How it Works: A Field Study of Non-Technical Users Interacting with an Intelligent System

Joe Tullio
Applications Research Center
Motorola Labs
Schaumburg, IL 60196
jtullio@acm.org

Anind K. Dey, Jason Chalecki†
Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213
anind@cs.cmu.edu

James Fogarty
Computer Science & Engineering
University of Washington
Seattle, WA 98195
jfogarty@cs.washington.edu

ABSTRACT
In order to develop intelligent systems that attain the trust of their users, it is important to understand how users perceive such systems and develop those perceptions over time. We present an investigation into how users come to understand an intelligent system as they use it in their daily work. During a six-week field study, we interviewed eight office workers regarding the operation of a system that predicted their managers' interruptibility, comparing their mental models to the actual system model. Our results show that by the end of the study, participants were able to discount some of their initial misconceptions about what information the system used for reasoning about interruptibility. However, the overarching structures of their mental models stayed relatively stable over the course of the study. Lastly, we found that participants were able to give lay descriptions attributing simple machine learning concepts to the system despite their lack of technical knowledge. Our findings suggest an appropriate level of feedback for user interfaces of intelligent systems, provide a baseline level of complexity for user understanding, and highlight the challenges of making users aware of sensed inputs for such systems.

Author Keywords
Intelligent systems, context-aware, mental models, qualitative research, machine learning, field study.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous. I2.m. Artificial Intelligence: Miscellaneous.

INTRODUCTION
Intelligent systems continue to find their way into everyday applications. These systems gather information about their users, reason from what they have learned in the past, and generate predictions or decisions based on this reasoning. Examples include spam filtering applications and product recommenders. While these applications are quite useful, intelligent systems promise to do much more in terms of automating tasks and increasing awareness by inferring the states of users and environments. Such applications demand a higher degree of trust from users before they are willing to delegate important decisions or personal information to a software system [8, 17]. This trust comes from an ability to predict the system's behavior through observation [20].

To predict and explain the behavior of a system, people construct mental models that may be more or less complete and accurate [22]. Therefore, designers must create intelligent applications that enable the formation of mental models that are predictable enough to merit their trust [3]. This presents a difficult challenge, since designers may not know the degree to which non-technical users can understand concepts related to intelligent systems. In addition, they do not know how these mental models may change over time as users gain experience with the system.

In the interest of increasing designers' knowledge of how users come to understand intelligent systems, we conducted a six-week field study intended to compare user mental models of an intelligent application with the actual system model. We designed, built, and deployed an application that employed user-trained sensor-based statistical models to provide estimates of worker interruptibility. These estimates were then displayed on screens mounted at their office doors. By interviewing the direct reports of managers who were using the application, as well as capturing *in situ* use through on-site surveys, we were able to document how their understanding of the system's operation developed over time. In addition, we were able to ground their stated beliefs about the system with actual instances of use.

Our findings show that even when some details of the system's implementation were revealed, higher-level beliefs about how the system operated remained surprisingly robust, even when those beliefs were flawed and new or contradictory evidence was presented. However, lower-level beliefs about what the system was sensing in order to make its inferences changed substantially over the course of the study period, with participants able to discount some of their early misconceptions before the study's conclusion.

† Jason Chalecki is now with Oracle Corporation.

In addition, we found that several of our participants gave lay descriptions attributing simple statistical or rule-based machine learning concepts to our application, despite having no technical knowledge of such systems.

These results expose the important challenge of dealing with the inertia of static mental models in intelligent UI design, suggesting that users may need additional, high-level feedback in order to adopt more correct structures. We provide some guidance to this challenge by illustrating the various intelligent constructs participants ascribed to our system as a baseline of the complexity users can grasp in intelligent systems. Lastly, while our participants improved their notions of what our system was sensing through experience, simple feedback in the interface about such inputs may not be enough accelerate this improvement.

In this paper, we first discuss related work on mental models and the use of explanations for intelligent systems. We describe our six-week field study of an intelligent system, including the participants used, the technology implemented, and our data collection and analysis methods. We then present our results, showing how participants' mental models varied across participants but remained stable over time. Finally, we discuss how our results can be used, both in terms of what level of understanding to expect from users and avenues that could mitigate incorrect or incomplete mental models.

