

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/221607740>

Toward establishing trust in adaptive agents

Conference Paper · January 2008

DOI: 10.1145/1378773.1378804 · Source: DBLP

CITATIONS

154

READS

832

3 authors, including:



[Alyssa Glass](#)

College of William and Mary

10 PUBLICATIONS 222 CITATIONS

[SEE PROFILE](#)



[Deborah L McGuinness](#)

Rensselaer Polytechnic Institute

458 PUBLICATIONS 31,911 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Health Web Science [View project](#)



W3C PROV-WG [View project](#)

Toward Establishing Trust in Adaptive Agents

Alyssa Glass

Knowledge Systems, AI Lab
Stanford University
Stanford, CA 94305 USA
glass@ksl.stanford.edu

Deborah L. McGuinness

Knowledge Systems, AI Lab
Stanford University
Stanford, CA 94305 USA
dlm@ksl.stanford.edu

Michael Wolverton

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025 USA
mjw@ai.sri.com

ABSTRACT

As adaptive agents become more complex and take increasing autonomy in their user's lives, it becomes more important for users to trust and understand these agents. Little work has been done, however, to study what factors influence the level of trust users are willing to place in these agents. Without trust in the actions and results produced by these agents, their use and adoption as trusted assistants and partners will be severely limited. We present the results of a study among test users of CALO, one such complex adaptive agent system, to investigate themes surrounding trust and understandability. We identify and discuss eight major themes that significantly impact user trust in complex systems. We further provide guidelines for the design of trustable adaptive agents. Based on our analysis of these results, we conclude that the availability of explanation capabilities in these agents can address the majority of trust concerns identified by users.

Author Keywords

Trust, explanation, user study, automated assistants, complex agents, adaptive agents, evaluation.

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces – Evaluation; Prototyping; Training, help, and documentation. I.2.1 [Artificial Intelligence]: Applications and Expert Systems – Office automation. H.4.1 [Information Systems Applications]: Office Automation.

INTRODUCTION

Adaptive agents and intelligent assistant systems are becoming increasingly complex (for instance, [5, 6, 12, 19, 22]). They often include highly distributed reasoning systems such as task processors, hybrid theorem provers, and probabilistic inference engines; multiple learning components employing a wide range of logical and

statistical techniques; and multiple heterogeneous, distributed information sources underlying the processing. Despite the sophistication, however, they typically provide little transparency into the computation and reasoning being performed.

At the same time as these systems are becoming more complex, they are also taking more autonomous control. They are being asked to assist user actions, but also to act autonomously on behalf of their users. The DARPA Personalized Assistant that Learns (PAL) program [20] describes such agents as being able to “reason, learn from experience, be told what to do, explain what they are doing, reflect on their experience, and respond robustly to surprise.”

As researchers build these systems to plan for the achievement of abstract objectives, execute tasks, anticipate future needs, aggregate multiple sensors and information sources, and adapt its behavior over time, there is an underlying assumption that there will be a user in the loop whom the agent is serving. This user would need to understand the agent's behavior and responses enough to participate in the mixed-initiative execution process and to adjust the autonomy inherent in such systems. The user would also need to trust the reasoning and actions performed by the agent. Few have considered, however, how the user would interact with the agent, and what requirements may exist in order for a user to trust and rely on such a complex system, particularly when the underlying knowledge, behavior, and assumptions of the system are constantly changing and adapting through the use of machine learning.

To better understand the factors that influence the trust and understanding of these adaptive agents, we conducted a trust study among a set of users of one of these agents. We used as the basis of our study the Cognitive Assistant that Learns and Organizes (CALO) system [5], an adaptive personalized assistant that performs a wide range of office-related tasks involving calendars, address books, email, documents, and the Web. The system includes several interlinked components providing a range of capabilities, the majority of which include various methods of machine learning. The CALO system used by the participants in our study was in an advanced research prototype phase, and the participants were asked to be tolerant of typical prototype

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IUT08, January 13-16, 2008, Maspalomas, Gran Canaria, Spain.
Copyright 2008 ACM 978-1-59593-987-6/08/0001 \$5.00

issues like software bugs and system crashes. We further asked participants to look beyond such problems to begin to understand what it means for users to trust and rely on such a system in their daily life.

Previous work investigating trust representation, reasoning, and presentation in complex systems [16, 28] has revealed complexities related to understanding the notion of trust. This work also has shown how explanations of reasoning can help users to establish and build trust. Our study — which consisted of structured interviews with a variety of CALO users and analysis of their answers — shows that explanation of adaptive agents, particularly in the presence of learned behaviors and information, can similarly be key to addressing user concerns about understandability, and to helping users to build trust in these complex assistants.

