Assessing Demand for Transparency in Intelligent Systems Using Machine Learning

Eric S. Vorm
Indiana University Purdue University of Indianapolis
Department of Human Centered Computing
Indianapolis, Indiana USA
esvorm@iu.edu

***Abstract*—Intelligent systems offering decision support can lessen cognitive load and improve the efficiency of decision making in a variety of contexts. These systems assist users by evaluating multiple courses of action and recommending the right action at the right time. Modern intelligent systems using machine learning introduce new capabilities in decision support, but they can come at a cost. Machine learning models provide little explanation of their outputs or reasoning process, making it difficult to determine when it is appropriate to trust, or if not, what went wrong. In order to improve trust and ensure appropriate reliance on these systems, users must be afforded increased transparency, enabling an understanding of the systems reasoning, and an explanation of its predictions or classifications. Here we discuss the salient factors in designing transparent intelligent systems using machine learning, and present the results of a user-centered design study. We propose design guidelines derived from our study, and discuss next steps for designing for intelligent system transparency.**

***Index Terms*—Artificial Intelligence, Machine Learning, Transparency, Explainability, Intelligibility, Intelligent Systems**

I. Introduction

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. 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). Historically, these systems involved building a knowledge base, either developed by expert user input, or through aggregation of data in some other form, such as automated collaborative filtering. This knowledge base is the heart of a traditional rule-based intelligent system, and represents the total knowledge the system knows about a subject area, such as network troubleshooting or medicine.

Recently, the simultaneous maturing of powerful multi-threaded graphics processors and the availability of very large labeled training datasets have enabled a new generation of intelligent systems built on machine learning, which have given rise to the data-driven paradigm of deep learning on deep neural networks. Today, knowledge bases are built by letting the system develop its own rules and knowledge directly from the data, rather than the intensive efforts of knowledge engineering previously required. [1][h] Although todays intelligent systems built on sophisticated architectures such as convoluted neural networks may appear to have little in common with older generations of intelligent systems, fundamentally they still share much in common. Both digest data in order to output a recommendation to the user, whether in the form of a classification or prediction, and users in both cases must determine what to do with that output.

A key usability challenge for intelligent systems that has plagued the field for decades, however, is that these systems rarely offer sufficient explanations of their reasoning or logic to the user. In order for a person to make an informed decision in response to a computer-generated recommendation, they require some causal understanding of how a systems output was generated. Past work to provide meaningful explanations of system outputs have included methods to extract and trace system logic, often in the form of graphic decision trees, or narrative explanations that qualitatively explain to the user why the system arrived at its conclusion [1]. While this is possible with simple decision models such as logistic regression, this is no longer possible with machine learning due to the scale and complexity of their models. While several efforts are underway to develop methods that enable machine learning systems to explain their outputs [2]-[5], these efforts are still very much in their infancy, and are far from being codified.

Recently lawmakers and governments have begun expressed concern about emerging issues related to the increase in artificial intelligence assuming greater roles with more autonomy. Many researchers and potential users of such systems are concerned with the challenge of being able to detect whether or not an intelligent system may have developed bias in its decision rules [6], or the ability to validate that its reasoning is still within the parameters for which it was originally designed. To address these and other concerns regarding the transparency of machine learning, in 2016, the European Union passed the General Data Protection Regulation [7], a law that requires any decision made based on an algorithm to be explainable to the user. This law, and others like it, demand that users have a ”right to an explanation” concerning algorithm-created decisions that are based on personal information. These laws have

the potential to challenge the continued growth of artificial intelligence in industry, limiting its usefulness and utility.

At the same time, intelligent systems are expanding into areas of increasing risk, such as defense, medicine, and public safety, increasing the need for more transparency exponentially. This is particularly important in cases where unexpected system behavior can have a detrimental effect on user performance, such as aviation [8]. Users with questions like what is the system doing now? and what will it do next? [9] require answers in order to accomplish a common grounding with the system and establish appropriate and calibrated trust.

The need for intelligent system transparency is evident, but designers of transparent intelligent systems will require formative guidance to determine what information users require, how they prefer that information, and what information will have the greatest effect on establishing trust and promoting acceptance.

