An Explanation-Centric Approach for Personalizing Intelligent Agents

Todd Kulesza
Oregon State University
Corvallis, Oregon
kuleszto@eecs.oregonstate.edu

ABSTRACT
Intelligent agents are becoming ubiquitous in the lives of users, but the research community has only recently begun to study how people establish trust in and communicate with such agents. I plan to design an explanation-centric approach to support end users in personalizing their intelligent agents and in assessing their strengths and weaknesses. My goal is to define an approach that helps people understand when they can rely on their intelligent agents' decisions, and allows them to directly debug their agents' reasoning when it does not align with their own.

Author Keywords
Interactive machine learning, mental models, trust, end-user programming.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

CONTEXT AND MOTIVATION
Intelligent agents have moved beyond mundane tasks like filtering junk email. Search engines now exploit pattern recognition to detect image content (e.g., clipart, photography, and faces); Facebook and image editors take this a step further, making educated guesses as to *who* is in a particular photo. Netflix and Amazon use collaborative filtering to recommend items of interest to their customers, while Pandora and Last.fm use similar techniques to create radio stations crafted to an individual's idiosyncratic tastes. Simple rule-based systems have evolved into agents employing complex algorithms. These *intelligent agents* are computer programs whose behavior only becomes fully specified *after* learning from an end user's training data. Advances in machine learning and pattern recognition continue to unlock new applications for intelligent agents, but two challenges prevent end users from more fully utilizing the automated decisions and recommendations from these tools.

First, end users of intelligent agents need to understand when they can trust, their agent's work. Such trust is highly contextual—some of the agent's predictions may not matter at all to an end user, while others may matter a great deal. Further, the agent's reasoning is constantly changing as it learns from the user's behavior, so a system that was reliable yesterday may not be trustworthy today.

The second challenge is about personalization. When an intelligent agent's reasoning causes it to perform unexpectedly in the field, only the end user is in a position to personalize, or more accurately, *to debug,* the agent's flawed reasoning. Here, debugging refers to *mindfully and purposely* adjusting the agent's reasoning (after its initial training) so it more closely matches the user's expectations. Recent research has made inroads into supporting this type of functionality [1,8,10,13], but debugging can be difficult for even trained software developers—helping end users, who have knowledge of neither software engineering nor machine learning, is no trivial task.

I believe these two challenges—establishing contextual trust in an agent, and aligning its reasoning with a specific end user's—are inherently linked. My thesis work aims to understand how people reason about intelligent agents, how people communicate with intelligent agents, and how intelligent agents can improve their reasoning given these communications. I hypothesize that explaining an agent's reasoning to end users will enable them to form better judgments of the agent's reliability, and to provide feedback that can substantially improve the agent's future predictions, as illustrated in Figure 1.

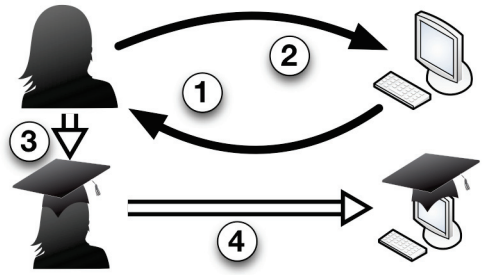

Figure 1: I envision a cyclic, explanation-based approach for users to learn about the agent's reasoning (1) and interactively adjust it (2). In the process, the user learns more about how to effectively steer the agent (3), with the eventual outcome of "more intelligent" intelligent agents (4).

IUI'12, February 14–17, 2011, Lisbon, Portugal.

RELATED WORK

Mental Models

Mental models are internal representations that people build based on their experiences in the real world. There are two main types of mental models: *Functional* (shallow) models provide the ability to use a system, whereas *structural* (deep) models provide a detailed understanding of how and *why* it works. These models do not have to be entirely *complete* to be useful, but they must be *sound* (i.e., accurate) enough to support effective interactions. Many instances of unsound mental models guiding erroneous behavior have been observed [15].

When things go wrong, the richness of a mental model can also have an effect. While a structural model can help someone deal with unexpected behavior and fix the problem, a purely functional model does not provide the abstract concepts required for such interactions [7]. Knowing how to *use* a computer, for example, does not mean you can *fix* one that fails to power on.