In this paper, we present the method and structure of a study for examining these issues; we describe the themes identified in the participant responses, and the implications of these themes on the practical use of adaptive assistants; and we discuss recommendations for how to build adaptive agents that will be trusted and used by everyday users. Our study's primary contribution is the identification of these trustability themes, and the resulting guidelines that these themes imply for designers of adaptive agents.

TRUST STUDY

We conducted a structured, qualitative trust study to do the following:

- identify what factors users believe influence their trust in complex adaptive agents;
- identify which types of questions, if any, users would like to be able to ask an adaptive assistant, to better understand the assistant's answers and behavior;
- evaluate the general usability of adaptive agents.

Procedure

Our study was conducted in two basic stages: the usage stage and the interview stage. For the usage stage, we piggy-backed on a broad study aimed at testing the learning capabilities within the CALO system. The broad study is part of a long term effort involving a large set of testers and researchers, extensive participant training for the use of various CALO components, and detailed analysis of complex data logs and learning algorithms. The focus of the broad study is to gather quantitative data about capabilities and learning improvements in the complete system and is done through intensive system use by dedicated users over approximately two weeks.

During the usage stage, each study participant used an individual CALO agent to assist in performing a variety of office tasks. Participants were given tasks to accomplish with their agents, such as scheduling meetings, preparing documents, or planning conference travel; for some tasks, users were given detailed scripts outlining how to use the agent to assist with the task. Participants were told that the

primary purpose of their CALO usage during this time was for the experimenters to collect data on the agent's behavior, and thus they were required to complete as many tasks as they could during the test period. Participants typically used the system for a full eight hour work day, each day, for the entire duration of the test period.

During the interview stage of our trust study, we interviewed each participant after the end of the usage period. Thirteen participants were interviewed within two days of the end of the test period; one participant was interviewed one week later. The interviews were structured to follow a fixed script for all participants. The script contained 40 questions — eleven questions using a typical 5-step Likert scale (from “hardly ever” to “extremely often”) and 29 open response questions. When individual participants provided detailed answers to particular portions of the script, we adjusted the interview to follow up on these answers in more detail. The script was organized around five main topics: failure, surprise, confusion, question-answering, and trust. Each interview was audio recorded. We used these recordings to make notes on the interviews and to organize the responses into common themes.

Participants were informed before the usage period that they would be interviewed about system failures, confusion, usability, and trust at the end of the usage period. A few of the participants took notes on these issues during the test period, which they referred to during our interviews. Two sets of users participated in the study. The first, main set of participants used CALO for approximately two weeks. A second, smaller set of participants used CALO for approximately one day, using a similar usage format to the first set, but concentrating on a small subset of the total system capabilities.

Agent Overview

The CALO system used by the participants in our study provided capabilities for a wide range of office-related tasks [19], including maintaining calendars and schedules of meetings [2]; managing contact information; scanning and sorting email and other documents [7]; performing Web searches; scraping information from the Web; helping to prepare new documents and presentations; managing tasks [8]; purchasing new equipment; planning travel; and learning new procedures for previously unknown tasks [1, 3].

Typical tasks the participants performed with the help of their agents included scheduling mutually convenient meetings with groups of people; planning detailed travel arrangements to attend conferences; and teaching their agents how to independently find and store contact information for colleagues using Web searches and scraping. In many cases, participants were provided with guidance about how best to use CALO to perform the tasks. In other cases, participants were free to use any portion of

the CALO system that they felt would best enable them to accomplish specific goals.

The main objective for the CALO project is to explore the use of machine learning techniques as applied in robust, complex assistant agents capable of reasoning, execution, explanation, and self-reflection. Questions of usability, though important when building a deployed personal assistant, were not central research questions for the project.

Participants

The longer, two-week study had ten participants, and the shorter study had four participants. All participants from both studies were employees of SRI International, twelve men and two women. Participants spanned a wide range of ages, falling approximately evenly into five decade-long age brackets. All participants had at least a bachelor's degree, and just over one third (five participants) held doctoral degrees. Most, but not all, had studied engineering-related fields.

The participants' prior experience with the CALO system varied widely. Several had helped to develop individual components of the system. Others were managers or researchers contributing to individual aspects of the system. A minority of participants had no development or research association with the system.

STUDY FINDINGS

After the interviews were completed, we grouped similar comments from multiple users into topics, and discarded topics that only a few participants commented on. The remaining topic areas are discussed below clustered into eight themes. While not all participants commented on all themes, each of the themes were significant to a majority of the participants. We further loosely categorize these themes into discussions of usability, explanation requirements, and trust, though we maintain that these three topics are often deeply intertwined with each other.