RELATED WORK

In this section, we discuss related work in both mental models and explanations in intelligent systems. We regard our work as unique in that it applies an examination of mental models to a field study of an intelligent system. Our intent is to provide designers with practical knowledge that is more broadly applicable to other systems.

Mental models

Mental models are a hypothetical construct defined as a mental representation of a real or imagined situation [13]. Norman describes them as internal representations that provide predictive and explanatory power to users [22]. While rooted in psychology, mental models have also been used to evaluate shared understanding among work teams [5], and to measure concept learning in educational settings [6]. Psychological studies of mental models aim to understand how people develop and use these models by focusing on simple physical systems [25] or devices [14] with well-understood, explicit models of operation. While our study borrows from some of the elicitation methods used in this work, it is focused on mental models in the domain of intelligent systems as opposed to fundamental knowledge about the nature of mental models.

Moray provides a theory of mental model development for expert users [18]. He states that instead of forming a completely accurate mental model, users will develop a reduced model that encompasses the majority of observed behavior. He describes these models as *homomorphisms* of the actual system model, where several inputs, or the relationships between them, may be coalesced into a single concept. By breaking down these homomorphic concepts into more detailed accounts of the system's operation, a user could increase their understanding of the system.

However, a particular model may become ingrained to the point that faults are addressed not by a rethinking of the user's mental model, but by an attempt to fit the new behavior into the existing model. This "cognitive lockup" can be difficult to break, especially if the user has a large body of experience with the system and therefore a high degree of confidence in their mental model.

As we explain later, the results of our study showed that while participants simplified models by reducing the number of elements they believed our system was sensing, the structural complexity of their mental models was stable throughout the study. In the discussion section of this paper, we examine this result in light of Moray's theories.

Trust, explanation, and accountability

Explanation has been one of the most frequently used methods for improving trust in intelligent systems [12] by making their behavior more observable [20]. Users have been shown to put greater trust in recommender systems when given additional explanations of the recommendations [11]. Suermondt and Cooper found that physicians using a medical expert system made fewer mistakes when provided with explanations of the system's diagnoses [23]. Antifakos *et al.* showed that confidence levels could increase user trust in an intelligent notification system [1].

Along these lines, Dourish observed that the way a system presents its own state of operation has a strong influence on the way users perceive how their tasks on that system are being accomplished [7]. He argues that systems should provide accounts of their own operation that capture the inherent structure of system processes, revealing features deemed relevant and hiding unnecessary details. Bellotti and Edwards reiterate this point with respect to intelligent, context-aware systems, stating that such systems must be intelligible as to their states and intent in order for people to control their behavior [2]. To design for this sort of accountability, it is necessary to provide the appropriate abstractions from the implementation to the interface. By examining user mental models of an intelligent system that incorporates some explanatory power, our results can support accountability in design by reporting on the abstractions developed by people during normal use.

Borgman used mental models to gauge the effectiveness of training techniques for an information retrieval system [4]. Her findings showed that subjects found it easier to describe the process of achieving tasks on the system rather than describing the system itself. More recently, Muramatsu and Pratt investigated how users' mental models of search queries could be improved to correctly use concepts such as logical operators and stop words [21]. By conducting open-ended interviews both before and after users submitted a search query, they were able to assess deficiencies in

mental models and design appropriate explanations to correct them. We take a similar approach, but to a different class of system that incorporates probabilistic inference and machine learning.

METHODS

For this study, we designed and built a system to present estimates of the interruptibility of managers in an office setting. Since this system was intended to be used intermittently throughout the workday, we planned our study as a six-week field deployment. In this way, our participants were afforded prolonged exposure to the system. By employing several methods of data collection, we were able to obtain a more realistic picture of how their mental models developed over time.

Participants

We recruited four managers and nine of their direct reports from a local university's human resources department. Their responsibilities ranged from payroll and benefits management for the university to coordination of temp services. Participants had no existing knowledge of programming or machine learning. The department was organized such that a small subset of our participants (managers) would generate estimates of interruptibility, while the remaining participants (direct reports) would use them in their work. In this situation, we would focus on the mental models of direct reports as opposed to the managers. Our rationale was that managers, by seeing their own interruptibility estimates, would have more time to monitor and even modify the estimates. In this way, they would develop an understanding of the system that was not achieved through normal work practice and would be less valid in the context of our study.