In this paper, we share findings from a user-centered design study that qualitatively assessed user information needs in the context of interactions with intelligent systems built on machine learning. Our contribution is to assess the demand for information: what questions users want answered, under what moderating circumstances, and what effect answering those questions may have on improving user satisfaction when interacting with intelligent systems.

We describe a formative experiment that used descriptive vignettes to expose participants to a range of hypothetical experiences with intelligent systems in order to investigate what types of information users want. We end with a discussion of why users of intelligent systems may demand certain types of information in different situations, and provide design recommendations for providing different information types to make intelligent systems explainable and more acceptable to users.

II. Background

The challenge of programming a computer system that can explain itself has been a concern since the invention of intelligent systems in the 1950s [10]. Early work in the psychology of explanations [11] and psycholinguistics served to guide our understanding of how machines might best communicate with their users. Clark, et. Al [12], [13] described the concept of common grounding amongst members of a team, which is the process through which groups tacitly agree to a common goal through some form of communication. In order to accomplish collaborative goals, team members must share knowledge, beliefs, and assumptions, and engage in continuous sharing of these qualities to prevent the breakdown of a common ground.

Klein, et al. [14] adopted this concept for human-machine teaming and human-computer interaction, and developed ten rules for team players that provide a useful framework to consider what intelligent systems must do to accomplish transparent interaction similar to human-human interactions. While these frameworks can guide our understanding of how an ideal discourse between computer and user may take place, developing methods that enable computers to explain their reasoning and share their knowledge of the problem space and decision environment have proven challenging.

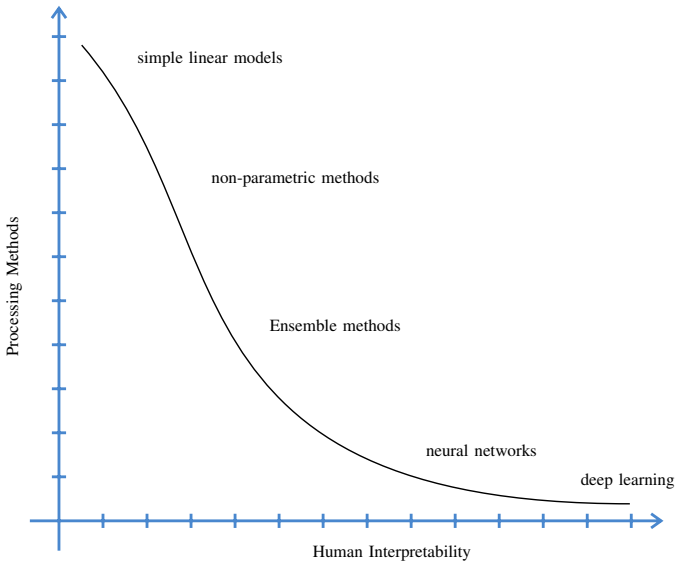

Fig. 1. Outputs from intelligent systems built on machine learning models are inherently challenging to explain to users.

Previous research on developing methods for having systems explain themselves has involved automatically generating explanations from the underlying knowledge base and deliver them to users. Early exemplars of these approaches, such as the MYCIN system that provided a certainty factor along with its recommendations, demonstrated the merit of these approaches, which have been since used in a variety of intelligent systems [1]. Research from other intelligent system domains such as context-aware systems [15], knowledge-based systems [16], and adaptive agents [17] are also approaches that have demonstrated merit.

These approaches, while promising in some cases, are somewhat limited in machine learning systems, however, because these models are often beyond human comprehension [15], [18]. Goodman and Flaxman [19] provided rank orderings of these systems on a scale of human intelligibility, and ranked machine learning approaches such as deep learning as the least intelligible or understandable, shown in figure 1 above.

While few efforts to make machine learning more intelligible and transparent have been demonstrated [4], [20], [21], few of these studies [22] [23] have included user evaluations to determine whether or not these explanations have a measurable effect on acceptance and usability. Without sufficient evaluation, designers of systems are left with little guidance in terms of what information is most important to the user, and will have the most impact on trust, usability, and acceptance. Thus, we have designed our study to explicitly assess user information needs in the context of intelligent systems built on machine learning that make recommendations to the user, who then must determine whether or not to use or ignore that recommendation.