Usability

Theme 1: High-Level Usability of Complex Prototypes. The most common usability comments concerned system performance and processing time; and the inability to “undo” or reverse actions. Other comments involved well-understood principles of human-computer interaction (for example, [11, 23]). While we were aware that these usability problems existed in the system, and addressing these issues was certainly not a central focus for the CALO project, we were nevertheless interested to discover the degree to which these research-oriented users were overwhelmingly frustrated by usability concerns even when they admit that they had very low usability expectations for a prototype-stage system.

Participant comments on the general usability of the system confirmed that there is a limit to how much cutting-edge systems are able to ignore usability issues while focusing on more advanced features. While the users we interviewed

were aware that the system was at the prototype stage, and they were generally sympathetic to its status as a research-quality system, the underlying usability problems were still able to block the acceptance of cutting-edge capabilities. Even when the advanced features were able to provide functionality that the users desperately wanted to use, they were unable to accept the software because of the usability issues. One user commented about the usability problems, “I can’t tell you how much I would love to have [the system], but I also can’t tell you how much I can’t stand it.” Nonetheless, this first theme can be distinguished from the remainder of the themes discussed in this paper by the degree to which it is effected by the prototype-stage condition of the system.

Theme 2: Being Ignored. Many participants complained about feeling ignored by the agent. After providing the system with personal preferences, as well as suggestions and feedback aimed at improving machine learning, many users were left with the impression that their effort was wasted and that the agent was ignoring them. Users complained that the agent was “not paying attention” during interactions. One user said, “You specify something, and [the system] comes up with something completely different, and you’re like, it’s ignoring what I want!”

Additionally, several users commented that the behavior exhibited by the system would have been fine for non-adaptive systems, but once they formed the expectation that the system was able to accept user guidance and obey preferences, they felt somehow betrayed when the system failed to do so. We return to this theme in our discussion of expectations.

Explanation Requirements

Theme 3: Context-Sensitive Questions. To investigate the value of different types of explanations to user needs, we asked our users to rate a list of question types according to how often they would have utilized questions of that type if an answer to it had been available during the test.

We used Silveira *et al.*'s “Taxonomy of Users’ Frequent Doubts” [24], an enumeration of user information needs, as our list of candidate question types, and each was ranked by each user on a Likert scale from 1 (“would never want to ask”) to 5 (“would want to ask extremely often”). We averaged the ratings to produce an overall score of usefulness for each question. The results are shown in Figure 1. The question types, along with sample questions of each type, in preference order were:

1. Choice (What can I do right now?)
2. Procedural (How do I do this?)
3. Informative (What kinds of tasks can I accomplish?)
4. Interpretive (What is happening now? Why did it happen?)
5. Guidance (What *should* I do now?)

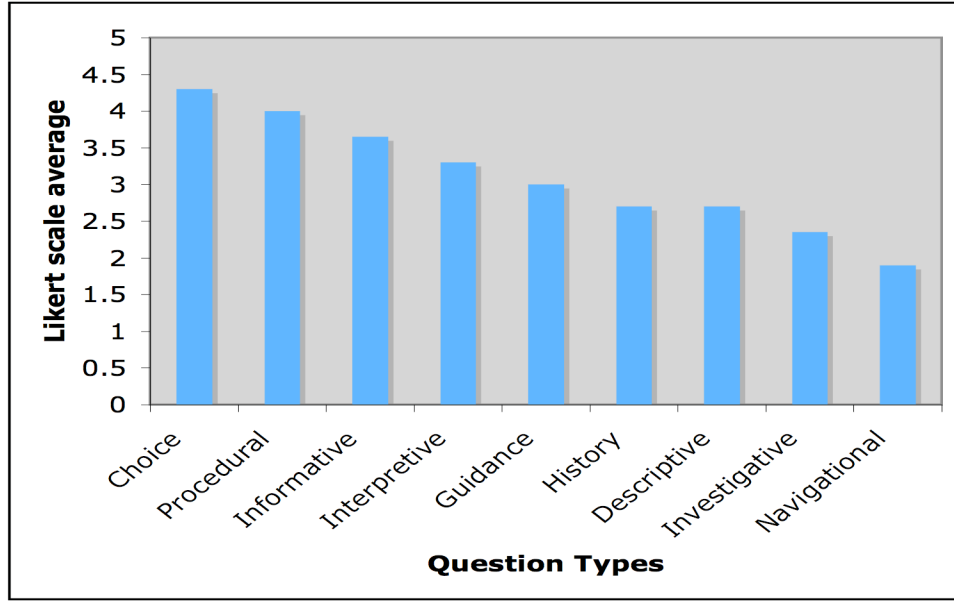


Figure 1: Average Likert scale responses for question types.