Our participants consisted of two six-person groups, each comprised of two managers and four of their direct reports (one of the direct reports, due to a hectic work schedule, had to leave the study after two weeks). These groups were split evenly into two different buildings about one city block apart, with little to no contact between them. Every manager had his/her own office and door, and all participants were located on the same floor of their respective buildings. Direct reports had either an individual office or cubicle. Communication was conducted primarily through email, office visits, and scheduled meetings, with occasional phone use. Instant messaging was not in use. Each direct report interacted primarily with only one of the managers in his/her group. All participants received movie theater gift cards as compensation.

Interruptibility Estimates

Interruptibility estimates were obtained using sensor-based statistical models developed with Subtle, an extensible toolkit developed by Fogarty and Hudson [10]. A statistical model of interruptibility was constructed for each manager during a three-month pre-deployment training phase. The training phase consisted of prompting each manager once per hour for a self-report of their interruptibility. Potential features were then created by using a small set of operators to explore the available sensor events. Sensor events included input events (*e.g.*, keyboard/mouse), window manager events, nearby wireless access points, and audio. Operators included the current value of a feature, its recent values within a given time window, and whether its value was above/below a learned threshold. The feature set was then filtered to remove equivalent, highly-correlated, or noisy features. Finally, the optimal feature set was selected using a wrapper-based selection method [16]. This approach has been previously demonstrated to produce reliable estimates of office worker interruptibility, with model estimates outperforming human observers [9].

	Full Model Accuracy	True Int	False Int	True Non-Int	False Non-Int	A' ROC Area	Presented Accuracy
Manager 1	98.2%	72	0	88	3	.996	86.5%
Manager 2	98.1%	21	2	134	1	.994	95.0%
Manager 3	93.8%	35	0	26	4	.976	93.8%
Manager 4	94.8%	41	3	14	0	.977	94.8%

Table 1. Summary of sensor-based estimate reliability.

While this paper is not focused on examining the reliability of sensor-based models of human interruptibility, it seems likely that a reasonable level of reliability is necessary before the consumers of interruptibility estimates will place confidence in those estimates. Table 1 therefore presents a very brief summary of the reliability of the interruptibility model for each manager. Model accuracy ranged from 93.8% to 98.2%, estimated using ten-fold cross-validation. We note that the individual models learned here contain some automatically-learned features that are clearly based on small nuances in an individual's data, are unlikely to be generally applicable, and are difficult to interpret. This seemed to be a larger problem when more self-reports were available for a particular manager, as this provided more data in which the learning system could find such small details. Based on the relative importance of each feature in each learned model, we chose a subset that we felt were both important and interpretable. Only those features were presented to participants in the remainder of this work, and the presented accuracy column in Table 1 presents the accuracy of only those features for each manager.

Existing Interruptibility Practices

Interviews conducted with direct reports before we deployed our system indicated that they rely chiefly on two pieces of information in order to determine interruptibility. The first is the state of the manager's office door. If it is closed, this is a strong indicator that the manager is not to be disturbed. If open, coworkers then listen for the sound of talking within the office. If no talking is heard, the manager is usually understood to be available. Otherwise, a judgment call is made based on the presence and identities of other people in the office, the content of the manager's

Figure 1. Door tags used onsite prior to our study.

conversation, and the importance or urgency of the information the coworker needs to communicate. For instance, a manager engaged in a conversation about last night's football game with the coworker across the hall would invite interruption, while being engaged in a business-related phone call would not.

To a lesser extent, some participants reported using the manager's online calendar as a means of saving a trip down the hall. They reported instances where they had walked to the manager's office only to find that the manager was either not in the office or holding a scheduled meeting there. In addition, several participants reported making eye contact with their managers and using subtle gestures to negotiate interruption, in a similar manner to that reported by Kendon and Ferber [15]. It should be noted that the system we deployed was not capable of sensing some of the cues participants relied on for determining interruptibility. For example, it could not sense door state, presence, the identity of a conversant, calendar information, or conversational topic. While this made it challenging for participants to apply their existing models of interruptibility, our results show that by the study's end, participants retained few of their initial notions about what the system could sense.

An interesting feature of the workplace was the existence of ceramic tags on each office door (Figure 1). These tags, participants told us, had been created to signal availability to coworkers who veered from the tacit open door/closed door policy that indicated interruptibility. The tags were painted green on one side and red on the other, acting as a "do not disturb" indicator when the red side was displayed.