III. Approach

Our hypothesis is that users will consider different information types more or less relevant or important depending on a variety of factors, including the context of the application, and the nature of the decision space. Previous research indicates that when users are provided with an explanation of system policies and reasoning strategies, they report greater levels of satisfaction and trust with automated collaborative filtering and recommender systems [24]-[26]. Our belief is that users will express similar desires for information in scenarios with intelligent systems involving machine learning, and that additional factors specifically related to machine learning may also be considered important and will lead to greater acceptance and trust.

We designed a study of user information needs for intelligent systems in order to assess how these demands align with or differ from previous research. Because intelligent systems employing machine learning approaches to decision support are not commonly available, we opted to use hypothetical scenarios. This allowed us to study user information needs across a wider range of potential intelligent system applications, and to do so in a more efficient manner.

A. Hypothetical intelligent systems

To assess user demand for information when interacting with intelligent systems, we developed five descriptive vignettes of hypothetical systems: 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.

1) 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.

2) 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.

3) 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.

4) 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.

5) 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.

B. Interaction vignettes

Participants were shown five short vignettes that described a first-person interaction with each hypothetical system. Each vignette described a scenario in which the participant was left at a decision point in which they would either have to act on the systems recommendation, or ignore it. All of these applications are machine-learning based, meaning they were said to have been trained on a representative data set of some sufficient size, and can reason about live data in order to provide a recommendation in the form of a prediction or classification to their users. Users were informed of the systems basic capabilities in the form of its output, but were not given any specific information, for instance where each system derives its data, or how each system represents user goals, etc.

IV. Methods

Participants in this study were graduate students in a human computer interaction program at a large mid-western university. The study began with a brief description of intelligent systems, specifically outlining machine learning applications of intelligent systems. Participants were then shown a vignette. Having read the vignette, participants were asked to write down any questions that they would want to ask the system that would aid them in determining whether or not to follow or ignore the systems recommendation. Participants were encouraged to consider these systems as being capable of discourse, 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.

At the conclusion of all vignettes, participants were instructed to post their sticky notes on a wall-sized white board in no particular order or arrangement, in accordance with traditional user-centered design [27]. Once all notes were posted on the wall, participants were asked to read through all sticky notes, and then asked to collaboratively discuss how the notes might be combined or grouped, and whether or not any moderating factors might be uncovered. 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 [28]. 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 white board. Open discussion in the format of a focus group was facilitated in order to capture qualitative insights from participant comments, adding a depth of understanding to why certain questions would be asked, and what those questions, if answered by the system, would accomplish in relation to user trust and willingness to act on system-generated recommendations. Participants were encouraged to rank order the vignettes according to different factors that they had identified. Once rank ordered, participants were asked to identify which information types were most important to them according to the context of the discussion. For example, when participants identified the criticality of the decision as a moderating factor, they were asked to rank order the vignettes according to the factor of criticality, and were then asked to indicate what information type would be most important to them in that particular context. Once the workshop concluded, all questions were entered into a spreadsheet according to their category, along with comments and notes taken by the primary investigator for analysis.

V. Results

A. Information Type Analysis

Using affinity diagramming, users identified seven distinct categories of information types: raw data, uncertainty, logic/reasoning, reliability, personalization, system confidence, and social. Using an open coding method [29], we arranged these information types into three categories: technological factors, personal factors, and social factors.

1) Technical factors: Information in this category pertains to the system itself, particularly relating to system policies, input sources, algorithm logic, and performance history. Users asked questions about the ***data itself, such as How current is the data used in making this recommendation? and How clean or accurate is the data used in making this recommendation? Questions about system ***uncertainty were also grouped into this category, and included questions like Is the system working with solid data, or is the system inferring or making assumptions on fuzzy information? Users asked several questions about the ***logic and inner workings of the system, such as Why is this recommendation the best option? and What are all of the factors (or indicators) that were considered in this recommendation, and how are they weighted? Questions of ***reliability such as Under what circumstances has this system been wrong in the past? were grouped into this category as well. These findings are supported by several other studies that have considered the impact of system history (i.e., its reliability) on user trust and acceptance of technology [30]-[33].