6. History (What have I already done?)
7. Descriptive (What does this do?)
8. Investigative (Did I miss anything?)
9. Navigational (Where am I?)

We note that questions can generally be divided into two categories. *Context-independent questions* have answers that are not dependent on the context in which they are asked; these questions can generally be addressed by standard help systems or simple mouse-over pop-up text boxes. *Context-sensitive questions* have answers that require the system to consider what is currently happening (“task sensitivity”) or high-level goals inferred from the user (“user-intent sensitivity”).

Of the question types presented, five of them had average scores of 3.0 or higher. Of these five question types, the Choice, Procedural, and Informative questions are context-independent and could be supported in a software system through the use of easily-accessed documentation. The other two top question types, the Interpretive and Guidance questions, are context-sensitive, and point to the need for more complex explanation capabilities to address these user needs.

In addition to rating the standard taxonomy of questions, we also asked our users to identify on their own the questions they most wanted to ask.¹ The most common questions

identified by our users as being potentially helpful to them during their use of the system were:

1. What are you doing right now?
2. Why did you do that?
3. When will you be finished?
4. What information sources did you use?

Of the 34 questions mentioned by our participants, 16 of them (47%) were variations on these four questions.

We also note two final observations about the explanation requirements identified by the participants. First, we note that the questions most often identified by the participants *before* being presented with the taxonomy of question types are entirely context-sensitive; the first three are Interpretive questions, and the final question (about information provenance) does not easily fit into the taxonomy, perhaps because of the relative complexity of CALO as compared to the systems discussed in [24]. We conclude that the majority of the confusion encountered by the participants cannot be solved with the use of simple help availability or better documentation, but rather requires a deeper solution.

Second, we were surprised by the reaction of our participants when presented with the question types from the taxonomy. Common reactions included comments like “I would love to ask that!”, “That’s a cool [question]... I’d use that if it existed!”, and “I was asking that [to myself] all day!” The majority of our participants expressed that these were questions that they would definitely want to ask and receive answers to, but it would not generally occur to them to ask the questions of the system, because they do not

¹ In the study, users provided their free-form questions before being presented with the taxonomy, to prevent being influenced by the standard question types.

expect computer systems to be able to provide useful answers. We contend, however, that these questions can be solved by context-sensitive explanation systems, and conclude that such systems would need to provide users with initial questions, so that these doubtful users are made aware of the supported question types.

Theme 4: Granularity of Feedback. When the agent did provide feedback to the participants, many of them commented that the feedback was at the wrong level of detail for their needs. Several of the agent components provided simple status messages indicating simply “Okay” or “Not Okay.” Participants found this type of feedback frustrating; when the status was “Not Okay,” they wanted additional feedback to explore details about the cause of the problem, and possible solutions. Participants commented, “[The component]² would say ‘I’m confused’ and there was no other feedback,” and “I definitely wanted to know WHY!” Several participants in particular mentioned that, when the status was “Not Okay,” the lack of detail prevented them from identifying whether they themselves had caused the error through their actions, or whether there was a problem with the system. Lacking more detailed feedback, they were unable to fix the problem, nor to avoid it in the future.

Equally frustrating to many participants were other system components that provided an overwhelming amount of feedback. The constant stream of status information from these components was so frequent and cumbersome that most users found them to be unhelpful even when there was a problem. One participant noted that, despite the large number of status messages from these components, he still “wasn’t getting answers,” and another said that he wanted to ask the system to just “explain to me what you think the current state of things are” in a simple and straightforward manner.

We expect that context modeling, to identify when additional detail would be useful, will help with the problem of identifying the proper level of granularity to use in different situations. We also expect that user modeling will help to adapt the granularity of feedback to the needs of different users.

Building Trust

Theme 5: Transparency. When asked what would help them to build trust in the system, the first thing most participants (71%) mentioned was transparency, and every participant (100%) mentioned transparency as a major factor affecting overall usability. Participants complained that the system was “too opaque” and “needs to be more comprehensible.” Several users noted that the components

of the system that they trusted the most were the ones that provided feedback about what they were doing, and that when actions were taken by the system for which the underlying computational reasoning was not apparent, the users mistrusted the results. One user commented that “the ability to check up on the system, ask it questions, get transparency to verify what it is doing, is the number one thing that would make me want to use it.”