Because much of this thesis rests upon end users' ability to understand and adjust an intelligent agent's reasoning, an understanding of how users build mental models of this reasoning is foundational. Outside the realm of intelligent agents, it has been argued that users should be exposed to transparent, intuitive systems and appropriate instructions to build mental models [18]. *Scaffolded instruction* is one method that has been shown to contribute positively to learning to use a new system [16], but has not been explored in the context of machine leaning.

To date, very little work has investigated end users' mental models of an intelligent system. One study found prolonged use of such a system induced plausible models of its reasoning, but these models were surprisingly hard to shift, even when people were aware of contradictory evidence [21]. My preliminary work has found that users will change their mental models for an intelligent agent when it makes its reasoning transparent [10], but some explanations by agents may lead to only shallow mental models [19].

Interactive Machine Learning

There has been recent interest in supporting the debugging of intelligent agents' reasoning [1,8,10,12,13,19]. This debugging consists of two components: explanations about the agent's current reasoning, and end-user corrections to align this reasoning with his or her own.

For users to debug an intelligent agent's reasoning, they must first be able to see it. Explanations of intelligent agent's reasoning have taken a variety of forms, such as relating user actions and the resulting predictions [3], detailing why a program made a particular prediction [14], or explaining how an outcome resulted from user actions [21,22]. However, these explanations are limited to account for the intelligent agent's past behavior and do not extend to accepting user corrections to adapt *future* behavior.

The second component of interactive machine learning consists of users debugging faults in the machine's reasoning. One such system, EnsembleMatrix [20], provides a visualization of a classifier's accuracy and the means to adjust its logic, but is targeted at machine-learning experts *developing* complex ensemble classifiers, rather than end users *working with* the resulting classifiers. ManiMatrix [8] provides an interactive visualization of a classifier's accuracy, but user interactions are restricted to modifying a classifier's cost matrix. Most closely related to my own work, Stumpf et al. have explored potential end-user debugging techniques for intelligent agents, finding that even simple corrections from end users have the potential to increase the accuracy of an agent's predictions [19]. Some participants' corrections, however, decreased the quality of the agent's predictions; there were barriers that prevented them from successfully debugging the agent. I conducted a study [11] building upon this work, identifying and categorizing these barriers and the information participants requested to overcome them.

Contextual Trust

Existing work has investigated how end users establish trust in intelligent agents [6], identifying a number of themes (e.g., transparency, user expectations) that impact users' trust in such systems. Other researchers have identified a positivity bias toward trusting intelligent agents (i.e., people trust the agent without initial proof that it is trustworthy) [5]; this same bias is well-established among human interactions [4].

Dzindolet et al. found that end users quickly lose trust in agents once even a handful of agent mistakes are observed [5]. One explanation for this erosion of trust comes from theories on expectations, or *schemas*, in which unexpected situations stand out as individual events in memory, while expected situations are lumped together as a single schema [2] (e.g., someone is unlikely to remember the specifics of brushing his or her teeth this morning, but everyone can likely recall a time they dropped their toothbrush someplace they did not want it to land). Thus, after only a few mistakes on the part of the intelligent agent, an end user may be unduly influenced by the memory of these aberrations, and so form a lesser degree of trust in the system than may be warranted. My preliminary work has also found evidence of this phenomenon—in a 10-minute assessment task that involved identifying an agent's mistakes, participants consistently reported that the agent was less reliable than it actually was [9].

STATEMENT OF THESIS

The purpose of my dissertation research is to improve peoples' experiences with intelligent agents in two specific ways: (1) by helping people correct errant agents, and (2) by helping people establish appropriate levels of trust in their agent's work. My central thesis is that two-way communication between the user and the intelligent agent will be necessary for each of these tasks; thus, much of my

research will investigate explanations (both from the agent to the user and from the user to the agent), their potential content, and the impact this content may have on user's mental models and the agent's reasoning.

RESEARCH GOALS AND METHODS

Because replicating experiments over the diverse range of intelligent agents is a prohibitively large undertaking, I propose to focus on the domain of music recommendation systems. Music recommendation embodies two of the components of intelligent agents I care most about—they are widely used, and their work is *beneficial* without being *critical* to their end users. This is a more challenging problem than studying agents that perform critical tasks, because non-critical systems cannot assume a high level of attention and effort from their users. Further, low-cost solutions to these problems are likely to applicable to high-criticality systems, whereas high-cost solutions are unlikely to be useful for non-critical domains.

