ended questions help to clarify and refine this impression. Participants explained their reasoning for why they prioritize this type of information

over others. For instance, one participant expressed “Because I want to understand the system’s reasoning in order to compare it to my personal experience.” Another added that “there are many, many factors that I would like to consider, and it would be worth the effort to explore the data. I could learn a lot about the system by just looking at its data structure,” and another said “If the system is in err, I want to prevent it from happening again. This seems the best way.” The placement of the highest rated questions along with these open-ended explanations help to define this group, which seems to primarily be interested in understanding system processes.

Analyzing the lowest rated questions revealed a general disregard for the social aspects of explanations, and the opinions or behaviors of others when considering what to do with computer-generated recommendations. Participants in factor group one 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.

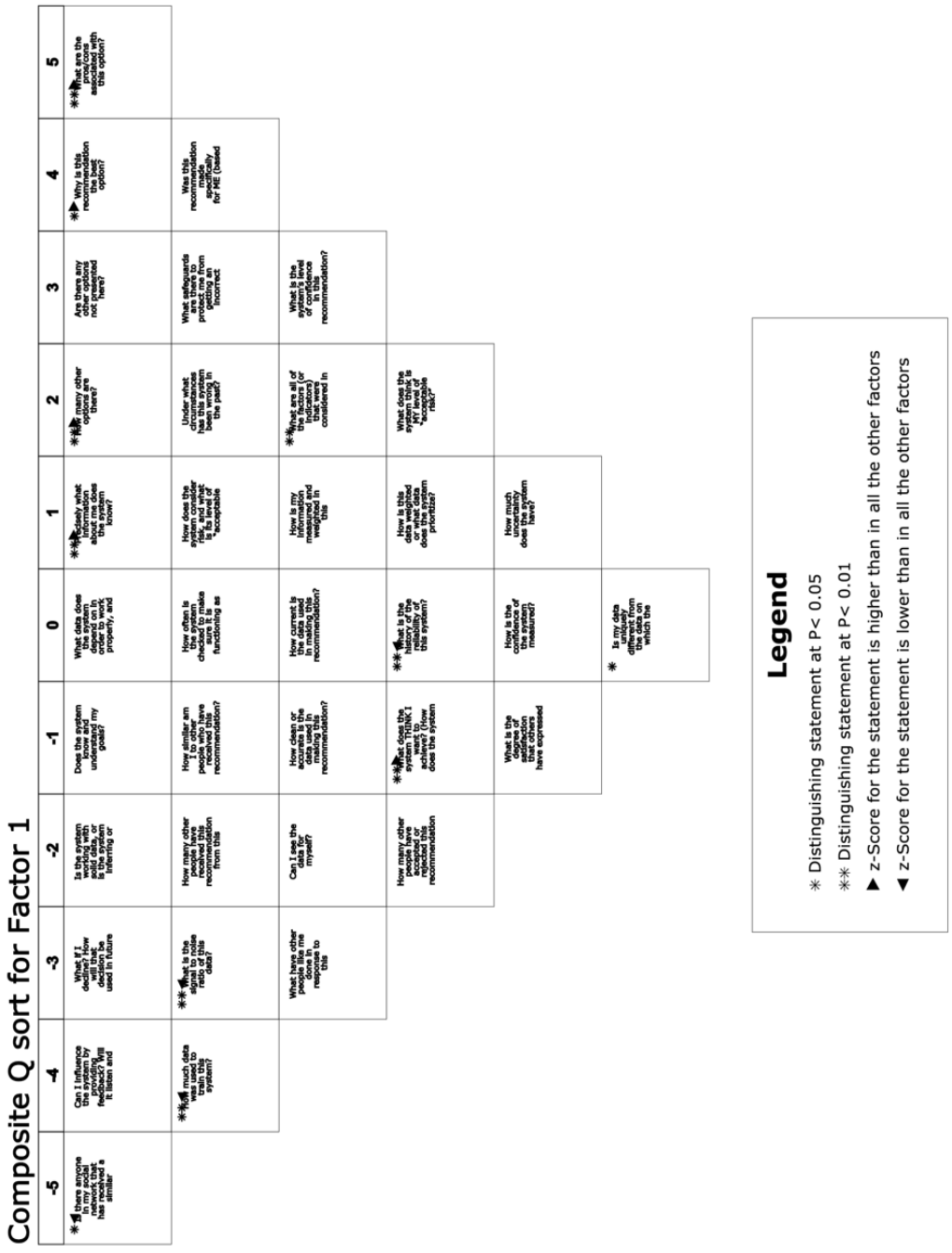

Figure 9: Composite sort for Interested and Independent group

Again, I used responses to the open-ended questions to help understand and refine my interpretation of this factor group. Comments such as “I don’t care about non-expert opinions,” “I just don’t care what others do,” and “although it might be interesting to know if others have received similar recommendations, I don’t know how valuable that information will be since their circumstances could be much different than mine” seem to confirm that these individuals do not place a high value or importance on details of what other users do in response to AI-based recommendations.

Factor group one therefore seems to be characterized by both a willingness and desire to learn a deeper knowledge of system functions, and an independent outlook that is typified by the tendency to rank socially-related information as least important to them. Hence, I named factor group one “Interested & Independent.”

Factor Group 2: Cautious and 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.

Participants in factor group two most want 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, therefore, seemed to possess a deep concern over a system's past performance and reliability, and seemed to prioritize information that can help them establish risk and minimize uncertainty.

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, $Z = 1$, $p < 0.01$). Factor Group 2's composite sort can be found below in figure 10.

Analyzing their open-ended feedback helped to clarify why these participants ranked these questions as their most important. For example, one person simply wrote "If not reliable, I don't care about it." Others supported this sentiment, for example, "I want to know when the system has been wrong in the past so I can compare it to my situation. Since there may be severe consequences, I need to know what could make the system wrong." Another wrote "The entire outcome mostly works based on the accuracy of the data. The first step in a predictive model is to make sure we are giving the system the right input."

What participants in factor group two seemed to devalue the most were questions that pertained to social information, or how information may be personalized for a user. The question this group listed as least important to them was "Is 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, $Z = -1.32$, $p < 0.01$).

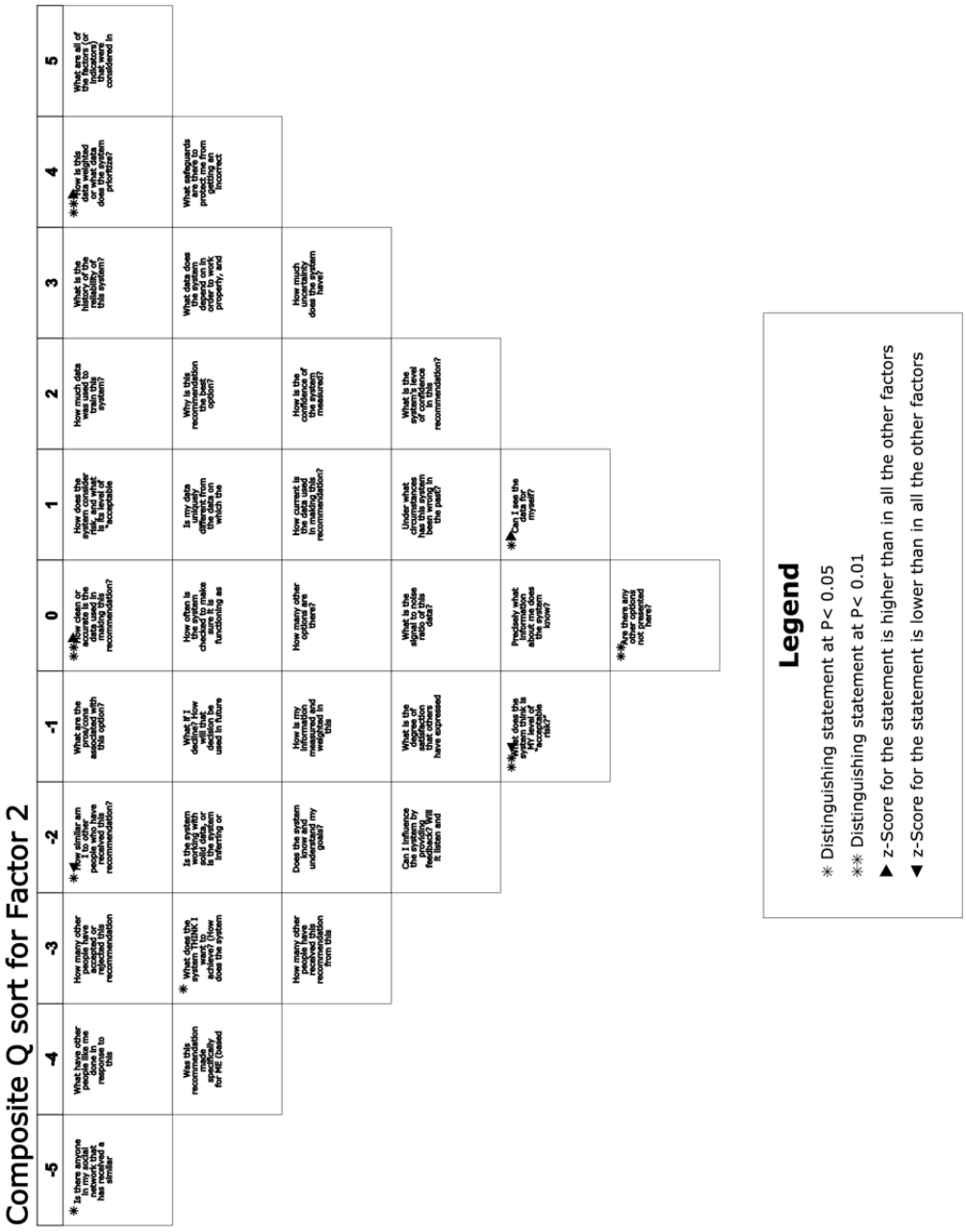

Figure 10: Composite sort for Cautious and Reluctant group

To better understand these findings, I analyzed their open-ended feedback. Although these questions appear somewhat disconnected from one another, factor group two's comments suggest a general distrust, or perhaps reluctance to engage with systems that embody artificial intelligence in the form of recommendations. For instance, "I don't want to have to reverse engineer. The system either knows what I want, or it is guessing. Either way, I am not that interested." Another said "I expect it to work, or at least to have someone who knows what they are doing behind the scenes. I am not interested in programming anything myself."

These comments, combined with the placement of questions, suggests that participants in this group harbor a general distrust and attitude of reluctance to engage with or trust artificial intelligence, or perhaps technology in general. For this reason, I named factor group two "Cautious & Reluctant."

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 ranked "Why is this recommendation the BEST option?" (composite score 5, $Z = 1.75$, $p < 0.05$) as their most important question, followed closely by "What are the pros/cons associated with this option?" (composite score 4, $Z = 1.25$, $p < 0.01$). Recall from earlier that this question was also the most important question to the Interested & Independent group. But where the Interested & Independent group sought additional details about system functions and logic to help

them further understand, factor group three seemed to use a different strategy. Their preference appeared to be to seek information about the opinions of others.

For example, the second most important question ranked by factor group three was “what is the degree of satisfaction that others have expressed when taking this recommendation?" (composite score 3, $Z = 0.9$, $p < 0.01$), followed by “how many other people have accepted or rejected this recommendation from this system? (What is the ratio of approve to disapprove?)" (composite score 1, $Z = 0.29$, $p < 0.01$). Factor Group 3’s composite sort is below in figure 11.

Analyzing the open-ended comments from factor group three provided deeper insights and helped to refine the interpretation of these data. One participant said “I want to know that the system has made the right choice for me and my lifestyle/preferences-has it really taken all my situations and personal feelings into consideration?" Another participant remarked, “I do not care about possibly situational circumstances in my social network. Rather, the input-output pairs, i.e., choosing to accept/decline and the result is more important for me.” Lastly, “I am curious what other people’s views are, and what they would do in the same situation,” and “I don’t want to go it alone. Knowing how many others have been in my situation would help to boost my confidence to make decisions.” It appears from these arrangements that both factor group three and the Interested & Independent group want some justification from the system before feeling

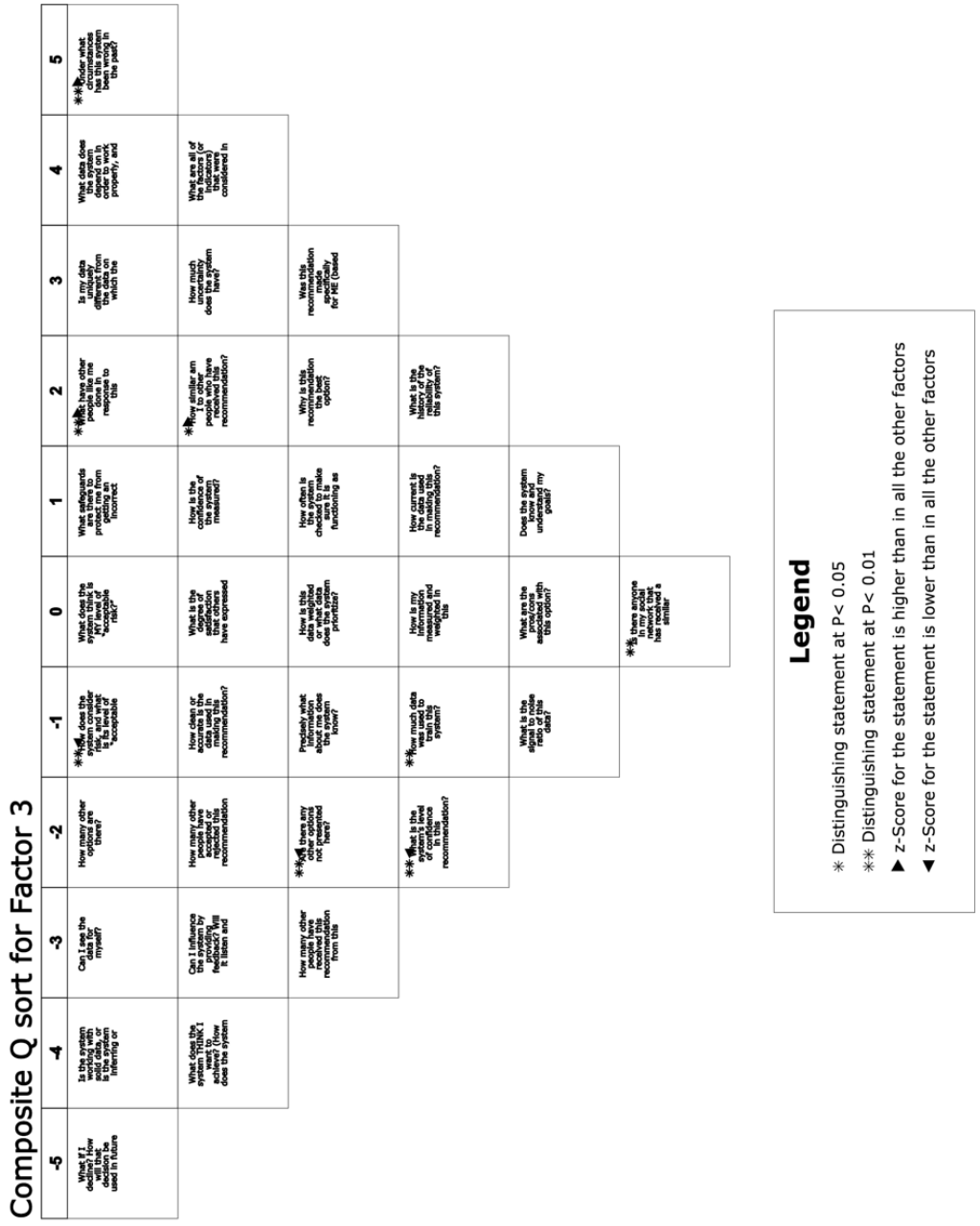

Figure 11: Composite sort for Socially Influenced group

willing to act on a recommendation. But factor group three places a significantly higher value on information about what others have done, indicating that they may use social heuristics in decision making more than others.

Observing what questions were least important to factor group three also helped to interpret these findings. Participants in this group appeared 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 open-ended explanations of why the above questions were least important to them helped to clarify the interpretation of their sorting. One participant remarked, “It is not a wise decision to go over a huge dataset to understand just one recommendation" Another said, “I am old and do not have the energy or skill to go over all the data" Similarly, another participant expressed, “I’m not so much worried about the data behind a recommendation, but more so the reasoning", and another said, “I don’t care how the system makes its choice; I want to know the reliability of the output."

Factor group three appeared to have a distinct preference for justifications that are supported in part by socially-related information, such as how frequently others have accepted or rejected similar recommendations. This preference was contrasted with what seemed to be a general distaste for technical information about the underlying data used

in making recommendations, as clearly expressed in the comments above. These interpretations of data led to factor group three being named, “Socially Influenced.”

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 factor group four ranked their most important question as “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, $Z = .98$, $p < 0.01$). These rankings indicated that factor group four appear to be most interested in understanding how recommendations relate to themselves, and others like them. Factor Group 4’s composite sort is available in figure 12 below.

Open-ended comments helped provide a better feel for how factor group four prioritized their questions. For example, one participant commented that “I want to

understand how the system incorporates my goals. By asking this I can compare it to what I know my goals are, then make sure it is really recommending for me." Another said, "Because I don't want to be listening to a model that works for the benefit of the corporate world. I want to be sure that I receive recommendations of MY interest, and not of my friends' or some corporate office." Others echoed this sentiment, for example, "The basis for the recommendation is most important, and that starts with what information is known about ME!" These comments appear to support the perception that this group seems very strongly motivated to understand how systems and their outputs relate to themselves. It is important to note that many of the questions ranked highest by this group were ranked lowest by the other three groups. The potential significance of this finding will be discussed later in the sections on consensus and disagreement.

Participants in factor group four ranked their least important question as "What are the pros/cons associated with this option?" (composite score -5, $Z = -1.99$, $p < 0.01$). Next least important was "how does the system consider risk, and what is its level of acceptable risk?" (composite score -4, $Z = -1.63$, $p < 0.01$), followed by "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, $Z = -1.17$, $p < 0.01$). Participants in this group definitely appeared not to care much for details about other options, or how the system considers the concept of risk.

Figure 12: Composite sort for Egocentric group

To examine this perception, I analyzed open-ended comments. Participants in this group explained “Most of the cases where I would feel comfortable using it do not involve much risk, so I don’t need to know this." Regarding the devaluing of options, one participant explained, “I don’t really need the number of other options. One or two is enough." And perhaps most informative, one participant commented, “What does it matter if I already don’t trust it? I would need a lot before these questions would even be relevant to me." These comments seem to characterize participants in factor group four as having a strong preference for information that relates to themselves, and contrasts that with a general devaluing for more details about system functions, ranking of options, or expressions of uncertainty or risk. This prioritization of user-centric information makes sense in contexts where personalization is most important to recommendations, such as movies, music, and shopping. This pattern perhaps underlies a viewpoint of intelligent recommender systems as they are today- primarily features of convenience in low-risk domains- as opposed to what they soon could be. Another potential interpretation of these findings might be that this group was comprised of mostly younger participants, which could indicate a generally positive and trusting attitude towards technologies such as artificial intelligence, such that the details or inner workings become less important in favor of information that directly reflects the needs and wants of the user.