The problem with these tags, we were told, was that new employees had to be educated on their use, and that a fair amount of diligence was required to keep the state of the tags accurate. By the time our interviews were conducted, the tags had been present for several years, but the original proponents had since moved to a different building. Only a small fraction of employees, and none of our managers, still used the tags, and several still had the mistaken belief that they were in use by all of their coworkers. It should be noted here that none of our direct reports were under the impression that their managers were using the tags, so this existing system did not compete with our technology.

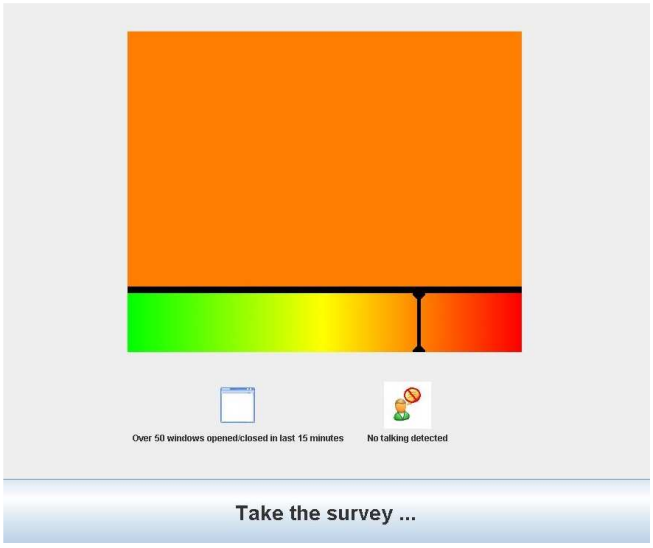

Figure 2. Screenshot of door display, with two relevant features displayed beneath the color-coded estimate.

Interruptibility Displays

In designing our application, we made use of the existing door tags. Since most participants were familiar with them, we used the door tags as the basis for our application design, with a solid color used to represent interruptibility (Figure 2). A gradient scale below the color shows the position of the estimate on the overall scale, which runs from red (very uninterruptible) to green (very interruptible). Our display used this gradient to indicate its confidence, with the middle yellow region corresponding to situations where the learned model indicated that either value was equally likely. Informal user testing with 11 subjects confirmed that color codes, along with the gradient scale, were more accurately matched to interruptibility levels than photographs using different levels of transparency, levels of saturation, or gaze directions of subjects in photographs. As an additional constraint, some participating managers had reservations about having their photographs displayed in our interface. Lastly, when no estimates had been received from the manager's computer for more than 10 minutes, the display would show a test pattern, usually indicating that the manager's computer was turned off.

While the system was capable of further refining its statistical models once deployed, we opted to keep the models static after the training phase. We made this decision in order to simplify the process of comparing participant mental models with the actual system models. With continuous learning, the system models would change frequently, potentially at different rates for each manager. By avoiding this "moving target" problem, we could make more valid comparisons across participants and groups.

In addition, feature representations within Subtle were not human-readable, requiring a level of abstraction to transform, for example, a particular computed property of ambient audio into a phrase such as "talking detected". While an abstraction of the low-level inputs, we feel these

human-readable descriptions did not diminish the accuracy of our portrayal of models.

Once the statistical models were built and their features labeled with human-readable text, they were installed on the managers' computers. A separate program continuously served updated interruptibility estimates based on the current sensor data. These estimates were delivered via the university network to the door displays outside each office.

Figure 3. Door display setup used in our study.

These displays took the form of 12" touch-sensitive color LCDs that were stationed outside the office doors of each of our four managers (Figure 3). Since participants in both sites were accustomed to frequent office visits and shared a common hallway, we were confident that they would have sufficient visibility for participants to notice them on a daily basis. The touch-sensitive screen was used for additional data collection, described later.

Providing accountability

Our two six-person groups were each given a different version of the door display user interface in order to examine whether the addition of feedback in the interface would improve users' mental models of the system's operation. In one condition, the system dynamically displayed up to three features that contributed the most information to the current interruptibility estimate, using a cross-entropy metric similar to that used by Suermondt and Cooper [23]. These features, which were specific to each manager's user model, were then displayed at the bottom of the interface along with an associated icon (*e.g.*, a keyboard icon for keyboard activity, speaker for audio activity, etc.). In this way, participants could see exactly what the most important sensors were, as well as their relative contributions to the estimates they were seeing. While estimates were updated continuously, the display itself was updated every 20 seconds to allow time to read the feature descriptions and minimize distraction. Participants in the other condition were able to see only the color-coded interruptibility estimates.