RQ1.1 (cost/benefit): How much will end users assess and provide feedback to intelligent agents when the perceived benefits are relatively low?

RQ1.2 (cost/benefit): How can interfaces lower the cost of assessing and providing feedback when the perceived benefits are relatively low?

I plan to study RQ1.1 via a longitudinal study of a deployed recommender system, while using low-fidelity prototypes to iteratively explore RQ1.2 in a laboratory setting.

My hypothesis is that sound structural mental models of the agent's reasoning will be critical to both assessment and feedback, so I will explore how mental models impact trust in and corrections of intelligent agents. Because this work will inform the design of explanations of an agent's reasoning, I will also explore how different attributes of explanations (salience, visibility, concreteness, etc.) influence the feedback users provide intelligent agents. I plan to develop a taxonomy of these attributes and the various trade-offs associated with them.

RQ2.1 (mental models): Do sound structural mental models of an intelligent agent's reasoning lead to more appropriate perceptions of its reliability?

RQ2.2 (mental models): Do sound structural mental models of an intelligent agent's reasoning lead to improved end-user debugging of its reasoning?

RQ2.3 (explanation attributes): What are common attributes of machine-generated explanations, and how do they influence the feedback provided by end users?

I plan to investigate RQ2.1 with a Wizard-of-Oz study. One of my goals for the Doctoral Consortium is to discuss methods for reliably exploring RQs 2.2 and 2.3. RQ2.3 seems particularly suited to a grounded theory approach, and will hopefully lead to a vocabulary and methodology for evaluating explanations.

I next intend to investigate methods for end users to communicate, either implicitly or explicitly, which of an agent's decisions matter most. Current work on active learning [17] views end users as tireless oracles—the main goal of such techniques is to shore up "fuzzy" decision boundaries by eliciting more details from the end user. I hypothesize the converse is attainable—end users could tell agents which of their decisions or decision boundaries need to be reliably accurate, so that any information the agents ask users to provide is directly related to what the user most needs the agent to get right.

RQ3.1 (evaluation abstraction): At what level of abstraction do users prefer to assess an agent's reasoning (e.g., decision boundaries, individual predictions, etc.)?

RQ3.2 (evaluation prioritization): Can end users tell intelligent agents which predictions matter to them in a way that allows the agent to apply active learning techniques to focus on particular classes or decision boundaries?

RQ3.1 will be investigated via a formative lab study involving low-fidelity intelligent agent prototypes. During this study I expect to gather participant feedback in a manner that will permit me to run offline studies exploring RQ3.2 via modifications to active learning techniques.

To conclude my dissertation research, I plan to use the results of RQs 1-3 to inform an approach for end users to assess and debug the reasoning of intelligent agents. This will be followed with a summative user study to evaluate the overall success of (and individual strengths and weaknesses inherent in) the approach. In essence, I will be attempting to answer these final research questions:

RQ4.1 (effectiveness): Can end users reliably assess the accuracy of an intelligent agent via an explanation-centric approach?

RQ4.2 (effectiveness): Can end users reliably improve the accuracy of an intelligent agent via an explanation-centric approach?

DISSERTATION STATUS

I am currently studying RQs 1 and 2 via a longitudinal study of end users steering a customizable music recommendation system. The preliminary results suggest that end users are initially willing to spend time and effort debugging an agent's reasoning, and that the more they learn about how the agent while debugging it, the more beneficial they view these debugging efforts. I have previously investigated parts of RQ 2 via both a qualitative study and a summative quantitative study, identifying a vocabulary end users wanted to employ when explaining corrections to an intelligent agent, and finding that presenting "run-time" debugging explanations to end users helped them debug the agent's reasoning more than "static" explanations of the agent's logic [10]. I have also begun exploring RQ3 with a statistical study exploring how end users responded to a systematic approach for assessing an

agent's decisions, and testing a technique for leveraging user assessments to test similar agent predictions [9].

I plan to use qualitative studies to explore RQs 1.2, 2.3, 3.1, and 3.2, which I hope to complete in the next 18 months. These results will iteratively inform an explanation-centric debugging and assessment approach and prototype. This approach and prototype will be evaluated by a final summative study approximately two years from now.

OPEN QUESTIONS

I have three primary questions to discuss with the researchers at the IUI Doctoral Consortium.

First, capturing mental models is not trivial [15]. I would like to discuss other researchers' experiences capturing mental models, including methods others have found useful and potential drawbacks to be aware of.