With these perceptions in mind, because of the apparent preference for information related to the user-centered viewpoint, I named this group “Egocentric.”

Discussion

I have described the initial analysis and interpretation of the four factor groups, outlining the strategies used in naming them. Further analyses used in interpreting Q-

Methodology studies can be virtually limitless, and are most commonly chosen based on the needs and purpose of the study itself. For the purposes of this study, which was to develop a detailed user typology of user knowledge goals, I used two secondary analysis techniques.

The first was to explore the data for questions that provided a high degree of consensus. Analyzing questions in the Q-set that were equally valued or devalued across all groups can immediately help prioritize the findings from study one, informing designers of explanations that are highly likely to be valued across all potential users, and conversely, those that are very likely to be considered meaningless.

The second technique was to look at questions that produced a high degree of disagreement between groups. Analyzing questions that had high variance in where they were placed by each group highlights questions that may require additional consideration and planning when incorporating them into interface designs. The questions that produced the greatest amount of variance will therefore indicate explanation vectors that are potentially polarizing, and may require adaptive and personalized approaches in order to satisfy end-users' need for information.

Consensus amongst groups

Consensus questions are those that do not distinguish between ANY pair of factor groups, and can be useful in determining what categories of information all users potentially see as valuable.

Comparing the relative rankings of all questions, the question "What are all of the factors (or indicators) that were considered in this recommendation, and how are they weighted?" was considered relatively important (average score 3.75, Z score variance

0.06) by all groups. This is not surprising, given that other studies have confirmed most individuals demand at least some degree of explanation and justification for system outputs (Biran & Cotton, 2017; Lim et al., 2009). Participants in this study also 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 found amongst participants in this study, these two questions were agreed upon by all as having at least moderate importance for users of intelligent systems that make recommendations. Questions of this sort should therefore be considered a high priority for interface designs that seek to incorporate explanations as a strategy to improving system transparency.

Contrary to the above questions, 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 .109) as being very important or valuable to them. These questions represent an extremely granular level of explanation, and are likely more important to programmers, who may appreciate this granularity of information about the underlying data, but they are unlikely to be meaningful to end users, or to improve trust or acceptance for most. Providing explanations about system processes that approach this level of granularity, therefore, is not likely to achieve the desired effects of improving system transparency, and should be avoided for front end systems designed for lay users in mind.

Disagreement amongst groups

In contrast to the analysis of consensus, this analysis examined questions that produced that greatest disagreement between groups. These polarizing questions can help identify potential design elements that may be points of contention to some users. This section first describes individual questions with high variance across all groups. Following this, I describe analysis of disagreement amongst explanation vectors. Recall from earlier that explanation vectors are categories of explanations. In this study, each explanation vector was represented by five to six individual questions. Therefore, by arranging all questions into their explanation vector categories, it is possible to examine disagreement amongst the factor groups.

Disagreement by question

The question with the highest Z score variance (2.078) between all groups 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.)?" This question is the most important question to the Egocentrics (composite score 5), and was the second most important question to the Socially Influenced (composite score 4). The Interested and Independent group, however, found it only moderately important (composite score 2), while the Cautious and Reluctant group thought it was decidedly unimportant to them (composite score -3). Similarly, the question with the next highest Z score variance (1.894) was "Is there anyone in my social network that has received a similar recommendation?", which was ranked moderately high by the Egocentrics and Socially Influenced (composite scores 3, 1), but was the

lowest ranked question by both the Interested and Independent and Cautious and Reluctant groups (composite scores -5).

Both of these questions are somewhat related as they both have a social component to them. Interested and Independents and Cautious and Reluctants both considered these questions to be of relatively little importance, preferring instead answers to questions about the qualities of data, justification of recommendation options, and more information about the inner workings of the system's algorithms. Conversely, the Socially Influenced and Egocentric groups thought these questions were very important to them, which were in line with other socially motivated questions which both groups ranked highly (e.g., "What have other people like me done in response to this recommendation?".

Using socially-related information to aid user decision making is a common practice in many intelligent systems, most notably automated collaborative filtering systems that are featured in many online commerce websites like Amazon.com, as well as media services that make recommendations such as Netflix and Spotify (Aliannejadi, Rafailidis, & Crestani, 2018; Herlocker et al., 2000; Marquez, Cummings, Roy, Kunda, & Newman, 2012; Swearingen & Sinha, 2001). This kind of grouping of individuals based on similar likes, dislikes, and behaviors is also central to most personalization features found on social media, such as Facebook's news feed (Bernstein et al., 2013; Yuji, 2017). It appears, according to this data, that providing details about how the user is modeled by the system, including what data is known about them, and how that data is used to derive characterizations, may be of some value to users who identify with either the Socially Influenced or Egocentric groups, while other users such as Interested and

Independents and Cautious and Reluctant may not find that data useful to them, and may instead consider it a nuisance. These findings suggest, therefore, that the decision to provide these kinds of data is likely to be highly context-dependent, and should be carefully considered and weighed against other priorities such as those discussed in the Consensus section.

Disagreement by Explanation Vector Category

System Parameters and Logic

Questions whose explanations describe the inner workings of a system, including its reasoning, logic, policies, and limitations, fell into the System Parameters & Logic explanation vector. 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 that they represent medium to low priority as design elements that are perhaps best delivered through menu options that can be accessed by those most interested. This finding is of potential interest because much of today's research into intelligibility and interpretability seeks to provide information of this type of end-users. That the participants in this study found these types of questions as relatively unimportant suggests that explanation interfaces may want to instead focus on other explanation vectors, such as those below, in order to improve user perceptions of transparency.

Qualities of Data

Overall, questions pertaining to the qualities of data, such as the age and provenance of data, 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 were especially interested.

User Personalization

Questions aimed at helping users understand how their personal data is known, collected, and used by the system to derive recommendations fall into the User

Personalization explanation vector. This category generated a wider range of sentiment (average Z score .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 .51),” and “precisely what about me does the system know? (average score .5, Z score variance .59),” suggests new efforts should be made towards affording users information about how their data is collected and used. Considering the strong prioritization of these questions by the Socially Influenced and Egocentrics, it is strongly suggested that 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, I have provided a screenshot of the Q-Concierge system developed for this research in Figure X. This proof of concept demonstrates one technique which many users may find useful, and others, such as those aligned with the Socially Influenced or Egocentric groups, may soon demand.

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 crowdsourcing becomes a more common feature in several domains. There are already several notable examples, such as citizen science (Thakur, Sparks, Li, Stewart, & Urban, 2016), personal wellness (Agapie et al., 2018), and even app design (Huang, Chang, & Bigham, 2018) 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 crowdsourcing, 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 techno centric generation assumes more of the user base.

Justification of Options

Closely related to explanations, justifications offer assertions about reasons for decisions or choices, examples, alternatives that are eliminated, or counterfactuals (Biran & Cotton, 2017). All factor groups in this 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 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 granular 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 (Sarter & Schroeder, 2001).

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 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 this 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 designers could possibly address these challenges, I have provided figure X, 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.

Design Implications

Of the 36 questions in the Q-set used for this study, virtually all were considered at least moderately important to one factor group or another, and (as demonstrated 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 would not be especially relevant to anyone. The user typology developed in this study presents a more practical solution.

Systems that can adapt to user preference represent a significant challenge, but also promise significant benefits. If such systems existed, then once a particular user's factor group was determined, the adaptive interface could prioritize information that is likely to be important to them. For example, an explanation and justification of options 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 13. Similarly, the Cautious and Reluctant group would likely 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 14.

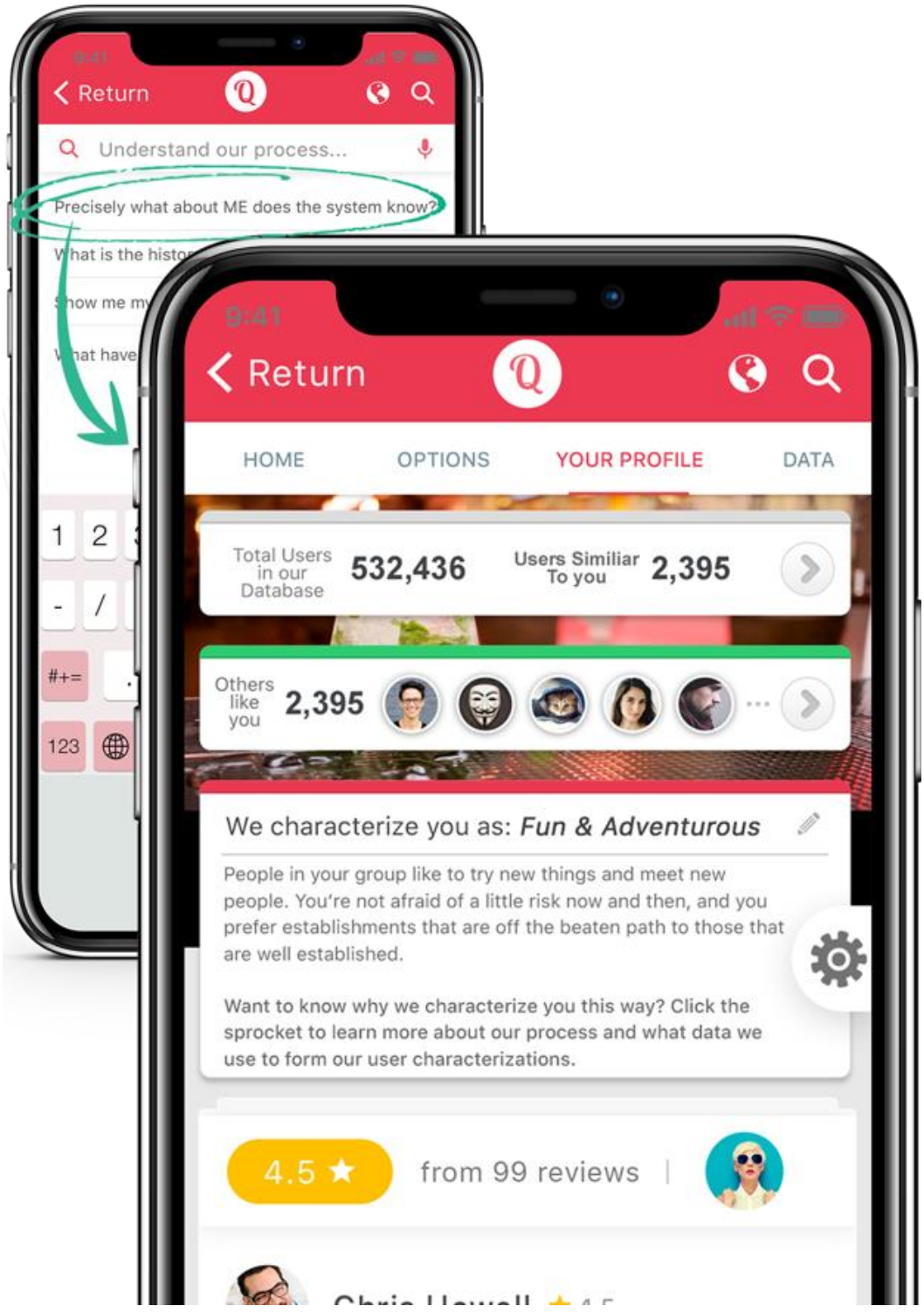

Figure 13: A mockup of the Q-Concierge system. This mockup demonstrates how social influence cues can be useful in providing transparency. This type of information is more likely to be considered valuable by individuals in the Socially Influenced group than by individuals in other factor groups.

These findings provide both immediate value towards design guidance, as well as areas that need further study. For immediate benefits, considering that User Personalization and Options explanation vectors yielded non-significant scores does not indicate that displaying or providing access to these categories of information is not meaningful or useful to users, rather that participants in this study found them equally important by the majority of all participants. This indicates that designers may consider ways they can incorporate these kinds of data into interfaces as a means of improving user trust and comprehension.

The finding that the Qualities of Data explanation vector ranked low was unexpected. It was theorized that this kind of information would be considered highly valuable before commencing the study, based in part on the prominence of this kind of explanation category in much of today's research into interpretability and explainability. Upon closer examination of the data, it appears that the granular level of information provided through the Qualities of Data explanation vector (e.g., fuzziness, provenance, etc.) reduces its potential value to most lay users. For these individuals, it is likely that more explicit methods of enhancing system transparency may be more appropriate, such information about how the user is modeled, or how their data is collected and used to derive recommendations.

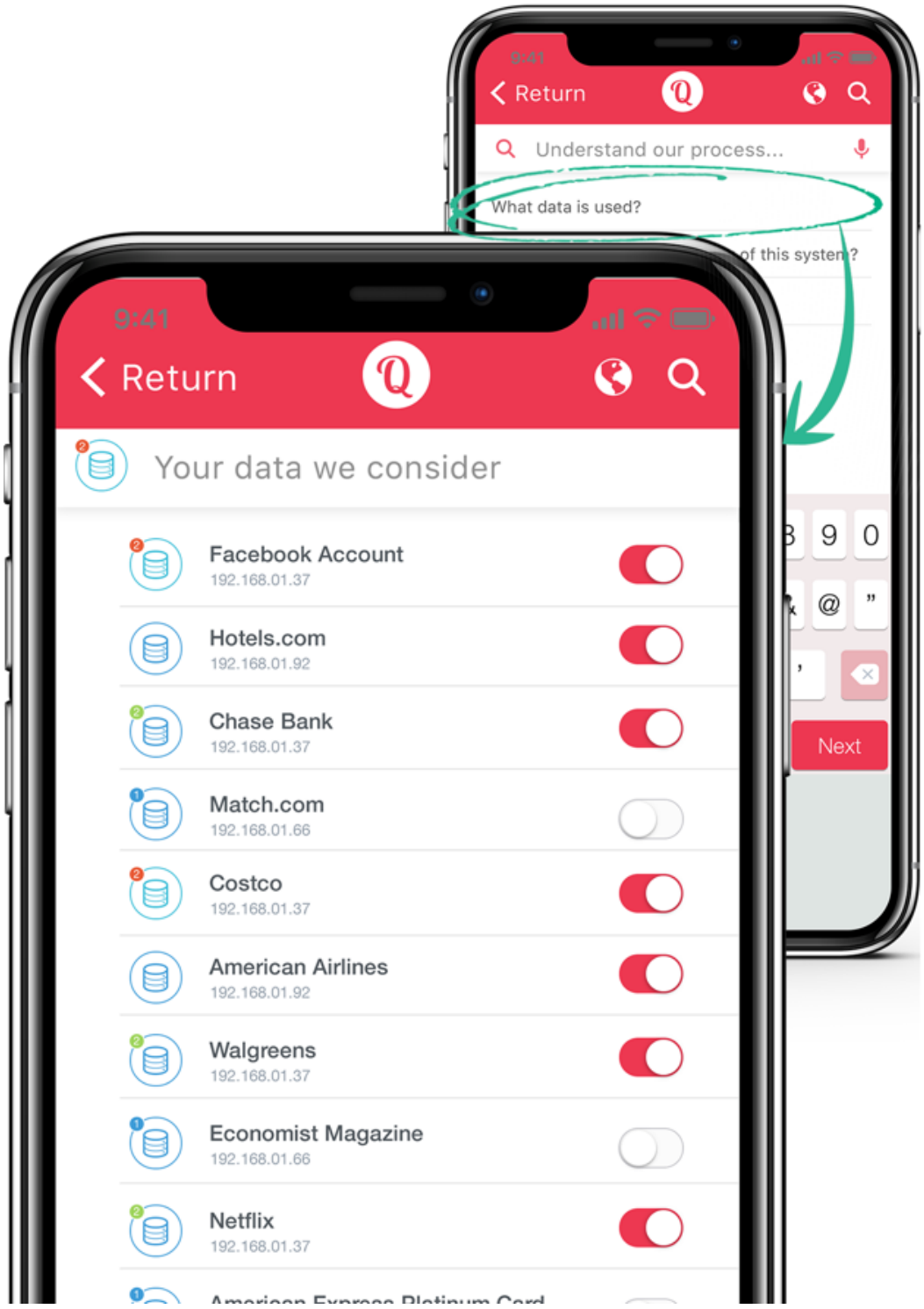

Figure 14: A second mockup of the Q-Concierge system. This mockup demonstrates how advanced controls over input data can be useful in providing transparency to end-users. Data from this study suggest that individuals in the Cautious and Reluctant group would benefit most from this type of approach to transparency.

Examining the role of social media and socially-related information as a means to increase transparency in recommender systems appears to play a polarizing role. Social-related information created high degrees of disagreement between participants in this study, indicating that it may be highly valued by some, while not as valuable to others. The value of social-related information has already been clearly demonstrated across a wide variety of recommender systems, and is most commonly applied in automated collaborative filtering systems (Herlocker, et al., 2000). It is possible that social media-related information, in the context of the vignettes used for this study, which describe interaction paradigms of recommender systems that are not yet commonplace, may not have seemed relevant to some participants.

Still, it is potentially worth noting that the individuals who ranked social media information as potentially valuable to them were all between the ages of 20-29, and had moderate-to-high levels of expertise in computer science, including recommender systems and artificial intelligence. This suggests that a younger population, raised in a data-driven, technology-centered landscape, may find information pertaining to social media a potentially valuable resource to help them understand and interact with recommender systems in the future.

Limitations

There are clear limitations inherent in the approach adopted in this study. The vignettes used to elicit user opinions were developed in study one using a technique known as design fiction. These vignettes intentionally introduced ambiguity and uncertainty in an overtly jarring manner, thus making the decision scenario more difficult for the users, and prompting them to seek additional information. The fictitious systems

developed for this study do not represent real world system in place today, though their development is currently underway. When these systems do arrive on the market, they will unlikely produce such jarring and unexpected results to their users. This may have an effect on user priorities and decision making, which could limit the reproducibility of these findings.

The participants in this study were split approximately 70-30 on expertise and knowledge of intelligent systems. There are no available census data to suggest what a representative sample would look like in terms of expertise. The inclusion of experts in the field of AI may have had an influence in the formation of the user typology because the focus of this study is on end-users who likely know little about the inner workings of intelligent systems.

Finally, the user typology was built by combining each participant's arrangement of cards across all five vignettes, thus averaging their responses to five different decision scenarios and contexts. This was done intentionally to obtain a broad range of attitudes and opinions towards intelligent systems. Some may argue that this averaging causes loss to the meaning of participant responses, potentially eliminating any nuance that may have occurred as a function of the decision context (i.e., each vignette). It would be useful, therefore, to revisit these data and analyze them by each vignette, potentially in order to create a user typology that is specific to each decision context. Doing so would enable a better understanding of the potential range of viewpoints in relation to different contexts of use. This is planned for future activities.