Data Collection

Our primary source of data collection was a series of five semi-structured interviews, each lasting between 30-60 minutes and recorded using both field notes and audio. The managers did not participate in these interviews; we were mainly interested in how their direct reports interpreted the interruptibility estimates generated for the managers. We did, however, interview managers at the close of the study for their thoughts on the system's usefulness.

The first of the five interviews were conducted prior to deploying our office displays. These interviews focused mainly on work practices, including job descriptions, the number of meetings per week with participating managers, meetings with other participants, and existing means of estimating interruptibility. The interview concluded with questions about how participants believed a system that predicts interruptibility *might* work, including what information the computer would use, and how this information might be synthesized into an estimate.

Eliciting technical information from non-technical users

After our system was deployed, we conducted four additional interviews at roughly one-week intervals. These interviews focused primarily on eliciting participants' mental models of the system's operation. Our initial concerns were that our participants, whose interaction with computers was primarily through a small set of office and communication applications and whose technical support was handled by the university's IT department, would not be able to venture an opinion on the operation of a fairly sophisticated machine learning system.

However, as Norman has claimed [22], mental models evolve naturally as users interact with a system. Most useful to our interviews was breaking down the problem into four areas. First we solicited the set of *sensors* that participants believed were being used by the system. In other words, what was the computer capable of sensing that was relevant to interruptibility? Second, participants were asked for their beliefs on the *relative importance* of these sensors. Third, they were asked to explain how they thought these sensors were *synthesized* into an interruptibility estimate. Finally, they were asked to *relate their experiences* to this model as a means of confirming its validity. Each week participants were asked to describe some instances in which they noticed or used the office displays, and the role they played in choosing whether or not to interrupt a manager.

As a supplement to these interviews, a short (five question, multiple choice) survey was included on the office displays themselves for collecting some *in situ* information on the use of the displays. These questions asked the user to rate how important his/her need was to speak with the manager (5-point Likert scale), whether he/she would actually interrupt the manager, and whether that decision was based on the display, the status of the manager, or the importance of the information. Lastly, three 5-point Likert ratings were requested for the participant's agreement with the estimate, the degree to which the display influenced their decision to talk to the manager, and their confidence in the display in absence of any other information.

Participants were encouraged, but not required, to complete the survey prior to visiting a manager's office, while the intent of the visit was still at the forefront of their attention.

This convention was not always followed, as some visits were time-sensitive and could not be delayed by the survey.

Upon the completion of a survey, a follow-up email to the participant would be triggered that would ask a few additional longer-form questions about the office visit. These questions asked what particular aspects of the office and the displays themselves contributed to the estimate on the display as well as the participant's own personal estimate of interruptibility.

Data analysis

The structures of our participants' mental models were coded by identifying several types of relationships and entities described during interviews. These included believed *sensors* used by the system that were listed by participants during interviews, *conditions* that involved making a decision based on the current state, *connections* that used the output of one element as the input to another, *priority lists* of activities or inputs were relevant, and *patterns* that consisted of activities or input levels that were learned over time and distinguishable by some software recognizer. In addition, participants mentioned *history* as a means by which patterns could be established. These coded models were then grouped by similarity into the four model types described in the next section.

RESULTS

After our deployment, we were left with a number of completed surveys documenting interactions with our door displays as well as coded interview data on participants' mental models of the system's operation. We present these results below.

Door display surveys

Forty-three surveys were completed at our door displays during the course of the study, with 19 of those completed anonymously. We learned later that two people in the office who were not officially part of the study had completed several surveys anonymously. All eight of the direct reports who participated completed at least one survey, with four of them completing at least four. Of the 43 surveys, 29 were completed in the first two weeks of the study.