Even when results appeared reasonable, they sought to verify the results rather than trusting them outright, fearful that a result may be coincidental or based on inappropriate information. These users identified explanations of system behavior, providing transparency into its reasoning and execution, as a key way of understanding answers and thus establishing trust.

We note as well that transparency is particularly useful in building trust in a system for which a baseline of trust does not already exist. In systems that are either widely used by others, or in systems that have already been used by the user without incident, the user may already have a base level of trust, regardless of the transparency revealed by the system. In contrast, system transparency can provide a building block on which to establish trust in systems for which no other basis already exists, as is often the case with new systems or inexperienced users, as in our study.

Theme 6: Provenance. Access to knowledge provenance (sources used to provide information, as well as meta-information about those sources) was also mentioned by many participants. Many users commented that knowing what resources were being used to provide answers, particularly when the system was learning new information based on these sources, would aid them in trusting the system. Several users reported that explanations of knowledge provenance would enable them to trust results without the need for extensive further verification. One user commented that, “in general, I wanted information about the source,” and another user said that “[the system] needs a better way to have a meta-conversation.”

We also found, somewhat surprisingly, that providing access to knowledge provenance would increase trust not only in the answers provided by the system, but also in the reasoning of the entire system itself. Most participants, when presented by the system with a perceived incorrect answer or conclusion, assumed that the cause of the error was flawed reasoning or logic deep in the system. Because these participants had an underlying assumption that “fixing” a reasoning problem would be difficult, their trust in the system was greatly eroded with even just one incorrect answer. For instance, one user complained, “You can’t talk to it! You can’t say ‘Why **didn’t** you learn something?’ It’s just a big black hole.”

These users, on the other hand, tended *not* to blame incorrect answers on errors in the data, or even a lack of sufficient data, as was often the case with statistical machine learning components in the system. These “data-

² The CALO system consists of a number of loosely integrated assistant technologies. Some user comments referred specifically to one of these components, rather than CALO as a whole.

driven” errors, however, are often easy to fix, and when errors could be properly identified as being data-driven rather than logic-driven, the users’ trust in the system as a whole was better maintained.

Theme 7: Managing Expectations. The majority of the participants had prior notions of what should be easy or hard for the system to do. When describing their interactions with the system that were confusing or frustrating, or times when they considered abandoning the system, almost every participant mentioned expectation mismatches of some kind.

Participants indicated that these times of expectation mismatches often directly led to a decrease in trust, particularly when something that they felt should have been easy suddenly appeared to be hard. One user commented, “Normal interactions with [office tools] don’t leave you with a lot of expectations, so I was always sort of wondering why [the test system] was spending so much time and seemed to be spinning its wheels.”

Discussing these moments with the participants, it became apparent that often the system was trying to do something complex, like adapt to a changing situation on the fly, or integrate many new sensory inputs into a large dataset, which were happening mostly in the background and thus were not apparent to the user. Because the goals of the system centered on its ability to use learning in a variety of situations to assist with tasks, the system often tried to perform tasks in an intelligent, adaptive way; however, because similar tasks could often be accomplished in a basic, non-adaptive way by a “dumb” system, the participants expected these tasks to be easy.

While the participants expected “easy” tasks to be accomplished quickly even by a complex, constantly adapting agent, their expectations were not because the participants did not understand the nature of the adaptation. To the contrary, their sometimes limited knowledge of what the agent was capable of learning led to expectation mismatches in the other direction as well; many complained that, once given the expectation that the agent would adapt to their desires, they became frustrated that the system was not adapting quickly enough. One participant said, “You would think that you could infer that something would work, but then that was not the case.” When the system would behave in a way typically associated with (and accepted in) non-adaptive systems, the participants would get increasingly upset that the system was not quickly getting better.

Other mismatches occurred when participants attempted to try something new, and discovered that they did not understand enough about what the system was doing to be able to complete their task. One user, on discovering that completing their goal was more complicated than expected, commented, “I was paralyzed with fear about what it would understand and what it would not.” Another simply concluded, “I had a misunderstanding of its capabilities.”

Theme 8: Autonomy and Verification. Though not often stated directly, most participants adopted a “trust but verify” approach to using the system. When asked how often they felt that they trusted the system, most participants responded that they trusted the system 25 to 60 percent of the time. When these responses were further investigated, however, it became clear that almost all participants actually meant that they *would* trust the system this often, but *only if* they were given mechanisms to verify the responses and override erroneous behavior when necessary. Typical participants said that they trusted the system when it “wasn’t too autonomous,” when the system performed “with supervision,” and when they could “check up on” the system.