Conclusions of Study Two

This study sought to prioritize potential design features for enhancing the intelligibility and transparency of intelligent systems by developing a detailed user typology of knowledge goals. This user typology describes the ways in which different individuals approach AI-based recommendation systems, and what information they consider more important in order to help them trust and determine to what extent they are comfortable using such systems. By comparing and contrasting these features, this typology provides a lens through which designers might determine efficient and valuable methods to providing information to end users across a broad spectrum of context.

This study also explored in finer detail the four independent groups of user preference related to understanding system functions and behaviors. These analyses indicate that 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.

CHAPTER SIX: DEVELOPMENT AND TESTING OF THE SYSTEM TRANSPARENCY EVALUATION METHOD (STEv)

Background and Motivation for Study Three

Providing a meaningful and impactful explanation to someone is inherently difficult because the perspectives of others are influenced, in part, by our own. Having knowledge of something inherently biases a person's perspective that others possess that same knowledge. Psychologists call this the "curse of knowledge" (Ross & Ward, 2018), and it has a considerable effect on communication because the sender often presupposes the receiver knows or understands more than they do. Thus, the effect of explaining something to someone may be reduced because it is incomplete, or leaves out details that are essential for understanding.

This overestimation in what others know is caused by what is known as "egocentrism," or the difficulty someone has imagining another's perspective, instead favoring their own (Keil, 2006). The effects of egocentrism in design work are manifested by latent assumptions that end users will sense and interpret design artifacts as they were intended, and thus the person's interaction with the system will function accordingly. This ideal interaction is rarely achieved, however, because what is intuitive to the designer of a system may not be intuitive to the end user, thanks in part to the effects of egocentrism.

HCI evaluation methods are intended to detect and address these biases by presenting prototype designs to potential users, and asking for their perspective and feedback. This allows developers to test assumptions, and iterate designs based on the perspective of users, rather than the perspective of the people behind those designs. An

evaluation method designed to measure and improve system transparency, therefore, would need also to assume the perspective of the end-user, because it is their perspective that ultimately determines a product's fitness for use and likelihood to achieve widespread acceptance. There are a number of different approaches to HCI evaluations, however, all of which need to be evaluated before designing a transparency-focused evaluation method could be accomplished. These issues are discussed in the following sections, which outline the theoretical developmental steps taken towards creating the STEv.

Overview of HCI Evaluation Methods

Expert versus User-Centered Evaluation Methods

Evaluation methods in HCI can be broken into two principal groups: expert evaluation methods, and user participation methods. Expert evaluations make use of people with expertise in a given domain, and rely on their knowledge and judgement to inform design. These methods include techniques such as cognitive walkthroughs (Preece et al., 2015) and heuristic evaluations (Dix et al., 2004), in which an expert describes their work or task requirements, which are mapped onto an interface that is designed to assist them with their task. Other expert evaluations are model-based, such as the use of the Goals, Operators, Methods, and Selection (GOMS) model to identify trouble areas in interface design, and eliminate unnecessary or confusing features that hinder general usability. These evaluation techniques work best in domains with well-established patterns of interaction, and users with standards of interaction (i.e., training dictates how tools are used).

For non-expert domains, or domains where technologies are less defined by specific use cases, user participant methods are more appropriate. These often involve observational techniques, where evaluators observe users as they interact with a system and attempt to infer where trouble occurs. This can be accomplished passively as well using software that records mouse-clicks and patterns of interaction, or even might use eye-tracking to create heat maps of focus. These can be blended with think aloud techniques, which ask users to describe what they are doing and verbalize intentions. These techniques are limited to how well users can articulate their issues and concerns, however. Another useful technique is the query methods. Query methods rely on asking the user about the interface directly, and can be useful in eliciting detail of the user's view of a system.

Objective and Functionally-Grounded Measures

Since transparency is both a multi-dimensional and relative construct, it follows that any attempt to measure transparency must also account for these complexities. Attempts have been made to develop accurate and valid measures of transparency in various domains. Their approaches can be broken down into two primary categories: objective measures, and subjective measures, which can be further broken down into application-grounded or functionally-grounded techniques.

Objective measures of transparency involve the quantification of indicators thought to be associated or related to the concept. Once identified, these metrics are measured, and the results are aggregated into an index, providing an overall 'transparency score' for whatever the scale's intended target. For example, da Cruz, et al., (da Cruz et al., 2015) created such an index of transparency in the local government of Portugal using

76 metrics arranged in seven dimensions of transparency (Organizational information, plans and planning, local tax regulations, relationship with citizens as customers, public procurement, economic and financial transparency, and urban planning and land use management). Hollyer, et. al., (2014) similarly created a measure of transparency by examining 172 indicators of country development, which was used to accurately predict a wide range of international governmental outcomes.

Transparency in sociotechnical systems is often measured objectively using performance-based measures. These methods attempt to evaluate a system's transparency by measuring its effect on a person's ability to perform some function, such as identify system errors (Mercado et al., 2016). The effects of modifying elements of an interface such as providing more lines of code, or programming the system to provide uncertainty information alongside a recommendation are then measured using human performance metrics. The assumption is that performance (the dependent variable) is a direct reflection of the changes to the system (i.e., greater transparency, the independent variable).

Objectively measuring transparency in computer systems can be accomplished in much the same way. For example, Owotoki, et al., (2007) created a measure of transparency based on a ratio of three indicators, (1) comprehensible rules and explanations, (2) a confidence measure provided by the model, and (3) visualizations of the relationships between variables. The resulting "transparency quotient" reports the overall level of transparency in a system, 0 being completely opaque, and 1 being completely transparent. The above is a good example of a functionally-grounded measure of transparency, in that it is designed to measure transparency in an abstract,

mathematical model. Good results on these functional models, therefore, are assumed to translate well into good results of interaction.

Much research in the computer science literature on transparency and interpretability falls into the functionally-grounded category. Mathematical approaches to quantifying transparency such as this, while important in many respects, unfortunately, have their limitations in HCI. As discussed earlier, a common goal of improving transparency is to enhance a user's experience by improving their understanding of what systems are doing and how they are doing it. Creating indices based on indicators such as model accuracy or confidence levels may not actually translate to something meaningful to users.

Another limitation is that while making more information available to users is an important aspect of improving system transparency, as discussed in Chapter Two, the more-information-equals-greater-transparency approach does not hold well because much of the information involved in discussions of system transparency is unintelligible to most users, and in many cases would only serve to clutter or distract users from their goals. This is especially important because research has repeatedly confirmed that most people do not require or want a full accounting of all causal factors when they ask for an explanation (Lombrozo & Vasilyeva, 2017). In most cases, a simple, concise, well-formed explanation that is relevant to the user's underlying knowledge goals will be the most successful.

While there are certain cases where objectively measuring the concept of transparency is a valid and appropriate approach, its role in information systems is more dependent on the end-user's perspective and perception of transparency. In other words,

measuring transparency in AI-based recommender systems is more appropriately accomplished using subjective as opposed to objective measures.

Subjective and Application-Grounded

As has been discussed, two primary limitations appear to affect the development of a widely accepted and useful measure of system transparency. The first is that definitions of system transparency are often stove-piped, and can involve conflicting meanings (both to make invisible, or make visible, depending on the context). The second is that measures of transparency are most commonly developed for backend validation (i.e., functionally-grounded) and are not explicitly designed to enhance user understanding or experience.

Seeking to objectively measure a concept such as transparency by creating indices is inherently challenging because the concept itself pertains more to a person's perceptions than it does to any objective object. A government may be ranked highly on some measure of transparency, but those measures only have merit if that government's citizens agree that they are important. In other words, it is the perception of the users that truly defines a thing as transparent- be it a government, or a computer system. Any conceptualization of transparency in terms of interface and interaction design, therefore, should also consider the user's subjective perception as its base.

Users build mental models through a combination of prior experience and subsequent interactions. These mental models come in the form of narratives, or personal stories that pose hypotheses for how things work, what things are connected to what, and their causal links. When system behaviors conflict with these narratives, users seek

additional information to refine their understanding. This information seeking comes in the form of questions, which are based on underlying knowledge goals.

It could be said, therefore, that a greatly simplified, but wholly appropriate definition of transparency might be:

> Transparency in computer systems is a measure of how easily users can find answers to their questions, and how well those answers satisfy their knowledge goals.

This extends and refines the proposed definition on page 24, reflecting the user-centered knowledge seeking focus that is a prevalent motivation in user interactions with intelligent systems. A measure of the above definition would be necessarily subjective, relative to the user and their knowledge goals, which would both take into account the user's needs, while also enabling application-grounded methods, rather than functionally-grounded.

Application-grounded evaluations involve real systems and use potential users to evaluate them. Since these approaches commonly involve methods such as interviews, ethnographies, and focus-groups, they tend to be time consuming, and the resulting data can be noisy, "messy," and difficult to interpret. Since each system is different, some methods and measurement tools must be adapted or modified in order to be most useful, thus further making application-grounded methods difficult to employ.

To summarize, because system transparency in intelligent systems is primarily concerned with the perspective of the end-user, evaluation techniques that are user-centered, subjective in nature, and grounded in real-world applications hold the most

promising approaches useful in the development of a transparency-specific evaluation method.

Research activities up until this point have focused on exploring and identifying the kinds of information that can impact an end-user's perception of transparency when dealing with AI-based recommender systems. The findings from studies one and two suggest that there are a wide variety of concerns that affect end users, and that identifying solutions to improving transparency in these kinds of systems involves a multidisciplinary approach that accounts for both the decision context, as well as user preference.

Implementing these findings into existing or future designs, however, will take special consideration and care. One roadblock to successfully implementing these findings is that HCI currently lacks evaluation methods or tools that are specifically designed to assess the transparency of AI-based recommendations. This is due, in part, because the term transparency in HCI is often confounded by conflicting definitions. Recall from Chapter two that the term transparency in HCI can sometime mean either to make something invisible or visible to users, depending on the context. So as a design feature or construct, transparency is not well standardized, and thus few methods exist that can assist to measure it.

To address this apparent gap, this study sought to develop an evaluation technique with the intent of being as widely applicable and appropriate to help designers evaluate and improve the transparency of their designs. This technique is known as the System Transparency Evaluation method (STEv), which is introduced in more detail in subsequent sections.

Overview of the System Transparency Evaluation Method (STEv)

The STEv was developed with the above characteristics in mind, borrowing from existing evaluation techniques that embody approaches that are user-centered, subjective, and application grounded. My goal was to develop a method that would allow designers to evaluate the transparency of their designs without imposing on them an arbitrary set of scales or definitions that may or may not apply to their designs. The STEv, therefore, is intentionally very modular in its design, allowing designers to define transparency that is context-specific to their use case, accounting for the nature of their system, its intended users, the degree of risk involved, and other related factors.

The STEv is a user-centered query method. The foundational approach of the STEv is to provide users with questions, and invite them to try to find answers to those questions through a system interface. This is a novel approach to measurement and evaluation of computer systems because many existing evaluation techniques invite users to ask questions. The STEv approach, on the other hand, provides users with questions, and asks the user to find answers to those questions by navigating through the interface themselves.

The questions provided are meant to directly represent knowledge goals of users who want or need additional information (i.e., greater transparency) about how the system functions. These questions can therefore be developed from a variety of sources, such as focus groups or other evaluation sessions, in order that they represent reasonable knowledge goals of common users.

It is important to note that participants in evaluations using the STEv are not expected to have answers to these questions in their memory, rather, the STEv measures

the extent to which that information is (1) available to the user, and (2) how meaningful and useful the information provided actually satisfies what users want.

Using this technique, issues related to transparency can be easily detected. For instance, if users report they are unable to find an answer to a question, then designers can work to make that information available. If users report the information provided to them is accessible but does not meet their expectations or satisfy their knowledge goals, then designers can work to improve that information through choice of words and depth of explanation.

The goals of the user are modeled by the questions, and users in turn report on how easily they could find answers to those questions, and how well those answers satisfy their knowledge goals. This allows designers and programmers to identify what questions they want posed to participants, thus not restricting the method to one definition of transparency, but rather to adapt to whatever is relevant to the context of that system. And because the STEv only requires an interface for users to evaluate, it can be applied to systems at any stage in development, from paper prototype to fully mature product.

The STEv considers several associated metrics when evaluating interface designs. These metrics are described briefly below, followed by a discussion of how the STEv is scored.

Scoring Scheme

The STEv derives scores from how well each question or knowledge goal is addressed, creating a weighted average of the above four associated metrics, (1) a measure of how much effort is required to find an answer, (2) how understandable the

explanation or answer is, (3) how comprehensive or complete is that explanation, and (4) how satisfying users found the explanation. Each question is evaluated on these four metrics, each using a scale of 0-100. 100 represents full or completeness, whereas 0 represents the complete absence of something.

Additionally, because some users may not consider all questions equally important, as discovered in study two, the relevance of each question to each user may not be equal. This poses a challenge in comparing scores across participants who use the STEv, since one person's ranking might reflect problems in transparency, whereas another person's ranking might reflect their opinion of the question itself. In an effort to account for these individual differences at play, the STEv uses an additional metric, which is the amount of personal importance of each question to the user. This metric is gathered using the same 0-100 scale as the others, and can be used in a variety of ways. First, evaluators can use the personal importance metric to stratify participants into groups, allowing for more meaningful comparison of scores within groups of more like-minded individuals who hold similar opinions about the relevance of each question to themselves. The personal importance metric can also be used to weight scores, assigning a larger weight to participants who score questions as significantly important to them, and lower weights to participants who do not care much about the question.

For the purposes of study three, a simple scoring scheme was developed that allowed for testing and evaluation of the STEv. This scoring scheme did not attempt to aggregate each metric score into a participant composite score, but rather considered each metric score individually. Future development of a more sophisticated weighted

composite scoring for the STEv will be discussed in the future steps section, following the conclusion section of study three.

The following section outlines these metrics in more detail, and describes the motivation behind choosing them as evaluation criteria on the STEv.

Effort

Current interaction paradigms span a broad range of devices and platform types. While there are subtle differences in interface designs associated with different operating systems and software architectures, the primary currency exchanged in information retrieval is effort.

Measuring cognitive effort in information retrieval tasks is a common method of evaluating interface design, and is strongly associated with human performance and learning (Kratchounova, Fiore, & Jentsch, 2004). Methods to accomplish this vary, and can include evaluating factors such as the time a user spends on a given task of interest (i.e., time on task) (Leis, Reinerman-Jones, Mercado, Szalma, & Hancock, 2015), and also the use of physiological measures such as eye tracking (Chen & Epps, 2013). These measures are useful in objectively quantifying cognitive effort in relation to task-specific interactions. The STEv, however, is less concerned with objective measures of cognitive effort, and instead focuses on a user's perception of effort in its scoring.

The importance of satisfying a user's needs in terms of effort in information retrieval has been well documented (Wickens, 2014). Users who must spend an excessive amount of time searching for information may decide to give up, which may mean they abandon the system altogether, or instead decide to rely on a less accurate, reductionist heuristic to guide their interactions (Keil, 2006). This has strong negative implications, as

inaccurate mental models and poor or inappropriate understanding of system functions have been routinely blamed for serious negative outcomes involving intelligent systems, such as aviation mishaps (Fisher & Kingma, 2001), and industrial accidents (Comfort & Miller, 2004).

Good design should provide efficiency and ease of access to necessary information (Rogers, 2012). Information that is manifest in a system, but is difficult to find, therefore, will cost more effort. The STEv assess a user's subjective perception of effort by asking users:

"How difficult was it for you to find an answer to this question?"

The user is provided a slider that ranges from zero to 100, with subjective interpretations of Very Easy (score of 0), Moderately Challenging (score of 50), and Extremely Difficult (score of 100).

The resulting variable is denoted Q_{effort}. Because the desired numerical value for effort should be low (reflecting lower effort), but the desired numerical values for all other variables is high (reflecting higher satisfaction, understanding, and completeness), we transform the effort variable by subtracting it from 100, therefore creating its reciprocal. The resulting variable is called "ease," and is denoted Q_{ease}.

$$100 - Q_{effort} = Q_{ease}$$

Motivation

Some users may have more motivation to search than others, and are therefore more likely willing to spend greater effort to retrieve necessary information (Lintern, 2013). Participants in an evaluation may therefore work harder to find answers if they are intrinsically motivated, or less hard if they are less intrinsically motivated. Because the

STEv uses subjective measures to derive its scoring, it is important to account for differences in user motivation.

The STEv accomplishes this by asking users:

"How important is this question to you?"

Users are provided a slider which ranges from zero to 100, with subjective interpretations of Not at all (score of 0), Fairly important (score of 50), and Critical (score of 100).

The resulting variable is denoted $Q_{importance}$. Because this measure may change as questions change, scores are calculated for each question separately (i.e., $Q_{importance\_x}$, $Q_{importance\_y}$, etc., where x and y are separate questions).

Understanding

Understanding reflects learning, in that a user's comprehension of a system's functions, that is, their mental model, directly reflects how well that system has informed the user through interactions with it. This is central to the design concept of Learnability.

Learnability, in the HCI sense, means how well features of an interactive system facilitate a user's (especially a novice user's) understanding of how to use them (Sears & Jacko, 2007). Designs that provide this quality of learnability help people discover new features and capabilities in a system, and improve their mental model, or mental image, of what the system does and can do, and how it works (Streitz, 1988).

Measures of comprehension, therefore, are central to the evaluation of the effectiveness of instruction, in that the extent to which a person understands what they have been taught (or explained) can be considered a direct reflection of the quality of that instruction (or explanation) (Ross & Ward, 2018). Hence, evaluating a user's

understanding of an explanation is a valuable way of establishing how well a design feature (i.e., text in an interface, graphical icons, etc) conveys the information it is intended to impart.

This measure does not attempt to account for the accuracy or validity of a user's understanding. For instance, if a user seeks to understand how weather in a mobile application is predicted, the application could provide a full accounting of the calculation of variables and their associated weights. This would be a comprehensive explanation, but may be too much for most lay users to understand or appreciate. Instead, the mobile application could provide a more parsimonious explanation, using simpler language.

The STEv does not attempt to claim that a user that reports a high level of understanding on an evaluation should therefore have a full, demonstrably comprehensive knowledge of that subject. These objective performance-based measures of understanding are important in the evaluation of interaction design. Our model, however, seeks to evaluate a user's perception of understanding, as that more accurately reflects the qualities of interaction design that add to a user's general sense of how things work and why (Preece et al., 2015).

The STEv evaluates a user's perception of understanding by asking the question:

"How understandable was the answer to you?"

Users are provided a slider that ranges from zero to 100, with subjective interpretations of Not at all (score of 0), Moderately (score of 50), and Completely (score of 100).

The resulting variable is denoted $Q_{understanding}$.