2) Personal factors: Information in this category pertains to the user, particularly how the system models the user, represents knowledge of the users needs or goals, expresses concepts such as risk or collateral damage, and whether or not users can influence or direct the system towards improved outputs (e.g., tractability). Questions of ***personalization, or whether or not a user is known by the system in terms of their needs and goals were included in this category. For example, Precisely what information about me does the system know? and 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 friends interests, etc.)? Users also considered questions of the systems ***confidence in this category, such as How does the system consider risk, and what is its level of acceptable risk? and What does the system think is MY level of acceptable risk?

3) Social factors: Information in this category pertains to the influence of social factors such as the prevalence of user behavior in relation to the systems output, how well the users profile matches the profiles of others receiving similar recommendations, and measures of satisfaction or success from other users who have previously interacted with the system in similar scenarios.

In many intelligent system contexts, particularly social collaborative filtering and recommender systems, the social dimension plays a particularly important role in user trust and acceptance. Users expressed similar preference for information related to this dimension for intelligent systems that use machine learning. Questions included How many other people have received this recommendation from this system? and How many other people have accepted or rejected this recommendation from this system? (What is the ratio of approve to disapprove?) It is plausible that answers to these questions, in some contexts, may serve as a meaningful heuristic to aid in user decision making.

B. Moderators

In addition to these categories of information, participants identified additional factors that could potentially moderate the preference for certain information, as well as the demand of that information. Participants identified three such moderating factors: external dependencies; the context of the decision on a scale of criticality, and the users role when interacting with the system.

1) External dependencies: Systems that were perceived to have high external dependencies were described as being less trustworthy by participants. Potential users may be skeptical of systems that rely heavily on external data, especially if the sources and qualities of that data are not made available to the user. Users expressed frustration with systems that intentionally hide this information, such as what data is used to determine ads in social media [34], and indicated the net effect of this encourages suspicion. Participants indicated that information related to raw data is the most valuable to them when interacting with systems that were perceived to to have high external dependencies. They also indicated that information about the quality of those dependencies, such as reliability and uncertainty information becomes increasingly important to determine appropriate decision making.

Information Categories	Information Types	Example Questions	Overall Indicated Preference	Moderators: Dependencies		Moderators: Criticality		Moderators: User Role		
				Low	High	Low	High	Personal End User	Expert End User	Secondary End User
Technical Factors	Data	How current is the data used in making this recommendation?	Highest	⇑						⇑
	Uncertainty	Is the system system inferring or making assumptions on fuzzy data?	Moderate		⇑		⇑		⇑	
	Logic	What are the pros/cons associated with this option?	Highest							
	Reliability	Under what circumstances has this system been wrong in the past?	Moderate		⇑			⇓		
Personal Factors	Personalization	Was this recommendation made specifically for ME?	Moderate			⇑		⇑		⇑
	Confidence	How is the confidence of the system measured?	Highest				⇑		⇓	
	Tractability	What if I decline?	Lowest	⇑	⇓	⇑				
Social Factors	Social Filtering	How alike am I to other people receiving this recommendation?	Lowest			⇑			⇓	

TABLE I

TABLE OF FINDINGS FROM USER-CENTERED DESIGN STUDY. HOW TO READ: IN GENERAL, USERS DEMAND INFORMATION ABOUT DATA, LOGIC, AND CONFIDENCE MOST. MODERATORS SUCH AS SYSTEMS WITH HIGHLY CRITICAL DECISIONS INCREASE THE VALUE OF UNCERTAINTY AND SYSTEM CONFIDENCE INFORMATION. INFORMATION RELATED TO SOCIAL FACTORS BECOME LESS VALUABLE FOR USER ROLES SUCH AS EXPERT END USERS.