Participants also were extremely reluctant to trust the system to perform any task that changed the world (for instance, making a purchase with a credit card, as opposed to simply looking up or calculating an answer). One user noted, “I trust [the system’s] accuracy, but not its judgment.”

In addition to the difficulty of verifying correct behavior for these world-changing tasks, many participants noted that they would only trust the system to perform these tasks after closely observing the system perform small parts of the task on its own, in a mixed-initiative manner. Participants noted that “trust is an earned property” that the system would only earn when its behavior has been verified.

DISCUSSION AND RELATED WORK

For simplicity of discussion, we list the eight identified themes here:

T1: High-Level Usability of Complex Prototypes

T2: Being Ignored

T3: Context-Sensitive Questions

T4: Granularity of Feedback

T5: Transparency

T6: Provenance

T7: Managing Expectations

T8: Autonomy and Verification

We are interested in how explanation systems can address issues of user trust in complex adaptive agents. We observe that an explanation system that provides context-sensitive explanations of adaptive agents (for instance, as in [17]) is capable of addressing the concerns of five of these eight themes:

T3: by providing the user with the ability to ask context-sensitive questions.

T4: by intelligently modulating the granularity of feedback based on context- and user-modeling.

T5: by supplying the user with access to information about the internal workings of the system.

- T6:** by providing the user with the information provenance for sources used by the system.
- T8:** by enabling the user to verify the autonomous steps taken by the system.

In addition, we note that the information provided by explanation systems can help to partially address the concerns of two more of these themes:

- T2:** by providing justifications for the actions taken by the system, particularly when they are contrary to the expectations of the user.
- T7:** by explaining to the user the capabilities and limitations of the system.

The final theme, T1, can only be addressed by diligent design and attention to user interaction principles throughout the prototype development stage.

Even without the use of an integrated explanation system, these themes suggest guidelines for the developers of all complex, adaptive agents. Computer users are increasingly interacting with systems that they do not fully understand. As these systems become increasingly complex, and users are expected to trust them with a wider range of responsibilities, the developers of these systems need to become aware of the trust requirements they are placing on users.

Our findings show that users have specific requirements in terms of transparency and verification that they expect from such systems before they are willing to trust the outputs and actions of the system. In addition, as these systems become more intelligent and adaptive, they must increasingly support the ability to have a “dialogue” or “conversation” with users, to provide them with reassurances about the system’s understanding of the user and the world. Several theoretical frameworks for modeling these dialogues have

been suggested; Walton [26] compares several of these models and suggests one such model that is particularly useful for modeling explanation dialogues of the type that would address the themes we identified here.

Recommendations for addressing each of the themes are summarized in Figure 2. In this figure we identified a technology as addressing a theme only if the theme is a core capability of that technology (as opposed to a rarely-addressed side issue). The intention is not to minimize the importance of Traditional Help Systems or UI Design — those technologies address many additional issues that are beyond the scope of this paper — but rather to demonstrate how explanation is central to the themes identified by the users in our study. We believe that a combination of all three technologies is necessary to fully address trust issues in complex agents.

In the HCI community, some previous work (for example, [14]) has taken a high-level view of what it would take to make an intelligent system usable. The guidelines that we present here build on the broad themes of transparency and trust mentioned in [14] and provide more grounded, concrete guidance for designing interfaces for adaptive agents.

We also note that trust can be built in multiple ways. Some systems that do not follow these guidelines do eventually earn user trust, generally through reputation-based methods such as trusted recommendations from others, or through (often forced) personal usage over long periods of time. In many cases, however, an existing user base is not available to provide “references” for new software, and users are often reluctant to spend long periods of time using systems that they do not already trust. Thus, these themes can be thought of as “short-cuts” to building trust. For new systems, or for systems as complex as CALO, providing these trust short-cuts can establish a base level of trust that

		Traditional Help System	Complex Explanation	UI Design
T1	High-Level Usability			X
T2	Being Ignored		X	
T3	Context-Sensitive Questions		X	
T4	Granularity of Feedback		X	
T5	Transparency	X	X	X
T6	Provenance		X	
T7	Managing Expectations	X	X	X
T8	Autonomy & Verification		X	

Figure 2: Summary of themes and possible solutions.

keeps users involved until longer, reputation-based trust can be established, or when these alternative sources of trust are not available. Our study was focused on identifying these methods for building trust in the absence of long-term methods. This approach is consistent with previous work, such as [27], on building initial trust in on-line e-commerce recommendation agents. Longitudinal studies that consider how user trust changes over time are needed to understand how reputation-based trust can also contribute to overall user trust.