Satisfaction

Methods of evaluating the quality of an explanation range from esoteric mathematical proofs (Achinstein, 1977), to proxies using behavior-based measures such as whether or not a doctor overrides a potential drug interaction alert (Bryant, Fletcher, & Payne, 2014). A simpler, and more general method of evaluating the "goodness" of an explanation is to determine whether or not it has an effect on its audience. In this case, the effect we are referring to is a measure of explanatory value—how well some information satisfies the goals of the audience. A poor explanation, according to this scale, is one that fails to satisfy, leaving the user with a sense of disappointment and unfulfilled goals. A principal component of what constitutes an explanation, therefore, is a measure of satisfaction. In this context, satisfaction is a subjective measure of how well an explanation aids a user in answering their question, thus helping them accomplish their underlying knowledge goals.

The use of satisfaction in a measure of system transparency does not claim to evaluate how well an explanation satisfies some formal criteria, such as a causal chain or logic argument (Eiter & Lukasiewicz, 2006). In this context, determining an explanation vector's level of satisfaction is determined by the user, which is in turn determined, in part, by a combination of their knowledge and goals, and the context of the explanation.

The STEv evaluates a user's perception of understanding by asking the question:

"How satisfying was the answer to you?"

Users are provided a slider that ranges from zero to 100, with subjective interpretations of Not at all (score of 0), Moderately (score of 50), and Completely (score of 100).

The resulting variable is denoted $Q_{satisfaction}$.

Completeness

People do not need to understand a complete causal chain to satisfy their goals for learning or understanding. In fact, research has demonstrated that people prefer simple explanations that "satisfice" to exhaustive ones that provide a full accounting of all causal and associated attributes (Pacer & Lombrozo, 2017). For example, explaining how gravity works could require a treatise of events going all the way back to the Big Bang, but most people can explain to children why a bouncing ball eventually comes to rest (Miller, 2017). The concept of completeness, as it relates to the quality of an explanation in our evaluations, therefore, does not refer to a measure of breadth or how comprehensively an explanation demonstrates causality. Rather, this usage of the term refers to how complete a user perceives an explanation to be. This reflects another commonly observed characteristic of what makes a good explanation: A good explanation, according to this measure, should not leave a person with unanswered questions, or more questions than they had to begin with.

This measure of completeness refers to another dimension of a user's degree of satisfaction with an explanation, and is considered here to be a companion of satisfaction. Designs should succinctly provide the right amount and quality of information, especially when users encounter something unexpected and demand answers.

The STEv measures the completeness of an explanation by asking users the following question:

"How complete did the answer seem to you?"

Users are provided a slider that ranges from zero to 100, with subjective interpretations of Not at all (score of 0), Moderately (score of 50), and Completely (score of 100).

The resulting variable is denoted $Q_{completeness}$.

Introduction to Study Three

With the items developed, the STEv was next implemented into a Qualtrics survey in order that it could be deployed. The goal for Study Three was therefore to pilot test the STEv under real life conditions to test the validity of its approach, begin to iteratively improve its design, and to develop a scoring strategy using pilot data.

Rather than developing a standalone testbed for this study, I decided to use four real world recommender systems as testbeds. These systems were Amazon Prime Video, Netflix, YouTube, and Spotify. These four systems were chosen because they all feature recommendations, and they are widely popular and publicly available. Although these recommender systems do not feature artificial intelligence and are not built on machine learning models, they feature architectures that are nonetheless complex, and so were considered good candidate systems to use for testing the STEv.

Pilot Study Results

In preparation for this pilot study, an internal preliminary analysis of explanation features of all four of these systems was conducted. To conduct this analysis, the research team used three questions that were derived from Study Two, and were found to be equally important to all participants during that study. The questions used for this analysis were:

- How is my personal data collected and used by the system?

- How are personalized recommendations made?
- How can I correct or influence the system when it makes incorrect suggestions for me?

The study team then attempted to find answers to these questions by exploring the main interface, all menu items, and any associated help pages (i.e., frequently asked questions or available knowledge bases) for the web-based, Smart TV (Apple TV, Roku, Google Chromecast), and mobile platforms (Apple iOS and Android) of each of the four systems.

This preliminary analysis revealed that two of these systems (Amazon Prime Video and YouTube) did a fair job at answering the questions. For example, each system either provided direct answers to the questions, or providing information that helped to answer them intuitively (i.e., interface controls allowed to modify recommendations by using a thumbs-up or thumbs-down feature). The other two (Netflix and Spotify) did a comparatively poor job answering those questions. For example, the team found that neither Netflix nor Spotify provided any direct answers to the questions, and the available information was sparse in comparison to Amazon or YouTube.

Based on these findings, I determined that the four systems would make a good testbed for comparison purposes to test the STEv. Because this study was only designed to evaluate the STEv, initial emphasis for study three was to determine whether or not the wording and flow of the STEv was appropriate for a user study, and whether or not the items chosen for scoring would provide the necessary sensitivity to detect the differences between systems that were discovered during preliminary testing.

Methods

Because this study required a large number of participants, I decided to use the online crowdsourcing platform, Amazon Mechanical Turk. Participants were paid $3 to participate. All respondents were required to have access to and be familiar with at least one of the available systems (Amazon Prime Video, Spotify, Netflix, and YouTube).

Each participant was asked to use one of the four systems for their evaluation, thus each participant would evaluate only one system using the STEv. Participants were given instructions for how to interact with the Qualtrics survey (see figure 15 below). Each participant was required to acknowledge that they were not allowed to use external sources (i.e., Wikipedia or other online sources) to locate answers to the questions on the STEv. Hence, only information directly provided through the system's interface (and associated help pages) would be evaluated.

Figure 15: Qualtrics study interface. Each question was presented, after which participants went in search of answers using only the interface and available help systems.

Once instructions were complete, each participant was shown a single question and asked to try to find answers to that question. If they reported they were unable to find an answer to that question, then the survey proceeded to the next question. If they reported they were able to find an answer to that question, then the survey proceeded to ask participants to rate that answer using the item criteria described earlier (how important is this question to you; how much effort was involved in finding the answer; and how understandable, satisfying, and complete was the answer).

Once the participant had rated all of the items for that question, they proceeded to the next question until all three questions had been answered. Once all questions were rated, the survey completed and the study concluded.

Results

329 people responded to the survey invitation. Of those, 98 cases were removed from analysis due to poor quality data, or incomplete survey results. The remaining 231 cases were collected over a 14-day period.

Average age of respondents was 24 years old. 147 were males, 84 were females (all participants freely identified as either male or female). The average time to complete the survey was 9 minutes, as shown in the histogram in Figure 16. There were 99 evaluations of YouTube, 77 evaluations of Amazon, 40 evaluations of Netflix, and 15 evaluations of Spotify. 188 evaluations were conducted using a desktop web interface; 27 were conducted using Android; 10 were conducted using iOS, and 6 were conducted using a TV streaming device, such as Apple TV, Roku, or Chromecast.

Table 4 below reports demographics for each system platform evaluated by device type.

Count of all Systems by type					
DEVICE TYPE	AMAZON	NETFLIX	SPOTIFY	YOUTUBE	TOTAL
Android	10	1	4	12	27
Desktop	64	28	11	85	188
iOS	3	5	0	2	10
Smart TV	0	6	0	0	6
GRAND TOTAL	**77**	**40**	**15**	**99**	**231**

Table 4: Demographics of participants in study three by system and device type

The following sections outline results by individual system. Brief discussions of each system will conclude each subsection, after which an overall discussion section will discuss the results of this pilot study.

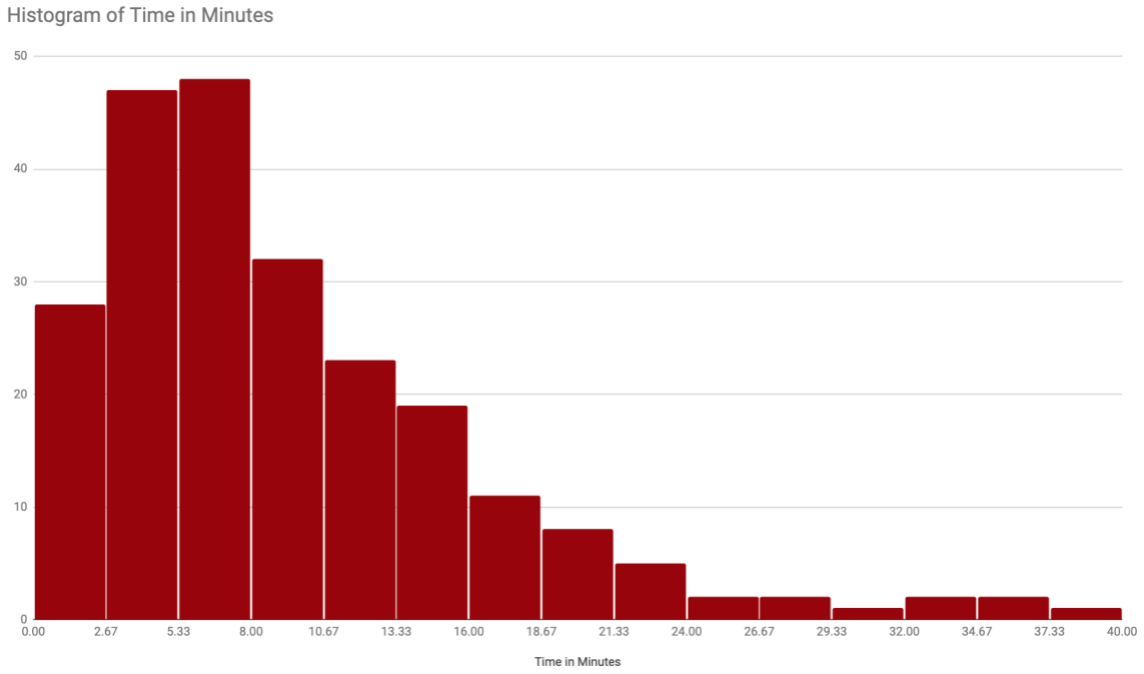

Figure 16: Histogram of time to completion for Study Three

Evaluated System 1: Amazon Prime Video

77 participants chose Amazon Prime Video for their evaluation. Of those, 64 (83.1%) used a desktop web interface, 10 (13%) used a mobile interface on Android OS, and 3 (3.9%) used a mobile interface on iOS. Participants were 64.9% male, 35.1%

female. Average time for completion for all Amazon Prime Video participants was 8.78 minutes.

Amazon Prime Video Question 1: How is my personal data collected and used by the system?

SYSTEM:	**AMAZON PRIME VIDEO**			
Question:	How is my personal data collected and used by the system?			
DEVICE TYPE	Android	Desktop	iOS	Grand Total
Found	7	49	3	59
Not Found	3	15	0	18
Found Ratio	70%	76.56%	100%	**76.62%**

Table 5: Breakdown of Amazon Prime Video participants finding answers to the above question by device type.

An average of 77% of participants were successfully able to locate an answer this this question. Android users demonstrated slightly lower scores than desktop or iOS users, though both Android and iOS participants were notably fewer, which should be considered in this analysis. Complete results are available in table 5 above.

SYSTEM:	**AMAZON PRIME VIDEO**			
Question:	How is my personal data collected and used by the system?			
DEVICE TYPE	Android	Desktop	iOS	Total Average
Average Effort (0-100)	60	50	39	**50**
Average Comprehension (0-100)	79.57	74.87	68.67	**75.13**
Average Satisfaction (0-100)	83.43	66.38	75.00	**68.88**
Average Completeness (0-100)	68.67	79.80	80.67	**81.34**

Table 6: Breakdown of Amazon Prime Video participant responses by qualities of effort, comprehension, satisfaction, and completeness.

Participants reported an average effort of 50/100 to find an answer to this question. This indicates a moderate amount of effort to locate an answer. Android users reported slightly more effort, while iOS users reported slightly less. Participants reported answers found to this question were fairly comprehendible, with an average user rating of 75 out of 100. Average reported user satisfaction was approximately 70 out of 100, indicating most users were satisfied with the answer(s) found to this question. Average user rating for completeness was 81 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete. Complete results are available in table 6 above.

Amazon Prime Video Question 2: How are personalized recommendations made?

SYSTEM:	AMAZON PRIME VIDEO			
Question:	How are personalized recommendations made?			
DEVICE TYPE	Android	Desktop	iOS	Grand Total
Found	10	54	2	66
Not Found	0	10	1	11
Found Ratio	100%	84.38%	66.67%	**85.71%**

Table 7: Breakdown of Amazon Prime Video participants finding answers to the above question by device type.

An average of 86% of participants were successfully able to locate an answer the question "How are personalized recommendations made?" Android users demonstrated slightly lower scores than desktop or iOS users, though both Android and iOS participants were notably fewer, which should be considered in this analysis. Complete results are available in table 7 above.

SYSTEM:	AMAZON PRIME VIDEO			
Question:	How are personalized recommendations made?			
DEVICE TYPE	Android	Desktop	iOS	Total Average
Average Effort (0-100)	65.12	47.01	48.22	**53.45**
Average Comprehension (0-100)	79.57	74.87	68.67	**75.13**
Average Satisfaction (0-100)	76.90	68.43	52.50	**69.25**
Average Completeness (0-100)	88.30	74.62	69.50	**76.57**

Table 8: Breakdown of Amazon Prime Video participant responses by qualities of effort, comprehension, satisfaction, and completeness.

As indicated in table 8 above, participants reported an average effort of 53/100 to find an answer to this question. This indicates a moderate amount of effort to locate an answer. Android users reported slightly more effort, while iOS users reported slightly less. Participants reported answers found to this question were fairly comprehendible,

with an average user rating of 75 out of 100. Average reported user satisfaction was approximately 69 out of 100, indicating most users were satisfied with the answer(s) found to this question. Average user rating for completeness was 77 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

Amazon Prime Video Question 3: How can I correct or influence the system when it makes incorrect suggestions for me?

SYSTEM:	**AMAZON PRIME VIDEO**			
Question:	How can I correct or influence the system when it makes incorrect suggestions for me?			
DEVICE TYPE	Android	Desktop	iOS	Grand Total
Found	8	50	1	59
Not Found	2	11	1	14
Found Ratio	80%	81.96%	50%	**76.71%**

Table 9: Breakdown of Amazon Prime Video participants finding answers to the above question by device type.

An average of 77% of participants were successfully able to locate an answer this question through an interface. Android and Desktop users both reported similar rates of success in finding answers, whereas iOS users reported a lower amount. It should be noted that there were only two iOS users for this question, which skew any results and should be considered in the interpretation of those results. Complete results are available in table 9 above.

SYSTEM:	AMAZON PRIME VIDEO			
Question:	How can I correct or influence the system when it makes incorrect suggestions for me?			
DEVICE TYPE	Android	Desktop	iOS	Total Average
Average Effort (0-100)	57.75	45.84	91.05	**64.88**
Average Comprehension (0-100)	78.87	79.46	73.00	**77.11**
Average Satisfaction (0-100)	82.50	72.64	51.00	**68.71**
Average Completeness (0-100)	86.38	81.30	71.00	**79.56**

Table 10: Breakdown of participant responses by qualities of effort, comprehension, satisfaction, and completeness.

As indicated in table 10 above, participants reported an average effort of 65/100 to find an answer to this question. This indicates a slightly higher amount of effort to locate an answer than the two previous questions. Android users reported slightly less effort, while iOS users reported significantly more (though due to low n these results should be carefully considered). Participants reported answers found to this question were fairly comprehendible, with an average user rating of 77 out of 100. Average reported user satisfaction was approximately 69 out of 100, indicating most users were satisfied with the answer(s) found to this question. Average user rating for completeness was 80 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

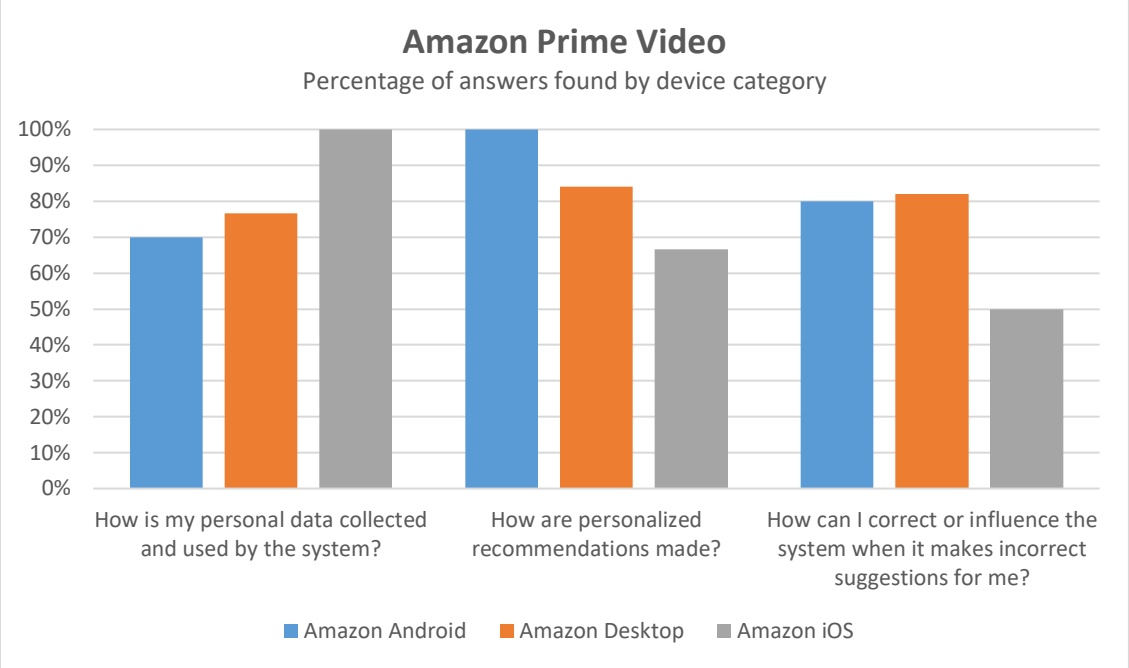

Figure 17: Overview of Amazon Prime Video responses by participants.

Evaluated System 1 Discussion

Participants who used Amazon Prime Video indicated that its overall transparency is relatively high. Across all devices and questions I asked participants to investigate, the majority (88%) were able to find answers that satisfied their curiosity, and were fairly easy to obtain. Figure 17 above highlights the totals for Amazon Prime Video.

Evaluated System 2: Spotify

Spotify Question 1: How is my personal data collected and used by the system?

SYSTEM:	**SPOTIFY**		
Question:	How is my personal data collected and used by the system?		
DEVICE TYPE	Android	Desktop	Grand Total
Found	4	9	13
Not Found	0	2	2
Found Ratio	100%	81.82%	**86.67%**

Table 11: Breakdown of Spotify participants finding answers to the above question by device type.

An average of 87% of participants were successfully able to locate an answer this this question. Android users were most successful, with desktop users reporting just over

20% failure rate. Because only 15 participants used Spotify for their evaluation, meaningful comparisons will be skewed. Complete results are available in table 11 above.