2) Criticality: Vignettes in which the decision was perceived as having a great degree of risk, either to the users themselves, or to others (e.g., secondary users of the system, such as job seekers in HR-Kit or patients in the ONNPAR scenario) were labeled as more critical than others. We arranged the vignettes on this dimension based on participant ranking, presented in Table II below. The importance of information about the systems confidence level and amount of uncertainty was expressed to be the most important and valuable information to users interacting with systems involving decisions perceived to be of high criticality.

Most Critical	ONNPAR
-	D-SAM
-	HR-KIT
-	Social Media
Least Critical	Q-Conc

TABLE II

PARTICIPANT RANK ORDERING OF HYPOTHETICAL VIGNETTES ON A SCALE OF MOST TO LEAST CRITICAL.

3) User Role: Participants reasoned that the value of information may depend, in part, also on the role of the user interacting with an intelligent system. Three generic roles were identified through our vignettes: a personal end user, or someone who uses the system for their own purposes; an expert end user, or someone who uses the system to accomplish a goal or task that affects others (such as a physician using a clinical decision support system to diagnose a patient); and a secondary end user, or someone who is the recipient of a computer-generated output, such as a person applying for a home loan or a patient in a hospital. In vignettes that described the user as a personal end user, such as the social media, Q-conc, and D-Sam vignettes, participants expressed preference for information that would help them understand the application itself, such as understanding how the system processes information (e.g., logic or reasoning). Participants also indicated that social factors such as whether or not friends or family members had used and been satisfied with such systems may play a role in determining their willingness to use those systems. Conversely, for vignettes that described the user as an expert end user, such as HR-KIT where the user is an HR manager who must decide who to hire, participants indicated an increased preference for uncertainty information, plausibly to help determine whether or not a system recommendation was accurate or not. For vignettes where the user was a secondary end user, such as ONNPAR where the user is a patient who receives a treatment recommendation generated by a computer, participants expressed preference for information related to the systems reliability, and whether or not the systems output was based on the users data (e.g., questions of personalization) as the most important information.

Discussions on this moderating factor were often contentious and spirited, and investigators did not feel there was a high degree of agreement amongst participants, based on the range of expression for information types. This may indicate that individual differences such as different information seeking schemas or styles of information seeking behavior may also play a role in determining the value of certain information types depending on the role of the user, or a combination of moderating factors. We will address this point further in the discussion below.

VI. DISCUSSION

Through five vignettes describing user interactions with hypothetical intelligent systems, we assessed the kinds of information users desire in order to support decision making and build appropriate trust. Our vignettes specifically described scenarios in which an intelligent system provides a recommendation to the user, which the user must then determine whether or not to follow or ignore (i.e., accept or reject).

Participants identified a range of information needs in the form of questions for which they want the system to answer. These questions were arranged and classified using participant-led discussions and guided by the affinity diagramming technique. These information categories were then re-assessed in relation to three moderating factors: a scale of criticality, or decisions that involve a degree of personal risk to selves or others; the degree to which each system relies on external dependencies; and the role of the user, either expert, novice, or recipient. All information types were then arranged into three categories of technological factors, personal factors, and social factors.

Preferred information

Our findings indicate that users of intelligent systems that offer recommendations demand a wide range of information types in order to feel comfortable accepting those recommendations. This willingness to accept a computer-generated recommendation is considered a proxy measure of trust. Participants generally expressed they want to know information about how the system operates and how its outputs

are generated, (e.g., logic/reasoning, system policies, etc.). Participants also consider information about how their data, needs, goals, and input is represented and taken into account by recommendations made by intelligent systems. Factors related to a users profile, including information about the behavior of other similar users in response to these systems is also something participants considered worthy of merit. When computer-generated recommendations are perceived to be incorrect or inappropriate, users want the ability to correct and train the system themselves.

Participants recognized that what information is most important to users may change as a function of different situations. Because of this, the value of different types of information may be moderated by the role of the intended user, the degree of criticality of the decision, and the amount of external dependencies present in the system.

Highly Critical Contexts Demand Different Information

User needs for information will likely change as the degree of criticality- or the amount of risk involved in the decision-increases. In addition to other highly valued information, details about the systems confidence level and level of uncertainty are preferred in highly critical contexts. This aligns with previous decision theory research that suggests decision makers seek information that can minimize or reduce uncertainty as a useful decision heuristic [1].