Though there is increasing interest in building complex adaptive systems, and much has been written about the algorithms and components that comprise these systems, little work has been done to evaluate their user acceptance. Cortellessa & Cesta [9] discuss this lack of research on what they call the “quality of interaction,” and provide results of an initial study focused on user trust and the use of explanation in mixed-initiative planning systems. In this particular domain, they found that the use of explanations was highly correlated with the number of failures experienced by users. They additionally found that novice users were more hesitant to trust the system, employing more mixed-initiative strategies than expert users, who trusted the system to perform more completely automated tasks.

Other studies of mixed-initiative systems, such as [4, 10, 13] have looked specifically at the use of mixed-initiative adaptation for customizing GUI components, reporting on how the adaptation influences the performance time of various tasks and overall satisfaction with the interface. While our study is less precise in measurement, it differs in scope. It attempts to evaluate overall user experience with a focus on trust in systems answers, and to identify high-level design principles for agents using adaptation not just in the user interface, but for the underlying processing of the agent itself. Most relevant for our study, Bunt *et al.* [4] provides for, but does not evaluate, a mechanism for simple explanations aimed at maintaining transparency. Our study extends that work to further understand how such an explanation mechanism can influence user trust in a broader range of adaptive systems.

User studies focused solely on understanding machine learning [21, 25] have looked at how explanations can increase acceptance and usability of these learning algorithms in isolation, by testing user understanding of a variety of machine learning algorithms when explanations of various forms are available. In this community as well, however, we are not aware of studies that have looked at these issues when the machine learning is integrated into larger hybrid systems.

Turning purely to issues of trust as one important aspect of acceptance and usability, Huang and Fox [15] provide a detailed definition and study of trust. They define the concept as “the psychological state comprising (1) *expectancy*: the truster expects a specific behavior of the

trustee such as providing valid information or effectively performing cooperative actions; (2) *belief*: the truster believes that expectancy is true, based on evidence of the trustee’s competence and goodwill; (3) *willingness to be vulnerable*: the truster is willing to be vulnerable to that belief in a specific context where the information is used or the actions are applied.” Their definition is consistent with our user interviews, and provides us with a framework in which to evaluate when user trust has been achieved.

Several issues remain open for investigation. We plan to implement the recommendations identified in this study by expanding ICEE (Integrated Cognitive Explanation Environment) [18], our complex explanation framework consistent with the model suggested in [26].

We also recognize the value in two different types of user studies: those aimed at guiding design and identifying design recommendations (like the one reported in this paper), and those that focus on evaluating an existing implementation. Thus, we plan a follow-on evaluative user study of a complex agent with explanation capabilities, both to investigate the effectiveness of the explanations and to verify the ability of the explanations to address the themes identified in this paper. In this way, we hope to measure both user satisfaction and user effectiveness in systems with and without explanation capabilities. We also plan to study how explanations can be used to direct learning in complex agents.

CONCLUSIONS

We have studied issues governing the trust and usability of complex adaptive agents. Without trust in the actions and results produced by these agents, they will not be used and widely adopted as assistants and partners. By interviewing users of these agents, we have identified several themes that describe the willingness of users to adopt and trust these agents, particularly as the agents employ increased autonomy to perform tasks and make decisions on behalf of their users.

Our study’s primary contribution is the identification of these themes, and the resulting guidelines for designers of adaptive agents. With these guidelines in mind, we show how the use of complex explanation systems can address the majority of the trust concerns mentioned by users, and thus can help to move adaptive assistants one step closer to use and acceptance by end users. These recommendations lay the groundwork for future work on evaluating how users interact with adaptive agents.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under contract #55-300000680 to-2 R2. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior – National Business Center (DOI-NBC). We thank

Karen Myers and the anonymous reviewers for their helpful comments and suggestions on earlier versions of this paper. We additionally thank Elizabeth Furtado and Aaron Spaulding for helpful suggestions on our structured interview format. Finally, we thank our study participants for their time and input.