SYSTEM:	**SPOTIFY**		
Question:	How is my personal data collected and used by the system?		
DEVICE TYPE	Android	Desktop	Total Average
Average Effort (0-100)	19.5	41.4	**34.69**
Average Comprehension (0-100)	86	76.67	**79.54**
Average Satisfaction (0-100)	78.5	67.22	**70.69**
Average Completeness (0-100)	80.75	78.56	**79.23**

Table 12: Breakdown of Spotify participant responses by qualities of effort, comprehension, satisfaction, and completeness.

Participants reported an average effort of 35/100 to find an answer to this question, indicating a very low amount of relative effort to locate an answer. Android users reported less effort than Desktop. Participants reported answers found to this question were very comprehendible, with an average user rating of 80 out of 100. As indicated in table 12, average reported user satisfaction was approximately 71 out of 100, indicating most users were satisfied with the answer(s) found to this question. Average user rating for completeness was 80 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

Spotify Question 2: How are personalized recommendations made?

SYSTEM:	**SPOTIFY**		
Question:	How are personalized recommendations made?		
DEVICE TYPE	Android	Desktop	Grand Total
Found	2	6	8
Not Found	2	5	7
Found Ratio	50%	54.54%	**53.33%**

Table 13: Breakdown of Spotify participants finding answers to the above question by device type.

An average of 53% of participants were successfully able to locate an answer this this question, indicating a significant struggle to locate an answer. Half of Android participants were unable to find an answer, while desktop users reported just over 50% success. Complete results are available in table 13 above.

SYSTEM:	**SPOTIFY**		
Question:	How are personalized recommendations made?		
DEVICE TYPE	Android	Desktop	Total Average
Average Effort (0-100)	36	41.83	**40.38**
Average Comprehension (0-100)	87.50	74.16	**77.5**
Average Satisfaction (0-100)	72.50	77	**75.86**
Average Completeness (0-100)	85	84.50	**84.63**

Table 14: Breakdown of Spotify participant responses by qualities of effort, comprehension, satisfaction, and completeness.

Table 14 above outlines detailed findings. Participants reported an average effort of 40/100 to find an answer to this question, indicating a very low amount of relative effort to locate an answer. Android users reported less effort than Desktop, though the difference is negligible. Participants reported answers found to this question were very comprehendible, with an average user rating of 78 out of 100. Average reported user satisfaction was approximately 76 out of 100, indicating most users were satisfied with

the answer(s) found to this question. Average user rating for completeness was 85 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

Spotify Question 3: How can I correct or influence the system when it makes incorrect suggestions for me?

SYSTEM:	**SPOTIFY**		
Question:	How can I correct or influence the system when it makes incorrect suggestions for me?		
DEVICE TYPE	Android	Desktop	Grand Total
Found	1	4	5
Not Found	3	7	10
Found Ratio	25%	37%	**34%**

Table 15: Breakdown of Spotify participants finding answers to the above question by device type.

As seen in table 15 above, an average of 34% of participants were successfully able to locate an answer this this question, which is the lowest success rate of all questions evaluated on all systems during this study.

SYSTEM:	**SPOTIFY**		
Question:	How can I correct or influence the system when it makes incorrect suggestions for me?		
DEVICE TYPE	Android	Desktop	Total Average
Average Effort (0-100)	8	48.75	**40.6**
Average Comprehension (0-100)	84	77.25	**78.6**
Average Satisfaction (0-100)	96	71	**76**
Average Completeness (0-100)	100	92.75	**94.2**

Table 16: Breakdown of Spotify participant responses by qualities of effort, comprehension, satisfaction, and completeness.

Participants reported an average effort of 41/100 to find an answer to this question, indicating a low amount of relative effort to locate an answer. Android users

reported less effort than Desktop, though this data is limited to one participant in the Android group. Table 16 above details that participants reported answers found to this question were very comprehendible, with an average user rating of 79 out of 100. Average reported user satisfaction was approximately 76 out of 100, indicating most users were satisfied with the answer(s) found to this question. Average user rating for completeness was 94 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

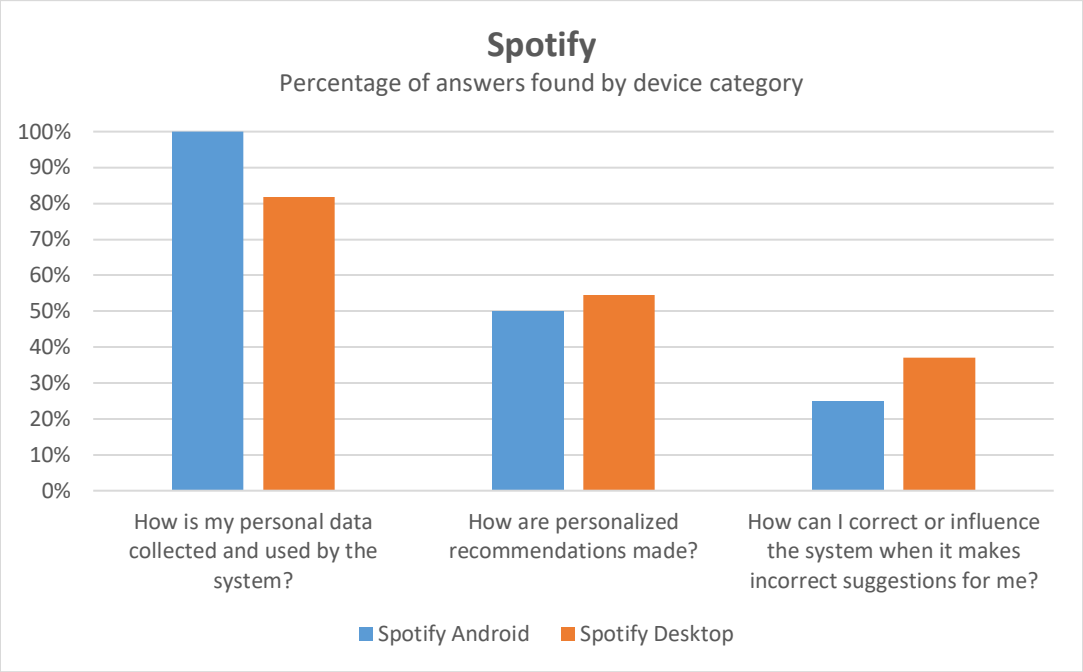

Figure 18: Overview of Spotify responses by participants.

Evaluated System 2 Discussion

Participants who used Spotify for their evaluation indicated a mixed review of Spotify's overall transparency (see figure 18 above). While some variance did exist across devices and questions, the majority (49%) were unable to find answers to their questions, and feedback indicates that those who could find answers were largely dissatisfied with the quality of the answers found.

Evaluated System 3: Netflix

Netflix Question 1: How is my personal data collected and used by the system?

SYSTEM:	**NETFLIX**				
Question:	How is my personal data collected and used by the system?				
DEVICE TYPE	Android	Desktop	iOS	Smart TV	Grand Total
Found	1	26	2	4	33
Not Found	0	3	1	2	6
Found Ratio	100%	89.65%	66.67%	66.67%	**84.61%**

Table 17: Breakdown of Netflix participants finding answers to the above question by device type.

An average of 85% of participants were successfully able to locate an answer this this question. Data is heavily skewed towards the Desktop group, which will affect the meaningfulness and reliability of these analyses. Complete results are available in table 17 above.

SYSTEM:	**NETFLIX**				
Question:	How is my personal data collected and used by the system?				
DEVICE TYPE	Android	Desktop	iOS	Smart TV	Total Average
Average Effort (0-100)	10	31.96	27	25.50	**30.21**
Average Comprehension (0-100)	60	78.23	85	66	**76.60**
Average Satisfaction (0-100)	65	66.38	55	70.50	**66.15**
Average Completeness (0-100)	80	77.11	90	81.75	**78.54**

Table 18: Breakdown of Netflix participant responses by qualities of effort, comprehension, satisfaction, and completeness.

Participants reported an average effort of 30/100 to find an answer to this question, indicating a low amount of relative effort to locate an answer. Android users reported less effort than Desktop, with iOS and Smart TV participants reporting roughly the same levels of effort. Participants reported answers found to this question were very

comprehendible, with an average user rating of 77 out of 100. Average reported user satisfaction was approximately 66 out of 100, indicating most users were satisfied with the answer(s) found to this question. Table 18 above shows that the average user rating for completeness was 79 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

Netflix Question 2: How are personalized recommendations made?

SYSTEM:	**NETFLIX**				
Question:	How are personalized recommendations made?				
DEVICE TYPE	Android	Desktop	iOS	Smart TV	Grand Total
Found	1	23	3	3	30
Not Found	0	3	0	3	6
Found Ratio	100%	88.46%	100%	50%	**83.33%**

Table 19: Breakdown of Netflix participants finding answers to the above question by device type.

An average of 83% of participants were successfully able to locate an answer this this question. Again, data for the Netflix system is heavily skewed towards the Desktop group, which will affect the meaningfulness and reliability of these analyses. Complete results are available in table 19 above.

SYSTEM:	**NETFLIX**				
Question:	How are personalized recommendations made?				
DEVICE TYPE	Android	Desktop	iOS	Smart TV	Total Average
Average Effort (0-100)	25	33.08	48	0	**31.18**
Average Comprehension (0-100)	55	78.38	82	94.33	**79.45**
Average Satisfaction (0-100)	50	67.53	75.33	94	**70.12**
Average Completeness (0-100)	60	76.23	88.33	94	**78.45**

Table 20: Breakdown of Netflix participant responses by qualities of effort, comprehension, satisfaction, and completeness.

Table 20 reports that participants reported an average effort of 31/100 to find an answer to this question, indicating a low amount of relative effort to locate an answer. Android users reported less effort than Desktop, with iOS and Smart TV participants reporting roughly the same levels of effort. Participants reported answers found to this question were very comprehendible, with an average user rating of 79 out of 100. Average reported user satisfaction was approximately 70 out of 100, indicating most users were satisfied with the answer(s) found to this question. Average user rating for completeness was 78 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

Netflix Question 3: How can I correct or influence the system when it makes incorrect suggestions for me?

SYSTEM:	**NETFLIX**				
Question:	How can I correct or influence the system when it makes incorrect suggestions for me?				
DEVICE TYPE	Android	Desktop	iOS	Smart TV	Grand Total
Found	0	14	1	2	17
Not Found	1	15	2	4	22
Found Ratio	0%	48.28%	33%	33%	**43.59%**

Table 21: Breakdown of Netflix participants finding answers to the above question by device type.

As can be seen in table 21, an average of 44% of participants were successfully able to locate an answer this this question, indicating most participants were unable to find an answer. Though the data for the Netflix system is heavily skewed towards the Desktop group, even the desktop group was only able to locate a question 48% of the time.

SYSTEM:	**NETFLIX**				
Question:	How can I correct or influence the system when it makes incorrect suggestions for me?				
DEVICE TYPE	Android	Desktop	iOS	Smart TV	Total Average
Average Effort (0-100)	0	43.07	40	22	**40.41**
Average Comprehension (0-100)	0	75.86	75	68	**74.88**
Average Satisfaction (0-100)	0	63.5	75	63.5	**64.18**
Average Completeness (0-100)	0	73	85	57.5	**71.88**

Table 22: Breakdown of Netflix participant responses by qualities of effort, comprehension, satisfaction, and completeness.

For those participants who were able to locate an answer, they reported an average effort of 40/100 to find an answer to this question, indicating a low amount of

relative effort to locate an answer. Participants reported answers found to this question were very comprehendible, with an average user rating of 75 out of 100. Table 22 highlights that the average reported user satisfaction was approximately 64 out of 100, indicating users able to find a question were only somewhat satisfied with the answer(s). Average user rating for completeness was 72 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

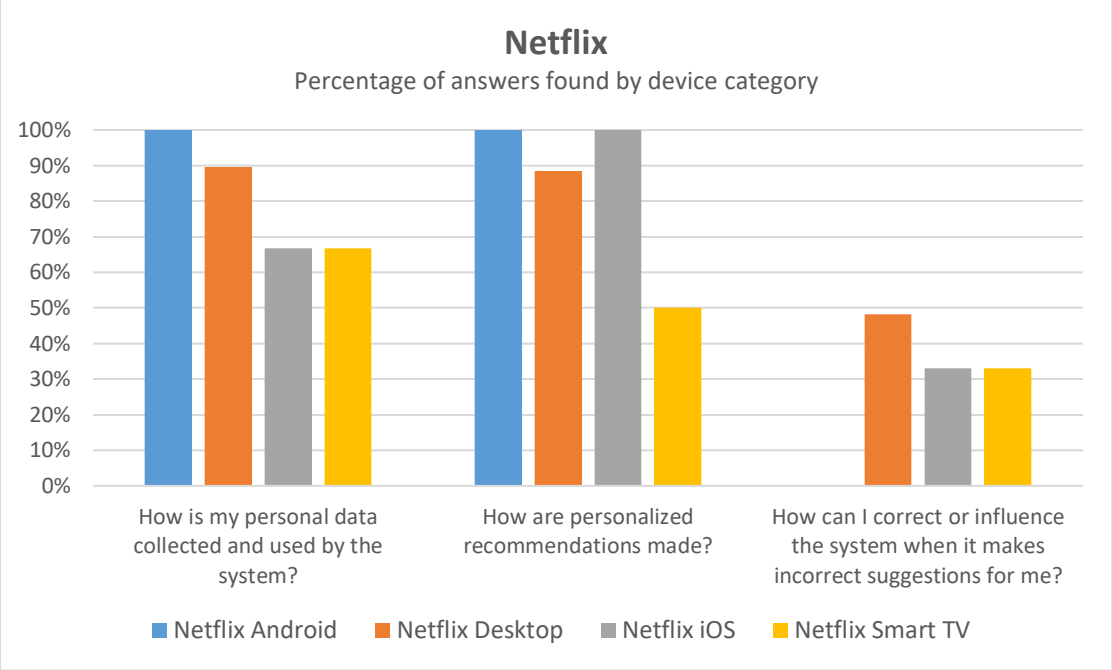

Figure 19: Overview of Netflix responses by participants.

Evaluated System 3 Discussion

Participants who evaluated Netflix reported a mostly positive level of transparency, reporting that most (69%) were able to find answers to their questions regardless of device used, as seen in figure 19 above. Only one question, "How can I correct or influence the system when it makes incorrect suggestions for me?" proved to be challenging. This finding provides reasonable design feedback that can be used to enhance the transparency of the Netflix interface.

Evaluated System 4: YouTube

YouTube Question 1: How is my personal data collected and used by the system?

SYSTEM:	**YOUTUBE**			
Question:	How is my personal data collected and used by the system?			
DEVICE TYPE	Android	Desktop	iOS	Grand Total
Found	9	67	1	77
Not Found	3	18	1	22
Found Ratio	75%	78.82%	50%	**77.78%**

Table 23: Breakdown of YouTube participants finding answers to the above question by device type.

An average of 78% of participants were successfully able to locate an answer this this question, indicating a high number of participants were able to find an answer, though the data for the YouTube system is heavily skewed towards the Desktop group, which should be considered when interpreting these results. Full results are available in table 23.

SYSTEM:	**YOUTUBE**			
Question:	How is my personal data collected and used by the system?			
DEVICE TYPE	Android	Desktop	iOS	Total Average
Average Effort (0-100)	43.78	32.79	64	**34.48**
Average Comprehension (0-100)	70.89	71.18	49	**70.86**
Average Satisfaction (0-100)	68.78	62.21	50	**62.82**
Average Completeness (0-100)	69.11	74.03	92	**73.69**

Table 24: Breakdown of YouTube participant responses by qualities of effort, comprehension, satisfaction, and completeness.

For those participants who were able to locate an answer, they reported an average effort of 35/100 to find an answer to this question, indicating a low amount of

relative effort to locate an answer. Participants reported answers found to this question were very comprehendible, with an average user rating of 70 out of 100. Average reported user satisfaction was approximately 63 out of 100, indicating users able to find a question were only somewhat satisfied with the answer(s). As can be seen in table 24, the average user rating for completeness was 74 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

YouTube Question 2: How are personalized recommendations made?

SYSTEM:	**YOUTUBE**			
Question:	How are personalized recommendations made?			
DEVICE TYPE	Android	Desktop	iOS	Grand Total
Found	11	65	2	78
Not Found	1	20	0	21
Found Ratio	91.67%	76.47%	100%	**78.78%**

Table 25: Breakdown of YouTube participants finding answers to the above question by device type.

An average of 79% of participants were successfully able to locate an answer this this question, indicating a high number of participants were able to find an answer. Below are the results of the various qualities of the answer(s) found. Full results are available in table 25.

SYSTEM:	**YOUTUBE**			
Question:	How are personalized recommendations made?			
DEVICE TYPE	Android	Desktop	iOS	Total Average
Average Effort (0-100)	37.27	42.72	53	**42.22**
Average Comprehension (0-100)	70.82	74.75	75	**74.21**
Average Satisfaction (0-100)	66.82	68.35	58.50	**67.89**
Average Completeness (0-100)	71.64	74.92	56	**73.97**

Table 26: Breakdown of YouTube participant responses by qualities of effort, comprehension, satisfaction, and completeness.

For those participants who were able to locate an answer, they reported an average effort of 42/100 to find an answer to this question, indicating a moderate amount of relative effort to locate an answer, as can be seen in table 26 above. Participants reported answers found to this question were very comprehendible, with an average user rating of 74 out of 100. Average reported user satisfaction was approximately 68 out of 100, indicating users able to find a question were only somewhat satisfied with the answer(s). Average user rating for completeness was 74 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

YouTube Question 3: How can I correct or influence the system when it makes incorrect suggestions for me?

SYSTEM:	**YOUTUBE**			
Question:	How can I correct or influence the system when it makes incorrect suggestions for me?			
DEVICE TYPE	Android	Desktop	iOS	Grand Total
Found	10	57	2	69
Not Found	2	28	0	30
Found Ratio	0%	48.28%	33%	**70%**

Table 27: Breakdown of YouTube participants finding answers to the above question by device type.

Results in table 27 show that an average of 70% of participants were successfully able to locate an answer this this question, indicating a high number of participants were able to find an answer. Below are the results of the various qualities of the answer(s) found.

SYSTEM:	**YOUTUBE**			
Question:	How can I correct or influence the system when it makes incorrect suggestions for me?			
DEVICE TYPE	Android	Desktop	iOS	Total Average
Average Effort (0-100)	54.10	40.46	63	**43.09**
Average Comprehension (0-100)	69.10	80.19	64	**78.12**
Average Satisfaction (0-100)	66.30	68.91	61.50	**68.31**
Average Completeness (0-100)	71.80	73.42	79.50	**73.36**

Table 28: Breakdown of YouTube participant responses by qualities of effort, comprehension, satisfaction, and completeness.

For those participants who were able to locate an answer, they reported an average effort of 43/100 to find an answer to this question, indicating a moderate amount of relative effort to locate an answer. Table 28 shows that participants reported answers

found to this question were very comprehendible, with an average user rating of 78 out of 100. Average reported user satisfaction was approximately 68 out of 100, indicating users able to find a question were only somewhat satisfied with the answer(s). Average user rating for completeness was 73 out of 100, indicating that the majority of participants considered the answer(s) provided to this question to be mostly complete.