User Role Affects Information Needs

Determining what information is valuable or critical to display to a user will depend, in part, on the role that user is in. Users who are using the system to accomplish personal end goals, such as determining how best to manage their financial portfolios, will likely seek information that assists their understanding of how the system processes data to arrive at its outputs (e.g., system logic, reasoning, and policies). These users may also benefit from relating information of other similar users, including how others in similar circumstances have fared when interacting with similar system-generated recommendations. Users who may be making decisions on behalf of others, such as physicians diagnosing or treating patients, may require additional information that serve to justify the systems recommendation. Finally, users who are recipients of system-generated recommendations, such as patients themselves, prefer information that indicate the systems prior performance, such as reliability data, as well as information that helps users determine whether or not their needs and goals are known by and represented in the system (e.g., personalization).

Tractability

In cases when the system may be incorrect, some users may be willing to continue to use (trust) it, provided they have a means to influence and direct its learning themselves. This concept of tractability was found to be a highly desirable feature by our participants, and designers can anticipate these desires by providing interfaces that afford this functionality.

VII. Limitations

Our user-centered workshop was conducted with participants that, while familiar with human-computer interaction and UX best practices, were not subject matter experts. In future iterations of our workshop, we plan several changes, including the use of Q-methodology [35] to allow for enhanced statistical evaluation of user preference and demand for different types of information. We will also conduct these workshops with subject matter experts in intelligent system design in order to increase the validity of our findings. Due to the preliminary nature of literature review, and the non-exclusive inclusion criteria chosen, traditional analyses of effects sizes and variance could not be performed. Future efforts to provide quantitative analyses of the literature on intelligent system transparency is already

An additional dimension we did not assess is 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.

Participant interactions and sentiment expressed revealed the possibility of different information schemas. We are interested in evaluating these potential individual differences, but will need additional information in order to assess them definitively. The current format did not allow for any quantitative evaluation of information styles amongst participants, and so we are planning this for a future study.

While the data derived from this study is useful and descriptive, our design did not allow for measures of agreement or effect sizes to be evaluated amongst participants. Future studies will enable these factors to be quantified, which we believe will compliment these findings.

VIII. Conclusion

We have studied issues governing the trust and usability of intelligent systems offering decision support. The utility of intelligent systems is evident, but adoption can be hindered when users cannot understand the systems 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 will not be used and widely adopted.

Using a qualitative approach, we identified themes that describe the willingness of users to adopt and trust these agents, particularly in the context of the decision theoretic. Having identified these themes, our study recommends design guidelines for designers of intelligent systems using machine learning. These guidelines provide a theoretical framework for future work on evaluating the effect of improving intelligent system transparency on user interactions.

REFERENCES

[1] E. J. Horvitz, J. S. Breese, and M. Henrion, Decision theory in expert systems and artificial intelligence, International Journal of Approximate Reasoning, vol. 2, no. 3, pp. 247302, Jul. 1988.

[2] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, Principles of Explanatory Debugging to Personalize Interactive Machine Learning, presented at the the 20th International Conference, New York, New York, USA, 2015, pp. 126137.

[3] F. Doshi-Velez and B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, AirXiv, 2017.

[4] J. Zhou, M. A. Khawaja, Z. Li, J. Sun, Y. Wang, and F. Chen, Making machine learning useable by revealing internal states update - a transparent approach, International Journal of Computational Science and Engineering, vol. 13, no. 4, pp. 378389, 2016.

[5] W. Yuji, The Trust Value Calculating for Social Network Based on Machine Learning, presented at the 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2017, pp. 133136.

[6] B. Berendt and S. Preibusch, Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence, Artificial Intelligence Law, vol. 22, no. 2, pp. 175209, Jan. 2014.

[7] EU, Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016, pp. 188.

[8] A. Degani, The Crash of Korean Air Lines Flight 007, in Taming HAL, no. 4, New York, 2004.

[9] L. Sherry, M. Feary, P. Polson, and E. Palmer, What's it doing now?: Taking the covers off autopilot behavior, presented at the 11th International Symposium on Aviation Psychology, 2001.