REFERENCES

- Allen, J., Chambers, N., Ferguson, G., Galescu, L., Jung, H., Swift, M., and Taysom, W. PLOW: A Collaborative Task Learning Agent. *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, 2007.
- Berry, P., Conley, K., Gervasio, M., Peintner, B., Uribe, T., Yorke-Smith, N. Deploying a Personalized Time Management Agent, *Proceedings of AAMAS'06*, 2006.
- Blythe, J. Task Learning by Instruction in Tailor, *Proceedings of the International Conference on Intelligent User Interfaces*, 2005.
- Bunt, A., Conati, C., and McGrenere, J. Supporting Interface Customization using a Mixed-Initiative Approach. *Proceedings of the International Conference on Intelligent User Interfaces (IUI-2007)*, ACM Press, 2007.
- CALO, 2007. <http://www.ai.sri.com/project/CALO>
- Chalupsky, H., Gil, Y., Knoblock, C., Lerman, K., Oh, J., Pynadath, D., Russ, T., and Tambe, M. Electric Elves: Applying Agent Technology to Support Human Organizations. *AI Magazine* 23(2), 2002.
- Cheyen, A., Park, J., and Giuli, R. IRIS: Integrate. Relate. Infer. Share. *1st Workshop on the Semantic Desktop*, International Semantic Web Conference (ISWC'05), 2005.
- Conley, K. and Carpenter, J. Towel: Towards an Intelligent To-Do List, *Proceedings of the AAAI Spring Symposium on Interaction Challenges for Artificial Assistants*, 2007.
- Cortellessa, G. and Cesta, A. Evaluating Mixed-Initiative Systems: An Experimental Approach. *ICAPS-06*, 2006.
- Debevc, M., Meyer, B., Donlagic, D., and Svecko, R. Design and Evaluation of an Adaptive Icon Toolbar. *User Modeling and User-Adapted Interaction* 6(1), 1996.
- Dix, A., Finlay, J.E., Abowd, G.D., and Beale, R. *Human-Computer Interaction*, Third Edition, Prentice Hall, 2003.
- Ferguson, G. and Allen, J. Mixed-Initiative Systems for Collaborative Problem Solving. *AI Magazine* 28(2), 2007.
- Gajos, K. Z., Czerwinski, M., Tan, D.S., and Weld, D.S. Exploring the Design Space for Adaptive Graphical User Interfaces. *Conference on Advanced Visual Interfaces (AVI '06)*, ACM Press, 2006.
- Höök, K. Steps To Take Before Intelligent User Interfaces Become Real. *Interacting with Computers*, 12(4), 2000.
- Huang, J. and Fox, M.S. "An ontology of trust: formal semantics and transitivity," in *ICEC '06: Proceedings of the 8th international conference on Electronic Commerce*, pp. 259–270, ACM Press, 2006.
- McGuinness, D.L., Zeng, H., Pinheiro da Silva, P., Ding, L., Narayanan, D., and Bhaowal, M. Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study. *WWW2006 Workshop on the Models of Trust for the Web (MTW'06)*, 2006.
- McGuinness, D.L., Glass, A., Wolverton, M., and Pinheiro da Silva, P. A Categorization of Explanation Questions for Task Processing Systems. *AAAI Workshop on Explanation-Aware Computing (ExaCt-07)*, 2007.
- McGuinness, D.L., Glass, A., Wolverton, M., and Pinheiro da Silva, P. Explaining Task Processing in Cognitive Assistants that Learn. *Proceedings of the 20th International FLAIRS Conference (FLAIRS-20)*, 2007.
- Myers, K., Berry, P., Blythe, J., Conley, K., Gervasio, M., McGuinness, D., Morley, D., Pfeffer, A., Pollack, M., and Tambe, M. An Intelligent Personal Assistant for Task and Time Management. *AI Magazine* 28(2), 2007.
- PAL, 2007. <http://www.darpa.mil/ipto/programs/pal/>
- Pazzani, M.J. Representation of Electronic Mail Filtering Profiles: A User Study. *Proceedings of the 5th International Conference on Intelligent User Interfaces (IUI-2000)*, ACM, 2000.
- Rich, C. and Sidner, C. DiamondHelp: A Generic Collaborative Task Guidance System. *AI Magazine* 28(2), 2007.
- Shneiderman, B. and Plaisant, C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*, Fourth Edition, Addison Wesley, 2004.
- Silveira, M.S., de Souza, C.S., and Barbosa, S.D.J. Semiotic Engineering Contributions for Designing Online Help Systems. *SIGDOC'01*, Association for Computing Machinery, 2001.
- Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., Drummond, R., and Herlocker, J. Toward Harnessing User Feedback for Machine Learning. *Proceedings of the International Conference on Intelligent User Interfaces (IUI-2007)*, ACM Press, 2007.
- Walton, D. Dialogical Models of Explanation. *AAAI Workshop on Explanation-Aware Computing (ExaCt-07)*, 2007.

27. Wang, W. and Benbasat, I. Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs. *Journal of Management Information Systems* 23(4), 2007.
28. Zaihrayeu, I., Pinheiro da Silva, P., and McGuinness, D.L. 2005. IWTrust: Improving User Trust in Answers from the Web. *Proceedings of the 3rd International Conference on Trust Management (iTrust2005)*, Springer, pp. 384-392.