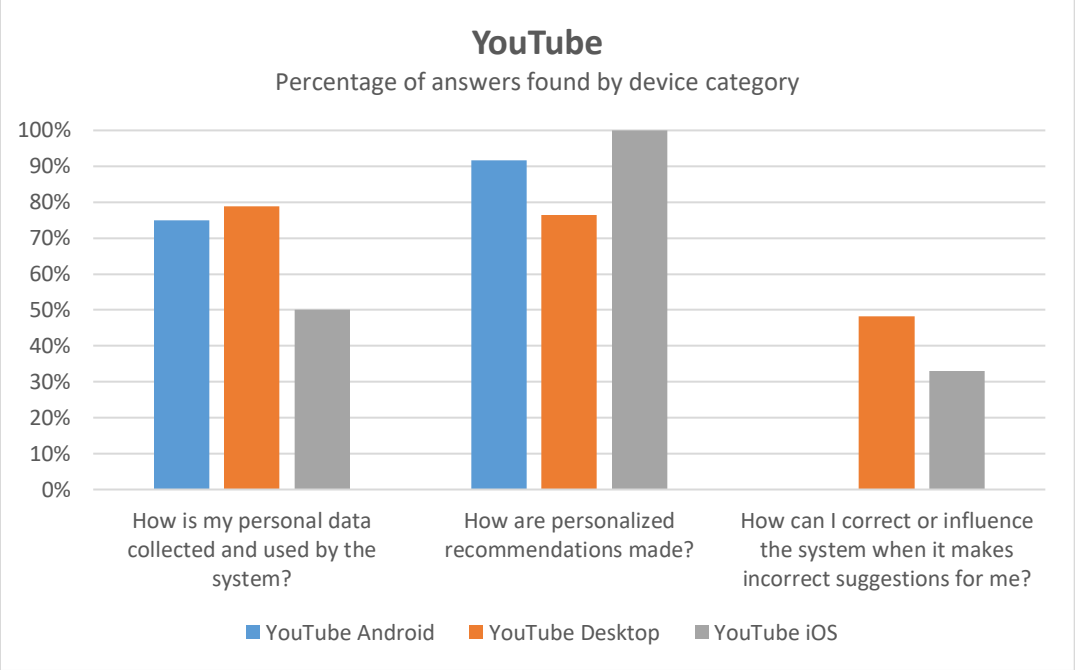

Figure 20: Overview of YouTube responses by participants.

Evaluated System 4 Discussion

As figure 20 shows, participants who used YouTube for their evaluations reported largely positive interactions with the system in terms of its transparency. Most (72%) were able to find answers to their questions, with the exception of the question "How can I correct or influence the system when it makes incorrect suggestions for me?" which was comparatively low across all devices.

Future Developments and Directions for the STEv

Development of a "Transparency Qualities Scale"

The results of my initial testing with the STEv suggests that an application-grounded, subjective evaluation of interface transparency is a viable and informative strategy. While these data were able to illustrate potential problem areas with regards to individual systems, the feedback available was limited. A logical next step in this research is to develop a quantifiable scale that addresses the qualities of transparency, and would enable objective and comparative evaluations between and amongst systems being assessed. Such a scale would facilitate a deeper, more nuanced measure of transparency to be obtained, thus providing more qualitative feedback which designers could use to improve their prototypes.

Combine STEv with Eye Tracking

A similar next step would be to take the STEv and combine it with existing eye tracking software in order to create a suite of transparency-related user evaluations. Bringing eye tracking into the STEv would greatly enhance the STEv's ability to evaluate how users are interpreting existing information, as well as give clues as to where they are expecting information to be (as measured by eye tracking patterns, for instance). There are a variety of existing software platforms that use eye-tracking to help with the design of interfaces and layouts (see, for example, Tobii, https://www.tobiipro.com/fields-of-use/user-experience-interaction/). What would be particularly informative is to discover whether or not eye tracking data could be used to predict whether or not a user would consider a system transparent or not. Beyond the general usability evaluation applications, using eye tracking to augment the STEv would provide an in-depth

evaluation or user information-seeking behaviors, which could then be used to further iterate and improve troublesome designs. This would be particularly important in use cases that involve a higher degree of risk to the user (such as financial decision making), or cases in which human-system trust is imperative.

Implement the STEv as Part of the Design Cycle of a Recommender System

While study three provided solid grounding for the STEv and demonstrated its viability as an HCI evaluation method, implementing the STEv in the development of a recommender system is a critical next step. This is because currently the STEv has only been used to evaluate existing designs, which limits the use of its feedback provided. The vision of the STEv, however, is that it will be used in conjunction with other evaluation techniques as part of an iterative design cycle, thereby influencing the overall design of a recommender system using user feedback. Work has already begun on a preliminary field trial of the STEv through collaborations between DARPA's XAI program and various US Government performers on the project. The focus of these efforts is again to demonstrate the utility of a transparency-focused, user-centered tool in being able to reduce or eliminate user confusion towards improving user acceptance and trust.

CHAPTER SEVEN: SUMMARY OF RESEARCH ACTIVITIES AND CONCLUSIONS

In this final chapter, I revisit the premise of this work, provide a summary of contributions, and discuss future directions for research to further this work. The scientific literature concerning the concept of system transparency is diverse. Both the objectives of these studies and the methods used suggest that transparency is not a binary concept, but rather represents several distinct ideas that must be identified, defined, and operationalized before meaningful progress can be made. Table 29 below highlights the contributions of this dissertation to the greater scientific community, in specific to the testing and evaluation of human-machine interfaces.

This research approached transparency in AI-based recommender systems from the end-user perspective, and sought to understand what potential users of AI-based recommender systems require in order to understand and ultimately act on AI-based recommendations.

Study one asked users to verbalize what questions they would ask if presented with anomalous or unusual recommendations from systems across a wide range of domains. Using a design-fiction approach, I created five descriptive vignettes, characterizing potential future interactions with AI-based recommendations. Utilizing a user-centered design workshop format, participant questions were recorded and coded, and used to develop a taxonomy of user knowledge goals, useful in helping to arrange and categorize different needs and goals of end-users interacting with AI-based recommendations. This taxonomy was then used to develop a framework of potential explanation approaches that could possibly be mapped to each knowledge goal. These

mappings, known as Explanation Vectors, describe separate approaches that system designers may adopt to provide improved transparency of AI-based recommendations to end-users.

Using this explanation vector framework as a foundation, **study two** examined patterns in user interactions with AI-based recommendations, and sought to describe individual differences that may help determine a user's subjective impression of transparency when dealing with AI-based recommendations. Using a novel mixed-method approach known as Q-methodology, these patterns of interaction were analyzed and used to create a detailed user typology of user information needs. This typology describes different ways in which users value different types of transparency information, depending in part on a mix of personality characteristics and posture towards AI-based technologies. The resulting typology can be used to identify user preferences for information, which can then be used by system designers to prioritize explanations.

Borrowing from this data, **study three** combined lessons learned from studies one and two in order to develop and test a new evaluation method for assessing transparency in AI-based recommender systems. The resulting system transparency evaluation method (STEv) was piloted using four real-world recommender systems, and was able to demonstrate efficacy in discriminating between systems with varying levels of transparency pertaining to specific questions of privacy, personalized recommendations, and system tractability. The STEv also demonstrated its ability to provide meaningful design feedback to system designers, validating its approach as viable in both prototype and fielded designs.

This research attempted to push the boundaries of existing research in a number of ways. Firstly, its focus on the needs and goals of end-users is a departure from much of the existing literature on system transparency in artificial intelligence applications, which primarily focus on transparency for developers, rather than lay users. Assessing these needs and goals through the direct input of potential users is also a unique method not commonly used in transparency-related literature, and expanded the concept of system transparency in AI-based recommendations beyond the common tenets of understanding algorithms, to include new dimensions such as the role of social influence, the importance of privacy, and understanding how to modify and adjust personalized recommendations. This research also sought to build on existing assessment techniques, and to contribute a new evaluation method that would be useful across a wide range of transparency-related applications and scenarios. The STEv represents a rapid, agile, and scalable evaluation technique that designers can employ alongside other existing HCI evaluation methods, or can use as a standalone evaluation to improve system transparency.

While the limitations of this research has already been discussed in each chapter, there are several persistent challenges related to system transparency in AI-based recommendations that require further research in order to address.

Persistent Challenges Related to System Transparency

User Algorithmic Literacy

In my research, I specifically chose to explore the role of individual differences in transparency. This decision was informed by many decades of HCI-related work in usability that have attempted to identify, map, and adapt systems to people's preferences

in order to make them more usable (for example, see (Nielsen, 1994)). While the typology of user information needs is a good first step towards understanding how individuals prioritize information needs differently, there is another dimension that may also influence these needs. In this case, people's knowledge of how algorithms work may determine their need for transparency. Studies have shown that people overestimate what robots or AI can do (Keil, 2006). Users of such systems often assign the artificial intelligence much greater capabilities than actually exist. This underlies what could be an overall lack of understanding of the role of algorithms in these systems, and how they work. My Q study attempted to parse this out somewhat by assessing people's level of expertise as it pertained to AI-based recommendations. A dedicated study that conducted a detailed and nuanced assessment that includes people's knowledge of algorithms would be very beneficial, as this variable may be a significant predictor of how important is transparency to each individual.

Trust and Its Proxies

While the arguments made in this dissertation have generally been in favor of more transparency, there are cases where improving transparency may hinder or hurt customer satisfaction, depending on the nature of the domain/task. Too much information about how processes are conducted and how automated decisions are made can sometimes backfire. For instance, Cramer, et al. (2008) conducted a study specifically exploring how providing explanations of how a recommender system made its recommendations. Results showed that explanations of why the recommendation was made improved user's acceptance of those recommendations, but did not improve trust, and actually lowered reported user confidence. Dzindolet, et al, (2003) conducted a

similar experiment showing that when participants were provided with an explanation of how a decision aid worked (indicating its limits and methods that can cause false alarm, i.e., transparency), participants' reported trust and reliance on the decision aid were decreased. The authors point out that this finding is not necessarily negative, as it is an indication that trust in the aid may have in fact been more appropriately calibrated- a common goal of much systems evaluation work involving trust. Nevertheless, there are several situations in which more transparency may negatively affect the usability of a system. Establishing guidelines for when greater transparency is called for, and when it should be tempered will be an important task for future research.

Context-based Transparency

Much, if not most, of the scientific research on transparency in artificial intelligence has focused on the developer- the individuals responsible for building a system. This focus has limited the methods and techniques of improving transparency to a category that only benefits those who are literate in computer programming. If current trends in intelligent systems development continue, however, then future consumers of everyday technology may very soon require transparency that benefits them, e.g., insights into machine reasoning and logic that aid users in understanding how a recommendation was made in order that they can intelligently decide what to do with it. Hence, this dissertation research focused on improving transparency for the end-user, and attempted to develop methods and techniques suitable for lay users with little or no knowledge of computer programming. There is a third class of user, however, which should be considered within the scope of a discussion on the importance of transparency in artificial intelligence. This class of user is the deployer, or the individuals that employ an AI

system. Deployers of said systems will likely face transparency challenges both internal and external, and will need to determine to what extent transparency should be provided to end-users. Concerns over trade secrets, intellectual property rights, human ethics, and privacy are all dimensions that may affect how much transparency deployers of these systems provide. Will HCI provide guidance to deployers? Is there research to support drawing a line with regards to balancing the above concerns with a user's desire for (or in some cases right to) transparency? These questions remain unanswered, and should be addressed in future research.

The discussion on the differences in roles of AI-based recommender systems is central to any discussion on the importance of transparency, largely because these roles can often be in conflict with one another. To give an example of how, consider, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, a decision support tool that predicts criminal recidivism. COMPAS is used by courts and judges across the United States to determine sentencing and parole for accused criminals. Its developer, Northpointe, needs to understand their system and be able to debug or improve it, and be able to sell it to prospective customers. Its deployer, various state and federal district courts, needs to have some understanding of how predictions are reached in order to feel comfortable defending decisions made on those predictions. Its user, various state and federal judges, need to have some understanding of the limits of the systems, and be able to determine when the system is functioning properly or is out of bounds in order to feel comfortable acting on its recommendations. Its beneficiary, accused criminals (and their legal counsel, along with potential watchdog groups interested in justice reform), need some explanation of how a prediction was reached, and

possibly be able to determine whether or not that recommendation was biased in some way.

Transparency information, in this case, is very much determined by the audience(s), which may have conflicting needs and goals in mind in seeking this information. Determining, therefore, how best to provide this information, how much to provide, and to what degree it should be made available are questions that require extensive research before they can be answered.

Final Conclusion

The utility of AI-based recommender systems built upon machine learning platforms is evident, but adoption is hindered by an inability to provide information that help users understand the system's reasoning strategy. Increasing the transparency of these systems should improve trust and ensure appropriate reliance in most cases. This research has attempted to address some of the needs that end-users have with regards to transparency in AI-based recommender systems, and has provided a template to evaluate systems on the dimension of transparency that is designed to complement existing HCI evaluation techniques.

Contributions of Studies

Development	Contribution	Ch.#
Overall		
Integrated definition of system transparency	Transparency in computer systems is multi-dimensional and contextually-dependent, but core components allow for an integrated definition	2
Conceptual model of system transparency	Through a systematic literature review of research on transparency in 12 domains, a conceptual model helps define a term that has historically been treated ambiguously	2
Study One		
Interaction vignettes based on design fiction	Using design fiction to create interaction vignettes that can be used to elicit future design requirements	4
Taxonomy of user knowledge goals	Conceptual taxonomy that assists designers in targeting information to potential information seeking goals of the user	4
Explanation Vector Framework	Multi-dimensional framework that provides channels of information which can be used to provide explanations of system behaviors	4
Study Two		
Q-methodology for design requirements elicitation and prioritization	Further advanced the use of Q-methodology in HCI research; adapted Q-method using a question-based framework	5
Four-Factor user typology of interactions with artificial intelligence recommendations	Identified four distinct viewpoints of interactions with intelligent systems, and elicited information priorities for each; this typology further confirms that information is not homogeneous, and individual differences play a significant role in the usability and transparency of intelligent systems	5
Study Three		
System Transparency Evaluation Method (STEv)	Developed and piloted a theory-based, application-grounded, user-centered method of evaluating transparency in intelligent systems	6

Table 29: Summary of research activities and contributions generated by this dissertation.

APPENDICES

Appendix A: Interactive Vignettes Used in User-Centered Design Workshop

D-SAM

You have had an illustrious career and are nearing retirement. You and your partner are already making plans for what to do next: travel, buying a new home in a quiet village near the beach, and doing lots of shopping and playing golf.

You receive a notice from your financial management firm. They are announcing that all accounts are being moved to a fully-automated financial trading system, D-SAM (deep securities and accounting management). D-SAM is built on a state-of-the-art neural network that processes millions of bits of financial information per second in order to make predictions on growth and future financial opportunities in the marketplace. D-SAM has been shown to outperform humans on both investment strategy and long-term growth, and was recently featured on the cover of a prominent financial magazine that you read.

Two weeks later you receive a notice that D-SAM recommends you move your mutual fund to a different index. You check the recommended index and read perspectives from other sources. None of them suggest to you that moving your money to this index is an obvious decision.

Since yours is a managed fund, you are given five days to either consent or decline this recommendation, after which time D-SAM will consent by default and your accounts will be moved. If you decline D-SAM's recommendation, your accounts will remain where they are.

NEXT GENERATION SOCIAL MEDIA

It has been a productive morning at work. You can hear other people coming and going from their cubicles, which means it must be lunchtime. You decide to eat lunch at your desk instead of going out.

While munching on your favorite sandwich, you open up your favorite social media website to check in on your network of friends. A political advertisement in the lower corner of the screen begins to autoplay.

The ad opens with a controversial statement that represents a fringe view that is opposite of your own. Your computer volume is louder than you expected. The ad continues to play as you try to close it and lower your volume.

Just as you find the sound settings in order to mute the ad, the narrative states a view that you continue to listen to, incredulous that these sorts of views exist.

As you press mute you realize that someone is standing behind you. You swivel in your chair to see your boss quickly moving away from your desk.

Q-CONCIERGE

You and your significant other are engaged. You are meeting your significant other's parents in a neutral town that neither of you have been to before. The task of choosing which restaurant is left up to you.

You decide to test out a new system you heard talked about on the radio, a system that uses artificial intelligence to make suggestions for things like shopping and restaurants, all based on information gathered through your personal social media and internet browsing history.

You connect to the system and provide it with the information it needs- name, social media handles, and an email address. Once you have registered, the system takes you through a brief tutorial and describes how you can use natural language to ask the system anything, and it will give you an accurate recommendation. As part of the tutorial, the system shows that you can ask the system something like, 'what is the best place to get a steak on a rainy afternoon in springtime?' and then it provides an answer, along with a long list of positive comments about that restaurant.

You type in the search bar "what is the best restaurant to go for meeting future in-laws?"

The system recommends a restaurant called "The Kraken," which describes itself as "a modernist pirate-themed decor, serving the very freshest fish and shellfish, or whatever lands on our docks." The restaurant reviews, although mostly positive, are somewhat mixed.

The day before the event, your significant other calls you. They sound nervous, and inform you their parents are actually quite difficult to please, have allergies, are notoriously picky, and are “kind of snobs.”

ONNPAR

Your significant other has just been diagnosed with a very serious malignant cancer, and is given a life expectancy of less than 6 months. Your doctor is recommending an unconventional treatment. She explains that this treatment has proven highly effective, with survivability rates more than double that of other treatments in patients with similar biological profiles diagnosed with these types of cancers. This treatment is not without risks, however.

While many patients showed improvements and lived longer with this treatment, in some rare cases patients suffered more, and in some cases died earlier than patients undergoing more traditional therapies.

The doctor explains that the treatment recommendation came from the oncological neural network prognosis and recommendation system (ONNPAR). ONNPAR is an artificial intelligence system that scans millions of medical records, and is able to make millions of layers of associations between variables- it sees connections that humans often miss, the doctor explains.

During testing, the system outperformed human diagnoses of several different types of diseases. You ask about taking the more traditional treatment. Your doctor informs you that while other treatments are available, you would have to obtain a second opinion, which insurance does not cover.

HR-KIT

For 11 years, you have been the head of the HR department at your company. All hiring decisions are up to you. Since you are always on the lookout for the best talent, you recently requested that the company purchase a license for HR-KIT (Human Resources Key Indicators of Talent). HR-KIT is a machine learning algorithm that processes data from millions of datasets and predicts success in the workplace. It was recently showcased at a conference you attended, and featured by a number of prominent businesses around the world- many of which are competitors with your company. HR-KIT was a major investment for your company, and the decision to purchase it took many months of convincing by you.

A new position recently opened up and you begin to narrow down a list of qualified candidates. The executive board is aware that this hiring decision will be the first test of HR-KIT, and they are eager to hear your report.