[10] B. Buchannan and E. Shortliffe, Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, MA: Addison Wesley, 1984.

[11] N. Pennington and R. Hastie, Reasoning in explanation-based decision making, Cognition, vol. 49, pp. 123163, 1993.

[12] A. A. Clarke and M. G. G. Smyth, A co-operative computer based on the principles of human co-operation, International Journal of Man-Machine Studies, vol. 38, no. 1, pp. 322, Jan. 1993.

[13] H. H. Clark and S. E. Brennan, Grounding in Communication, in Perspectives on Socially Shared Cognition, no. 7, L. Teasley and S. D. Teasley, Eds. Washington, DC, 1991, pp. 127149.

[14] G. Klein, D. D. Woods, J. M. Bradshaw, R. R. Hoffman, and P. J. Feltovich, Ten Challenges for Making Automation a Team Player in Joint Human-Agent Activity, IEEE Intell. Syst., vol. 19, no. 6, pp. 9195, Nov. 2004.

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

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

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

[18] V. Bellotti and K. Edwards, Intelligibility and Accountability: Human Considerations in Context-Aware Systems, Human-Computer Interaction, vol. 16, no. 2, pp. 193212, 2001.

[19] B. Goodman and S. Flaxman, European Union Regulations on Algorithmic Decision-Making and a Right to Explanation, AI Magazine, vol. 38, no. 3, pp. 5057, Oct. 2017.

[20] B. M. Lake, R. Salakhutdinov, and J. B. Tenebaum, Human-level concept learning through probabilistic program induction, Science, vol. 350, no. 6266, pp. 13321338, 11-Dec-2015.

[21] P. Owotoki and F. Mayer-Lindenberg, Transparency of Computational Intelligence Models, Research and Development in Intelligent Systems XXIII, no. 29, pp. 387393, 2007.

[22] M. J. Pazzani, Representation of electronic mail filtering profiles, presented at the the 5th international conference, New York, New York, USA, 2000, pp. 202206.

[23] S. Stumpf, V. Rajaram, L. Li, M. Burnett, T. Dietterich, E. Sullivan, R. Drummond, and J. Herlocker, Toward harnessing user feedback for machine learning, presented at the the 12th international conference, New York, New York, USA, 2007, p. 82.

[24] M. C. Dorneich, S. D. Whitlow, C. A. Miller, and J. A. Allen, A superior tool for airline operations, Ergonomics in Design: The Quarterly of Human Factors Applications, vol. 12, no. 2, pp. 1823, 2004.

[25] J. L. Herlocker, J. A. Konstan, and J. Riedl, Explaining collaborative filtering recommendations, presented at the 2000 ACM conference, New York, New York, USA, 2000, pp. 241250.

[26] R. Sinha and K. Swearingen, The role of transparency in recommender systems, presented at the CHI '02 extended abstracts, New York, New York, USA, 2002, pp. 8302.

[27] Y. Rogers, HCI Theory: Classical, Modern, and Contemporary, vol. 5, no. 2. Synthesis Lectures on Human-Centered Informatics, 2012.

[28] K. Holtzblatt and H. Beyer, Contextual Design: Evolved, vol. 7, no. 4. United States: Morgan and Claypool, 2014.

[29] J. Corbin and A. Strauss, Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, Third. Thousand Oaks, CA: SAGE Publications, Inc., 2008.

[30] H. Atoyan, J.-R. Duquet, and J.-M. Robert, Trust in new decision aid systems, presented at the the 18th international conference, New York, New York, USA, 2006, pp. 115122.

[31] J. D. Lee and K. A. See, Trust in Automation: Designing for Appropriate Reliance, Human Factors, vol. 46, no. 1, pp. 5080, 2004.

[32] R. Parasuraman and C. A. Miller, Trust and Etiquette in High-Criticality Automated Systems, Communications of the ACM, pp. 5155, 2004.

[33] E. Kaltenbach and I. Dolgov, On the Dual Nature of Transparency and Reliability: Rethinking Factors that Shape Trust in Automation, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 61, no. 1, pp. 308312, Sep. 2017.

[34] M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig, 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, 2015, pp. 153162.

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