Second, I am still uncertain of how to design formative studies to answer RQ2, and would like to discuss design ideas with researchers who have performed similar formative work.

Finally, I am looking for different levels of abstraction to use when investigating RQ 3.1. The two I have in mind are general decision boundaries and individual predictions, but would like to hear ideas from researchers who have more extensive machine learning backgrounds than my own.

My dissertation research is gathering momentum, and with approximately two years left, I feel this is an ideal time to discuss my work at the IUI Doctoral Consortium. I have enough experience with human-computer interaction research to constructively discuss the work of fellow students, but am not yet too far along to integrate their advice into my own research.

REFERENCES

1. Amershi, S., Fogarty, J., Kapoor, A. and Tan, D. 2010. Examining multiple potential models in end-user interactive concept learning. *Proc. CHI*, 1357–1360.
2. Ashcraft, M.H. 1994. *Human memory and cognition*. Harpercollins College Div.
3. Billsus, D. and Hilbert, D. 2005. Improving proactive information systems. *Proc. IUI*, 159-166.
4. Bruner, J.S. and Tagiuri, R. 1954. The Perception of People. *Handbook of Social Psychology*. G. Lindzey, ed. Addison-Wesley.
5. Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., and Pierce, L.G. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies*. 58, 6 (Jun. 2003), 697–718.
6. Glass, A., McGuinness, D. and Wolverton, M. 2008. Toward establishing trust in adaptive agents. *Proc. IUI*, 227-236.
7. Johnson-Laird, P.N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press.
8. Kapoor, A., Lee, B., Tan, D. and Horvitz, E. 2010. Interactive optimization for steering machine classification. *Proc. CHI*, 1343-1352.
9. Kulesza, T., Burnett, M., Stumpf, S., Wong, W., Das, S., Groce, A., Shinsel, A., Bice, F. and McIntosh, K. 2011. Where Are My Intelligent Assistant's Mistakes? A Systematic Testing Approach. *Proc. IS-EUD*, 171–186.
10. Kulesza, T., Stumpf, S., Burnett, M., Wong, W.-K., Riche, Y., Moore, T., Oberst, I., Shinsel, A. and McIntosh, K. 2010. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. *Proc. VL/HCC* (2010), 41–48.
11. Kulesza, T., Stumpf, S., Wong, W.-K., Burnett, M., Perona, S., Ko, A. and Obsert, I. 2011. Why-Oriented End-User Debugging of Naive Bayes Text Classification. *ACM Transactions on Interactive Intelligent Systems*. 1, 1 (Oct. 2011).
12. Kulesza, T., Wong, W.-K., Stumpf, S., Perona, S., White, R., Burnett, M., Oberst, I. and Ko, A. 2009. Fixing the program my computer learned: barriers for end users, challenges for the machine. *Proc. IUI*.
13. Lim, B. and Dey, A. 2010. Toolkit to support intelligibility in context-aware applications. *Proc. Ubicomp*, 13-22.
14. Lim, B., Dey, A. and Avrahami, D. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Proc. CHI*, 2119-2128.
15. Norman, D. 1983. Some Observations on Mental Models. *Mental Models*. D. Gentner and A. Stevens, eds. Psychology Press.
16. Rosson, M. and Carrol, J. 1990. Smalltalk scaffolding: a case study of minimalist instruction. *Proc. CHI*, 423-429.
17. Settles, B. 2009. *Active learning literature survey*. University of Wisconsin-Madison.
18. Sharp, H., Rogers, Y. and Preece, J. 2007. *Interaction Design*. John Wiley & Sons Inc.
19. Stumpf, S., Rajaram, V., Li, L., Wong, W., Burnett, M., Dietterich, T., Sullivan, E. and Herlocker, J. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*. 67, 8 (Aug. 2009), 639–662.
20. Talbot, J., Lee, B., Kapoor, A. and Tan, D. 2009. EnsembleMatrix: Interactive visualization to support machine learning with multiple classifiers. *Proc. CHI*, 1283-1292.
21. Tullio, J., Dey, A., Chalecki, J. and Fogarty, J. 2007. How it works: a field study of non-technical users interacting with an intelligent system. *Proc. CHI*, 31-40.
22. Vig, J., Sen, S. and Riedl, J. 2009. Tagsplanations: explaining recommendations using tags. *Proc IUI*, 47-56.