After manually sorting, you decide to interview three people. The first two interviews are standard, and both appear to be well-qualified and a good fit for your organization. The third interview, however, is not so smooth, and you conclude that the third candidate is not a good fit for the organization.

You enter all the candidate's information into HR-KIT and wait for the results.

HR-KIT recommends the third candidate.

Appendix B: Question Responses from User-Centered Design Workshop

- On what data is this recommendation made?
- How clean/accurate is the input data?
- How much uncertainty does the system have?
- What is the provenance of data?
- How is the physical environment considered in the recommendation?
- What is the system's goal?
- Does the system's goal match my own?
- How is risk measured?
- What information does the system know ABOUT ME?
- What dependencies are used in these recommendations?
- How aware of *me* is the system?
- Does the system even have a concept of risk?
- Does the computer have a good track record?
- Can I see who else has this kind of thing?
- Have other people done this before? What did they think?
- What will happen if I say yes?
- How much data is this?
- Can I have more information?
- What does the system have on me? (What personal data is the system aware of and considering?)
- How many times does this thing fail?
- How much time do I have to think it over?
- Can I see the data?
- Is there hidden information the computer isn't telling me about?
- Did I do something to make the computer think I wanted this outcome?

- Are these subsystems measured for accuracy/fidelity, or is data from these systems considered infallible?
- Does it know & understand my goals?
- Does it understand my limits?
- Are these options static or dynamic?
- What criteria is used, and how is it weighted (i.e., what is the recipe)?
- What cost/benefit tradeoffs exist?
- How are the options rank ordered?
- What is the signal:noise ratio?
- How aware is the system of the physical operating environment?
- Why is this option the best?
- What is the system's level of confidence?
- What other options were considered?
- What is the ratio of false positives?
- What if I don't want what is presented? Can I change how the computer works?
- What are the odds you're right?
- Where are all the sources of data?
- Show me the data!
- Does every user get the same recommendation?
- What part of my profile does the computer care about most?
- Does this system get my personal data (credit cards, health records, etc.) or is it just data from when I use the system (likes on facebook, etc)?
- Can I block the system from getting my data?
- What are the pros/cons?
- What kind of software is this running?
- What is the model built on?
- What does this system do really well?

- What is the estimated outcome?
- What is the current risk?
- Who else has taken this suggestion?
- When is it wrong?
- How accurate is that system?
- How will my decision affect the system?
- How common is this suggestion?
- How much data has been fed into the system to teach it?
- How have others fared when this suggestion was accepted?
- How can I give feedback?
- How is my feedback incorporated or considered in future recommendations?
- How big is the library of options?
- Does the system have a concept of collateral damage?
- Can I evaluate the data myself?
- Is the data accessible to me?
- Is this what the system was designed for?
- What are the limits of this system?
- When was this thing checked for bugs?
- Can I see user ratings from other people?
- Why wasn't I warned about my data being collected?
- Is there a person behind this system, or is this 100% computer?
- Does the system think I like this kind of thing? If so, why?
- What are the other people like that have gotten this kind of recommendation?
- What was the next best option?
- How does the computer know what I like? How do I know the computer understands what I want?
- Where is this coming from?
- Who else sees this?

- Does the computer know me?
- Why was this even shown?
- What is the meaning of this?
- Why is this relevant?
- How can I see what's behind all of this?
- Is this data any good?
- Is this common?
- Is this unique?
- What links up with this recommendation?

Appendix C: Final Question Bank for Study Two

Category	Statements
Data	How current is the data used in making this recommendation?
Data	How is this data weighted or what data does the system prioritize?
Data	How clean or accurate is the data used in making this recommendation?
Data	Is the system working with solid data, or is the system inferring or making assumptions on 'fuzzy' information?
Data	What are all of the factors (or indicators) that were considered in this recommendation, and how are they weighted?
Data	How much data was used to train this system?
Data	What is the signal to noise ratio of this data?
Data	Can I see the data for myself?
Options	What if I decline? How will that decision be used in future recommendations by this system?
Options	What are the pros/cons associated with this option?
Options	Can I influence the system by providing feedback? Will it listen and consider my input?
Options	Are there any other options not presented here?
Options	How many other options are there?
Options	Why is this recommendation the best option?
Personal	How does the system consider risk, and what is its level of "acceptable risk?"
Personal	What does the system THINK I want to achieve? (How does the system represent my priorities and goals?)
Personal	How is my information measured and weighted in this recommendation?
Personal	Does the system know and understand my goals?
Personal	Is my data uniquely different from the data on which the system has been trained?
Personal	Precisely what information about me does the system know?
Personal	What does the system think is MY level of "acceptable risk?"
Personal	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.)?
Social	How similar am I to other people who have received this recommendation?
Social	What have other people like me done in response to this recommendation?
Social	What is the degree of satisfaction that others have expressed when taking this recommendation?
Social	Is there anyone in my social network that has received a similar recommendation?

Social	How many other people have accepted or rejected this recommendation from this system? (What is the ratio of approve to disapprove?)
Social	How many other people have received this recommendation from this system?
System	What safeguards are there to protect me from getting an incorrect recommendation?
System	Under what circumstances has this system been wrong in the past?
System	What data does the system depend on in order to work properly, and do we know if those dependencies are functioning properly?
System	What is the history of the reliability of this system?
System	How much uncertainty does the system have?
System	How often is the system checked to make sure it is functioning as it was designed (i.e., for model accuracy)?
System	What is the system's level of confidence in this recommendation?
System	How is the confidence of the system measured?

REFERENCES

Achinstein, P. (1977). What is an Explanation? *American Philosophical Quarterly*, *14*(1). http://doi.org/10.2307/20009644

Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association : JMLA*, *103*(3), 152–153. http://doi.org/10.3163/1536-5050.103.3.010

Adobe Inc. (2018). *Digital Intelligence Briefing: 2018 Digital Trends*. Adobe Inc.

Agapie, E., Chinh, B., Pina, L. R., Oviedo, D., Welsh, M. C., Hsieh, G., & Munson, S. (2018). Crowdsourcing Exercise Plans Aligned with Expert Guidelines and Everyday Constraints. Presented at the the 2018 CHI Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/3173574.3173898

Ahn, W. K., & Bailenson, J. (1996). Causal Attribution as a Search for Underlying Mechanisms: An Explanation of the Conjunction Fallacy and the Discounting Principal. *Cognitive Psychology*, *31*, 82–123.

Ajzen, I. (1996). The Social Psychology of Decision Making. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social Psychology Handbook of basic principles* (pp. 297–325). New York, NY.

Aliannejadi, M., Rafailidis, D., & Crestani, F. (2018). A Collaborative Ranking Model with Multiple Location-based Similarities for Venue Suggestion. Presented at the the 2018 ACM SIGIR International Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/3234944.3234945

Allen, M. (2018). Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates. Retrieved October 23, 2018, from

https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates.

Amat, F., Chandrashekar, A., Jebara, T., & Basilico, J. (2018). Artwork Personalization at Netflix. Presented at the 12th ACM Conference on Recommender Systems, Vancouver, Canada: ACM Press. http://doi.org/10.1145/3240323.3241729

Amatriain, X. (2016). Past, Present, and Future of Recommender Systems. Presented at the the 21st International Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/2856767.2856798

Amirkhiabani, G., & Lovegrove, W. J. (1996). Role of Eccentricity and Size in the Global Precedence Effect. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(6), 1434–1447.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. (2015) *Plos One*, *10*(7) http://doi.org/10.1371/journal.pone.0130140

Bae, J., Ventocilla, E., Riveiro, M., Helldin, T., & Falkman, G. (2017). Evaluating Multi-attributes on Cause and Effect Relationship Visualization. Presented at the International Conference on Information Visualization Theory and Applications, SCITEPRESS - Science and Technology Publications. http://doi.org/10.5220/0006102300640074

Bainbridge, L. (1983). Ironies of Automation. *Automatica*, *19*(6), 775–779.

Banasick, S. (2018). Ken-Q Analysis. Retrieved from https://shawnbanasick.github.io/ken-q-analysis/index.html

Barg-Walkow, L. H., & Rogers, W. A. (2016). The Effect of Incorrect Reliability Information on Expectations, Perceptions, and Use of Automation. *Human Factors*, *58*(2), 242–260. http://doi.org/10.1177/0018720815610271

Bellman, R. E. (1961). Adaptive control processes: a guided tour. Princeton, NJ: Princeton University Press.

Bernstein, M. S., Bakshy, E., Burke, M., & Karrer, B. (2013). Quantifying the invisible audience in social networks. *ACM CHI Conference on Human Factors in Computing Systems* (pp. 21–30). New York, New York, USA: ACM. http://doi.org/10.1145/2470654.2470658

Berry, J., Segall, M. H., & Kagitcibasi, C. (1997). Handbook of Cross-cultural Psychology: Social behavior and applications. (J. Berry, Y. H. Poortinga, J. Pandey, P. R. Dasen, T. S. Saraswathi, M. H. Segall, & C. Kagitcibasi, Eds.) (Second). Allyn & Bacon.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. Presented at the IJCAI-17 Workshop on Explainable AI XAI.

Blume, L. E., & Easley, D. (2008). Rationality. In S. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics*. Palgrave-Macmillan: London, UK.

Bryant, A. D., Fletcher, G. S., & Payne, T. H. (2014). Drug interaction alert override rates in the Meaningful Use era. *Applied Clinical Informatics*, *5*(3), 802–813. http://doi.org/10.4338/ACI-2013-12-RA-0103

Buchannan, B., & Shortliffe, E. (1984). Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, MA: Addison Wesley.

Byrne, R. M. J. (2017). Counterfactual Thinking: From Logic to Morality. *Current Directions in Psychological Science*, *26*(4), 314–322. http://doi.org/10.1177/0963721417695617

Chalasani, P., Jha, S., Sadagopan, A., & Wu, X. (2018). Adversarial Learning and Explainability in Structured Datasets. https://arxiv.org/abs/1810.06583.

Chen, J. Y. C., Barnes, M. J., Selkowitz, A. R., & Stowers, K. (2016). Effects of Agent Transparency on human-autonomy teaming effectiveness. Presented at the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE. http://doi.org/10.1109/SMC.2016.7844505

Chen, S., & Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine*, *110*(2), 111–124. http://doi.org/10.1016/j.cmpb.2012.10.021

Cheverst, K., Byun, H. E., Fitton, D., Sas, C., Kray, C., & Villar, N. (2005). Exploring Issues of User Model Transparency and Proactive Behaviour in an Office Environment Control System. *User Modeling and User-Adapted Interaction*, *15*(3-4), 235–273. http://doi.org/10.1007/s11257-005-1269-8

Comfort, L., & Miller, C. (2004). *Decision-Making Under Uncertainty: The Three Mile Island Nuclear Accident From Multiple Perspectives*. University of Pittsburgh Institute of Politics: Pittsburgh, PA.

Corbin, J., & Strauss, A. (2008). Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory (Third). Thousand Oaks, CA: SAGE Publications, Inc. http://doi.org/10.4135/9781452230153

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., et al.

(2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496. http://doi.org/10.1007/s11257-008-9051-3

da Cruz, N. F., Tavares, A. F., Marques, R. C., Jorge, S., & de Sousa, L. (2015). Measuring Local Government Transparency. *Public Management Review*, *18*(6), 866–893. http://doi.org/10.1080/14719037.2015.1051572

David, J.-M., Krivine, J.-P., & Simmons, R. (Eds.). (1993). Second Generation Expert Systems (Vol. 6, pp. 39–44). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-77927-5

Devore, J. (1995). Probability and Statistics for Engineering and the Sciences (Fourth). New York, NY: Brooks/Cole.

Dix, A., Finlay, J., Abowd, G. D., & Beale, R. (2004). Human-Computer Interaction (Third Edition). Harlow, England: Pearson Prentice Hall.

do Prado Leite, J. C. S., & Cappelli, C. (2010). Software Transparency. *Business & Information Systems Engineering*, *2*(3), 127–139. http://doi.org/10.1007/s12599-010-0102-z

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. http://doi.org/10.1145/2347736.2347755

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. https://arxiv.org/abs/1702.08608

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer*

Studies, *58*(6), 697–718. http://doi.org/10.1016/S1071-5819(03)00038-7

Edison. (2018). *2018 Smart Audio Report*. Edison Research.

Eiter, T., & Lukasiewicz, T. (2006). Causes and explanations in the structural-model approach: Tractable cases. *Artificial Intelligence*, *170*(6-7), 542–580. http://doi.org/10.1016/j.artint.2005.12.003

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., et al. (2015). "I always assumed that I wasn't really that close to [her]". Presented at the the 33rd Annual ACM Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/2702123.2702556

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. http://doi.org/10.1038/nature21056

FAA. (1980). *Aircraft Alerting Systems Standardization Study* (No. FAA-RD-80-68). *Defense Technical Information Service*.

Feltovich, P. J., & Coulson, R. L. (2001). Learners'(mis) understanding of important and difficult concepts: A challenge to smart machines in education. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart Machines in Education The coming revolution in educational technology*. Menlo Park, CA: Researchgate.net.

Fingas, J. (2018). Google AI Can Spot Advanced Breast Cancer More Effectively Than Humans. Retrieved October 23, 2018, from https://www.engadget.com/2018/10/15/google-ai-spots-advanced-breast-cancer/

Fisher, C. W., & Kingma, B. R. (2001). Criticality of data quality as exemplified in two disasters. *Information and Management*, *39*, 109–116.

Fiske, S., & Taylor, S. (1991). Social Cognition: From Brains to Culture. Reading, MA: Addison Wesley.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The Application of Exploratory Factor Analysis in Applied Psychology: A critical review and analysis. *Personnel Psychology*, *39*(2), 291–314. http://doi.org/10.1111/j.1744-6570.1986.tb00583.x

Frankel, T. C. (2015). New machine could one day replace anesthesiologists. *The Washington Post*.

Gedikli, F., Jannach, D., & Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. *Journal of Human Computer Studies*, *72*(4), 367–382. http://doi.org/10.1016/j.ijhcs.2013.12.007

Geng, J. J., & Behrmann, M. (2005). Spatial probability as an attentional cue in visual search. *Perception & Psychophysics*, *67*(7), 1252–1268. http://doi.org/10.3758/BF03193557

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. https://arxiv.org/abs/1806.00069.

Glass, A., McGuinness, D. L., & Wolverton, M. (2008). Toward establishing trust in adaptive agents. Presented at the the 13th international conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/1378773.1378804

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press: Boston, MA.

Grand, S., & Wiedmer, M. (2010). Design fiction: A method toolbox for design research in a complex world. Presented at the International Conference of the Design

Research Society DRS. Montreal, Canada.

Gregor, S., & Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, *23*(4), 497.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. Presented at the ACM conference on computer-supported cooperative work, CSCW 00, Philadelphia, PA: ACM Press. http://doi.org/10.1145/358916.358995

Herman, B. (2017). The Promise and Peril of Human Evaluation for Model Interpretability. https://arxiv.org/abs/1711.07414

Hoffman, R. R. (2017). A Taxonomy of Emergent Trusting in the Human–Machine Relationship. In *Cognitive Systems Engineering* (1st ed., pp. 137–164). Boca Raton : Taylor & Francis, CRC Press, 2017. http://doi.org/10.1201/9781315572529-8

Hollyer, J. R., Rosendorff, B. P., & Vreeland, J. R. (2014). Measuring Transparency. *Political Analysis*, *22*(4), 413–434. http://doi.org/10.2307/24573081?refreqid=search-gateway:361e27acb8d161932abd6540c9227a1f

Holtzblatt, K., & Beyer, H. (2014). Contextual Design: Evolved. (J. M. Carroll, Ed.) (Vol. 7). United States: Morgan & Claypool. http://doi.org/10.2200/S00597ED1V01Y201409HCI024

Horrigan, J. B. (2017). *How People Approach Facts and Information*. Pew Research Center.

Huang, T.-H. K., Chang, J. C., & Bigham, J. P. (2018). Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time. Presented at the the

2018 CHI Conference, Montreal, Canada: ACM Press. http://doi.org/10.1145/3173574.3173869

Jamuna, K. S., & Karpagavalli, S. (2010). Classification of seed cotton yield based on the growth stages of cotton crop using machine learning techniques. *International Conference on Advances in Computer Engineering*. http://doi.org/10.1109

Jonassen, D. H., & Hernandez-Serrano, J. (2002). Case-Based Reasoning and Instructional Design: Using Stories to Support Problem Solving. *Educational Technology, Research and Development*, *50*(2), 65–77.

Kakar, A. S. (2016). A User-Centric Typology of Information System Requirements. *Journal of Organizational and End User Computing*, *28*(1), 32–55. http://doi.org/10.4018/JOEUC.2016010103

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254. http://doi.org/10.1146/annurev.psych.57.102904.190100;subPage:string:Abstract;issue:issue:10.1146/psych.2006.57.issue-1;journal:journal:psych;wgroup:string:AR

Klein, D. A. (1994). Decision-Analytic Intelligent Systems : Automated Explanation and Knowledge Acquisition. Routledge. http://doi.org/10.4324/9780203772775

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems*, *19*(06), 91–95. http://doi.org/10.1109/MIS.2004.74

Kratchounova, D., Fiore, S., & Jentsch, F. (2004). Design of learning environments for complex system architectures. Presented at the Second Human Performance, Situation Awareness and Automation Conference HPSAA II, Daytona Beach, FL.

Krause, J., Perer, A., & Ng, K. (2016). Interacting with Predictions (pp. 5686–5697). Presented at the the 2016 CHI Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/2858036.2858529

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of Explanatory Debugging to Personalize Interactive Machine Learning. Presented at the the 20th International Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/2678025.2701399

Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012). Tell me more: The effects of mental model soundness on personalizing an intelligent agent. Presented at the 2012 ACM annual conference, Austin, TX: ACM Press. http://doi.org/10.1145/2207676.2207678

Landman, A., Groen, E. L., van Paassen, M. M. R., Bronkhorst, A. W., & Mulder, M. (2017). Dealing With Unexpected Events on the Flight Deck: A Conceptual Model of Startle and Surprise. *Human Factors*, *68*(1), 001872081772342–12. http://doi.org/10.1177/0018720817723428

Lawshe, C. H. (1975). A Quantitative Approach to Content Validity. *Personnel Psychology*, *28*, 563–575.

Leis, R., Reinerman-Jones, L., Mercado, J. E., Szalma, J., & Hancock, P. A. (2015). Workload Change Over Time for Nuclear Power Plant Operation Tasks (Vol. 59). Presented at the 56th Annual Meeting of the Human Factors and Ergonomics Society. http://doi.org/10.1177/1541931215591022Copyright

Lidz, J., Pietroski, P., Halberda, J., & Hunter, T. (2011). Interface transparency and the psychosemantics of most. *Natural Language Semantics*, *19*(3), 227–256.

http://doi.org/10.2307/43550288?refreqid=search-gateway:361e27acb8d161932abd6540c9227a1f

Lim, B. Y., & Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. Presented at the the 11th international conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/1620545.1620576

Lim, B. Y., & Dey, A. K. (2010). Toolkit to support intelligibility in context-aware applications. Presented at the the 12th ACM international conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/1864349.1864353

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). 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: ACM Press. http://doi.org/10.1145/1518701.1519023

Lintern, G. (2013). Cognitive Work Analysis: Cognitive Systems Design. Available at http://www.cognitivesystemsdesign.net/Tutorials/CWA%20Tutorial.pdf

Liptak, A. (2018). A facial recognition program used by British police yielded thousands of false positives. Retrieved October 23, 2018, from https://www.theverge.com/2018/5/6/17324496/south-wales-police-automated-facial-recognition-false-positives-privacy-security

Liu, S., Wang, X., Liu, M., & Zhu, J. (2017). Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics*, *1*(1), 48–56. http://doi.org/10.1016/j.visinf.2017.01.006

Liu, Z., & Stasko, J. T. (2010). Mental Models, Visual Reasoning and Interaction in Information Visualization: A Top-down Perspective. *IEEE Transactions on*

Visualizations and Computer Graphics, *16*(6), 999–1008.

Lombrozo, T. (2012). Explanation and Abductive Inference. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199734689.013.0014

Lombrozo, T., & Vasilyeva, N. (2017). Causal Explanation. In M. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning*. Oxford, UK.

Lyons, J. B. (2013). Being Transparent about Transparency: A Model for Human-Robot Interaction. Presented at the 2013 AAAI Spring Symposium.

Markoff, J. (2011, March 4). Armies of Expensive Lawyers, Replaced by Cheaper Software. *The New York Times*.

Marquez, J., Cummings, M., Roy, N., Kunda, M., & Newman, D. (2012). Collaborative Human-Computer Decision Support for Planetary Surface Traversal. Presented at the Infotech@Aerospace 2012, Reston, Virigina: American Institute of Aeronautics and Astronautics. http://doi.org/10.2514/6.2005-6993

Marwick, A. E., & boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. http://doi.org/10.1177/1461444810365313

McGuinness, D. L., Glass, A., Wolverton, M., & Da Silva ExaCt, P. P. (2007). A Categorization of Explanation Questions for Task Processing Systems. Presented at the AAAI Workshop on Explanation-Aware Computing ExaCt.

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors*, *58*(3), 401–415.

http://doi.org/10.1177/0018720815621206

Merriam-Webster. (2018). The Merriam-Webster Dictionary. Springfield, MA.

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. https://arxiv.org/abs/1706.07269. http://doi.org/1706.07269

Monk, A. F. (1985). Fundamentals of human-computer interaction. New York, NY: Academic Press.

Monk, K., Shively, R. J., Fern, L., & Rorie, R. C. (2015). Effects of Display Location and Information Level on UAS Pilot Assessments of a Detect-and-Avoid System. Presented at the 2015 Annual Meeting of the Human Factors and Ergonomics Society. http://doi.org/10.1177/1541931215591011Copyright

Moore, J. D., & Swartout, W. R. (1988). *Explanation in Expert Systems: A Survey* (No. AD-A206 283) (pp. 1–58).

Mosier, K. L., & Skitka, L. J. (1996). Human Decision Makers and Automated Decision Aids: Made for Each Other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance Theory and applications.* Lawrence Erlbaum: Newark, NJ.

Muhlbacher, T., Piringer, H., Gratzl, S., Sedlmair, M., & Streit, M. (2014). Opening the Black Box: Strategies for Increased User Involvement in Existing Algorithm Implementations. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1643–1652. http://doi.org/10.1109/TVCG.2014.2346578

Mumaw, R. J. (2017). Analysis of alerting system failures in commercial aviation accidents (Vol. 61). Presented at the 61st Annual Meeting of the Human Factors and Ergonomics Society. http://doi.org/10.1177/1541931213601493

Mumaw, R. J., Roth, E., Vicente, K. J., & Burns, C. M. (2000). There Is More to

Monitoring a Nuclear Power Plant than Meets the Eye. *Human Factors*, *42*(1), 36–55.

Muñoz, L. A., & Bolivar, M. P. R. (2015). Determining factors of transparency and accountability in local governments: A meta-analytic study. *Journal of Local Self-Government*, *13*(2), 129–160. http://doi.org/10.4335/13.2.129-160

Murmann, P. (2018). Usable transparency for enhancing privacy in mobile health apps. Presented at the the 20th International Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/3236112.3236184

National Transportation Safety Board. (2010). *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* (No. NTSB/AAR-10/01 PB2010-910401). National Transportation Safety Board.

Nielsen, J. (1994). Enhancing the Explanatory Power of Usability Heuristics. Presented at the ACM CHI Conference on Human Factors in Computing Systems, Boston, MA.

Nissenbaum, H. (2011). A Contextual Approach to Privacy Online. *The Journal of the American Academy of Arts and Sciences*, *140*(4), 32–48.

Norman, D. A. (1988). The design of everyday things. New York: Basic Books.

Ososky, S., Sanders, T., Jentsch, F., Hancock, P. A., & Chen, J. Y. C. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems (Vol. 9084). Presented at the SPIE Defense + Security, SPIE. http://doi.org/10.1117/12.2050622

Owotoki, P., & Mayer-Lindenberg, F. (2007). Transparency of Computational Intelligence Models. *Research and Development in Intelligent Systems XXIII*, 387–

421. http://doi.org/10.1007/978-1-84628-663-6_29

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761–1780. http://doi.org/10.1037/xge0000318

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part a: Systems and Humans*, *30*(3), 286–297. http://doi.org/10.1109/3468.844354

Park, H., & Blenkinsopp, J. (2016). Transparency is in the eye of the beholder: the effects of identity and negative perceptions on ratings of transparency via surveys. *International Review of Administrative Sciences*, *83*, 177–194. http://doi.org/10.1177/0020852315615197

Pega. (2018). *What Consumers Really Think About AI: A Global Survey*. Pegasystems, Inc. Retrieved from https://www1.pega.com/ai-survey?utm_campaign=RCE&utm_medium=oso&utm_source=tw

Preece, J., Sharp, H., & Rogers, Y. (2015). Interaction Design: Beyond Human Computer Interaction (4 ed.). Wiley: West Sussex, UK.

Ram, A. (1993). AQUA: Questions that Drive the Explanation Process. In R. C. Schank, A. Kass, & C. K. Risebeck, Eds. Inside Case-Based Explanation, Chapter 7, pages 207–261, Lawrence Erlbaum Associates, 1994.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Presented at the the 22nd ACM SIGKDD International Conference, San Francisco, CA: ACM Press. http://doi.org/10.1145/2939672.2939778

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender Systems Handbook. Boston, MA: Springer US.

Riveiro, M., Helldin, T., Falkman, G., & Lebram, M. (2014). Effects of visualizing uncertainty on decision-making in a target identification scenario. *Computers and Graphics*, *41*(C), 84–98. http://doi.org/10.1016/j.cag.2014.02.006

Rogers, Y. (2012). HCI Theory: Classical, Modern, and Contemporary. (J. M. Carroll, Ed.) (Vol. 5). Synthesis Lectures on Human-Centered Informatics. http://doi.org/10.2200/S00418ED1V01Y201205HCI014

Rose. (2006). Human-Centered Design Meets Cognitive Load Theory: Designing Interfaces that Help People Think, Presented at the 14th ACM International Conference on Multimedia, Santa Barbara, CA.

Ross, L., & Ward, A. (2018). Naive Realism: Implications for Social Conflict and Misunderstanding. In E. S. Reed, E. Turiel, & T. Brown (Eds.), *Social cognition: The Ontario Symposium* (pp. 305–321). Hillsdale, NJ.

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, *100*(3), 349–363. http://doi.org/10.1037/0033-2909.100.3.349

Sadler, G. G., Battiste, H., Ho, N., Hoffmann, L. C., Johnson, W., Shively, R., et al. (2016). 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), IEEE. http://doi.org/10.1109/DASC.2016.7777998

SAE. (2016). *Surface Vehicle Recommended Practice- Taxonomy and Definitions for Terms Related to Driving Automated Systems for On-Road Motor Vehicles* (No. SAE

J3016_201609).

Sarter, N. B., & Schroeder, B. (2001). Supporting Decision Making and Action Selection under Time Pressure and Uncertainty: The Case of In-Flight Icing. *Human Factors*, *43*(4), 573–583.

Sarter, N. B., & Woods, D. D. (1995). How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors*, *37*(1), 5–19.

Schwartz, B. (2004). The paradox of choice: Why more is less. Harper Collins: New York, NY.

Sears, A., & Jacko, J. A. (2007). Human-computer interaction handbook: fundamentals, evolving technologies and emerging applications; 2nd ed. Boca Raton, FL: CRC Press.

Sebok, A., & Wickens, C. D. (2017). Implementing Lumberjacks and Black Swans Into Model-Based Tools to Support Human–Automation Interaction. *Human Factors*, *59*(2), 189–203. http://doi.org/10.1177/0018720816665201

Silveira, M. S., de Souza, C. S., & Barbosa, S. D. J. (2001). Semiotic engineering contributions for designing online help systems (pp. 31–38). Presented at the the 19th annual international conference, Santa Fe, NM: ACM Press. http://doi.org/10.1145/501521.501523

Simonite, T. (2016, March 29). Automated Anesthesiologist Suffers a Painful Defeat. *MIT Technology Review*.

Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. Presented at the CHI '02 extended abstracts, New York, New York, USA: ACM

Press. http://doi.org/10.1145/506443.506619

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. http://doi.org/10.1016/j.jbusres.2016.08.001

Snyder, A. (2018, May 3). The AI farm experiment. Retrieved October 23, 2018, from

Spice, B. (2018). New CMU Degree Prepares Researchers for AI-Directed Experimentation. Retrieved October 23, 2018, from https://www.axios.com/artificial-intelligence-in-agriculture-6f066e94-c704-4bbd-85a7-e638b871adce.html

Sterling, B. (2009). Design fiction. *Interactions*, *16*(3), 20–24. http://doi.org/10.1145/1516016.1516021

Streitz, N. A. (1988). Mental Models and Metaphors: Implications for the Design of Adaptive User-System Interfaces. In *Human-computer interaction handbook : fundamentals, evolving technologies and emerging applications* (pp. 164–186). New York, NY: Springer, New York, NY. http://doi.org/10.1007/978-1-4684-6350-7_8

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. https://arxiv.org/abs/1703.01365.

Swartout, W. R., & Moore, J. D. (1993). Explanation in Second Generation Expert Systems. In J.-M. David, J.-P. Krivine, & R. Simmons (Eds.), *Second Generation Expert Systems* (pp. 543–585). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-77927-5_24

Swearingen, K., & Sinha, R. (2001). Beyond algorithms: An HCI perspective on recommender systems. *ACM SIGIR 2001 Workshop on Recommender Systems.*

Taleb, N. N. (2007). The Black Swan: The Impact of the Highly Improbable. Random House Group.

Thakur, G. S., Sparks, K., Li, R., Stewart, R. N., & Urban, M. L. (2016). Demonstrating PlanetSense (pp. 1–4). Presented at the the 24th ACM SIGSPATIAL International Conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/2996913.2996975

Theodoridis, S. (2015). Machine Learning. London: Academic Press.

Trapsilawati, F., Wickens, C. D., Chen, C. H., & Xu, X. (2017). Transparency and conflict resolution automation reliability in air traffic control. Presented at the 19th Annual Symposium on Aviation Psychology, Dayton, OH.

Trapsilawati, F., Wickens, C. D., Qu, X., & Chen, C. H. (2016). Benefits of Imperfect Conflict Resolution Advisory Aids for Future Air Traffic Control. *Human Factors*, *58*(7), 1007-1019. http://doi.org/10.1177/0018720816655941

Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(4), 453–458. http://doi.org/10.1126/science.7455683

Valverde, R. (2014). Principles of Human Computer Interaction Design. (R. Valverde, Ed.). Lambert Academic Publishing: Montreal.

van Exel, J., & de Graaf, G. (2005). Q Methodology: A Sneak Preview. Retrieved March 27, 2018, from www.jobvanexel.nl.

Viégas, F. B. (2006). Bloggers' Expectations of Privacy and Accountability:An Initial Survey. *Journal of Computer-Mediated Communication*, *10*(3), 00–00. http://doi.org/10.1111/j.1083-6101.2005.tb00260.x

Waddell, K. (2018, October 20). The big picture: Even scientists are being automated. Retrieved October 23, 2018, from https://www.axios.com/automating-science-2e43b5c8-cb4a-44cb-bfce-f26272b1bd4c.html

Walker, V. R. (2010). Transforming science into law: Transparency and default reasoning in international trade disputes. In W. Wagner & R. Steinzor (Eds.), *Rescuing science from politics Regulation and the distortion of scientific research* (pp. 165–192). http://doi.org/https://doi.org/10.1017/CBO9780511751776

Wang, F.-Y., Carley, K. M., Zeng, D., & Mao, W. (2007). Social Computing: From Social Informatics to Social Intelligence. *IEEE Intelligent Systems*, *22*(2), 79–83. http://doi.org/10.1109/MIS.2007.41

Watts, S., & Stenner, P. (2005). Doing Q Methodology: theory, method and interpretation. *Qualitative Research in Psychology*, *2*(1), 67–91. http://doi.org/10.1191/1478088705qp022oa

Wickens, C. D. (2014). Effort in Human Factors Performance and Decision Making. *Human Factors*, *56*(8), 1329–1336. http://doi.org/10.1177/0018720814558419

Wright, J. L., Chen, J. Y. C., Barnes, M. J., & Hancock, P. A. (2017). The Effect of Agent Reasoning Transparency on Complacent Behavior: An Analysis of Eye Movements and Response Performance (Vol. 61, pp. 1594–1598). Presented at the 61st Annual Meeting of the Human Factors and Ergonomics Society. http://doi.org/10.1177/1541931213601762

Yan, L. (2018, July 14). Chinese AI Beats Doctors in Diagnosing Brain Tumors. *Popular Mechanics*.

Yantis, S., & Jonides, J. (1990). Abrupt Visual Onsets and Selective Attention: Voluntary

Versus Automatic Allocation. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(1), 121–134. http://doi.org/10.1.1.211.5016

Ye, L. R., & Johnson, P. E. (1995). The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. *MIS Quarterly*, *19*(2), 157. http://doi.org/10.2307/249686

Yuji, W. (2017). The Trust Value Calculating for Social Network Based on Machine Learning (pp. 133–136). Presented at the 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), IEEE. http://doi.org/10.1109/IHMSC.2017.145

Zachary, W. (1986). A Cognitively Based Functional Taxonomy of Decision Support Techniques. *Human-Computer Interaction*, *2*(1), 25–63. http://doi.org/https://doi.org/10.1207

Zahabi, M., Kaber, D. B., & Swangnetr, M. (2015). Usability and Safety in Electronic Medical Records Interface Design: A Review of Recent Literature and Guideline Formulation. *Human Factors*, *57*(5), 805–834. http://doi.org/10.1177/0018720815576827

Zeller, A. F. (1970). Accidents and Safety. In K. B. DeGreene (Ed.), *Systems Psychology* (pp. 131–150). New York, NY.

Zhang, L., Zhang, H., & Hao, S. (2018). An equity fund recommendation system by combing transfer learning and the utility function of the prospect theory. *The Journal of Finance and Data Science*. *4*(4), 223-233. http://doi.org/10.1016/j.jfds.2018.02.003

Zhou, J., Khawaja, M. A., Li, Z., Sun, J., Wang, Y., & Chen, F. (2016). Making machine

learning useable by revealing internal states update - a transparent approach. *International Journal of Computational Science and Engineering*, *13*(4), 378–389. http://doi.org/10.1504/IJCSE.2016.080214

CURRICULUM VITAE

Eric Stephen Vorm

Education

2019, PhD, Human Computer Interaction

Indiana University Purdue University Indianapolis

2011, Master of Science, Educational Psychology

University of North Texas

2004, Bachelor of Science, Psychology

Oral Roberts University

Professional Experience

2012-Present: Aerospace Experimental Psychologist, US Navy

2005-2012: Fleet Marine Force Corpsman, US Navy

Presentations

Vorm, E.S., (2018) Assessing the Value of Transparency in Recommender Systems: An End-User Perspective. Presentation given at the 12th ACM Conference on Recommender Systems (RecSys), Vancouver, Canada.

Vorm, E.S., (2018) Assessing Demand for Transparency in Intelligent Systems Using Machine Learning. Presentation given at the IEEE Innovations in Intelligent Systems (INISTA) conference, Thessonaliki, Greece.

Vorm, E.S. (2017) Human Factors Considerations in Future Unmanned Aeromedical

Evacuation. Poster presented at the 2017 Military Health System Research Symposium (MHSRS), Orlando, FL.

Vorm, E.S. (2016). Approaches to Context-Based Proactive Decision Support. Presentation given at the Human Factors and Ergonomics Annual Meeting, Washington, D.C.

Vorm, E.S., Saitzyk, A., LeVan, J. (2014) Analysis of Self-Directed Violence on U.S. Navy Aircraft Carriers. Poster presented at the 85th Aerospace Medical Association (AsMA) Conference, San Diego, CA.

Vorm, E.S., Saitzyk, A. (2013). Not always fair winds and following seas: Preliminary findings of suicide-related events onboard US Navy aircraft carriers. Presentation given at the US Navy Aeromedical Conference, Pensacola, FL.

Publications

Combs, D.J.Y., Blincoe, S., Garris, C.P., & Vorm, E.S. (2016) They're Beyond WEIRD: Helpful Frameworks for Conducting Non-WEIRD Research. In J.V. Cohn, S. Schatz, H. Freeman, and D.J.Y. Combs (Eds), Modeling Sociocultural Influences on Decision Making. Boca Raton, FL: CRC Press.

Saitzyk, A., & Vorm, E.S. (2016). Self-Directed Violence Aboard U.S. Navy Aircraft Carriers: An Examination of General and Shipboard-Specific Risk and Protective Factors. Military Medicine. 181(4), 343-349. http://doi.org/10.7205

Vorm, E.S. (2018) Assessing Demand for Transparency in Intelligent Systems Using Machine Learning. In IEEE Innovations in Intelligent Systems (INISTA). Thessonaliki, Greece: IEEE Xplore, pp. 25-35.

Vorm, E.S. (2018) Assessing the Value of Transparency in Recommender Systems: An End-User Perspective. 12th ACM Conference on Recommender Systems (RecSys), Vancouver, Canada.

Vorm, E.S. (2014) Compliance and Social Influence. In Smith, K. (Ed), Trust Attitudes, and Social Influence: The Cross-Cultural Social Psychology of Counterinsurgency (62-79), Washington, D.C.: Springer.

Zachary, W., & Vorm, E.S. (2016) Approaches to Context-Based Proactive Decision Support. Proceedings of the Human Factors and Ergonomic Society 56th Annual Meeting, 60(1), 238-240. http://doi.org/10.1177/1541931213601053.