Across all surveys, there was a moderate degree of correlation between the influence of the system on participants' decision-making and their confidence in relying on it exclusively for deciding whether or not to interrupt (Pearson's r=0.48). There was a low correlation between the participants' agreement with the system's estimates and their confidence in relying on them exclusively (Pearson's r=0.33). These results seem to indicate that users would rely on the system more if they have confidence in its predictions, but that they tend to trust their own estimates more. Only in nine of the surveys did participants report relying on the display to determine whether to interrupt. Average ratings for confidence in the displays (3.0/5.0, ±1.1) and their degree of influence (2.93/5.0, ±1.6) were higher than for agreement (2.75/5.0, ±1.3), though not significantly. Standard deviations were high, indicating that participants had either high or low degrees of faith in the system.

Fifteen follow-up email surveys were returned from seven of our participants. About half of them (8) were in agreement with the estimate presented on our displays. Of the remaining seven responses, two indicated that the manager had left the office, but was displayed as interruptible. The remaining surveys indicated that the system estimated too conservatively in those cases, predicting lower interruptibility than was the case, at least from the direct reports' perspectives.

A shortcoming of our system was that we could not explicitly detect a person's presence in the office. Therefore, the software would estimate interruptibility based on available sensors independent of whether the manager was in the office or not. This was a frequently cited problem in interviews, with participants mentioning that it also affected their responses on the door display surveys. Subsequently, our results in terms of participant agreement and confidence with respect to the system were likely affected by this issue.

Mental models

In eliciting mental models from participants, we asked for their beliefs concerning both the sensors that were being used by our system and the mechanism(s) to convert these inputs into an estimate of interruptibility. In this section, we start with participants' responses on possible sensors before discussing the actual model structures.

System sensors

The most common factors participants believed were being used by our system included the calendar, the presence of talking, and keyboard/mouse activity. While the presence of calendar-related windows on the screen was influential in one manager's model, the actual schedule itself was not used by our software. Mouse/keyboard activity was an influential sensor for three of the four managers, and the presence of talking was important to all of them.

An analysis of variance showed a significant effect of interview date on the number of sensors participants believed were being used by our system, $F(1, 31) = 7.04$, $p < 0.05$, with fewer sensors reported as the study progressed. However, there was no significant difference between those who received feedback about relevant features from the system's statistical models and those who did not. By the final week, participants reported an average of only 1.75 of the sensors they believed were being used by the system during the first week. In interviews, participants reported ruling out a number of potential sensors for a variety of reasons. For example, one participant ruled out cameras as "too Big Brother". Another ruled out audio because she noticed no change in the display after she had conducted a conversation within her manager's office. Note that in the first case, the decision to rule out cameras made for a more correct model, while in the second, it became less correct since the system was in fact using audio.

We measured correctness by taking the sensors evaluated by each user model of interruptibility and comparing them to the inputs listed by participants. On average, participants listed nearly the same number of correct sensors (2.1) in the first week as the last week (2.0), with no significant difference between the groups with feedback and those without. Given that participants listed significantly fewer sensors at the end of the study, a higher percentage of those features were correct. An analysis of variance showed a significant improvement in correctness between participants' beliefs for the first two interviews and those for the last three interviews, $F(1,31) = 4.44$, $p < 0.05$.

Model structure

Despite using similar means to judge managers' availability prior to the study, participants reported a fairly diverse range of topologies in terms of how they thought the system was operating. These models varied in the ways they incorporated history, made use of statistics and in one case, whether the inferences were human or machine-generated.

Our expectation, however, was that these models would change with time as participants gained experience with the system. Further, those participants who were provided with additional information about the features being used by the system were expected to develop more accurate mental models. What we found was that the overarching structures of their mental models remained for the most part stable throughout the study and having additional information had little impact. Later in our discussion, we remark on why this was the case and propose how the interface could trigger higher-level changes to users' mental models.

Simple set of rules

Two of our participants (one from each condition) believed that the system worked on a simple set of conditional rules to determine interruptibility. An example from one participant is given below:

> Is the computer on?
> - **No**: Display test pattern
> - **Yes**: Next condition

> Is the manager on the phone?
> - **Yes**: Highly uninterruptible (bright red)
> - **No**: Next condition

> Are there voices in the room?
> - **Yes**: Uninterruptible (red)
> - **No**: Next condition

> Is there an event on the calendar?
> - **Yes**: Uninterruptible (red)
> - **No**: …

Such rules would continue to be applied until a default value of green (interruptible) was reached. Though neither participant could explain fully where these rules originated, they both assumed that the training phase had some role in formulating them.

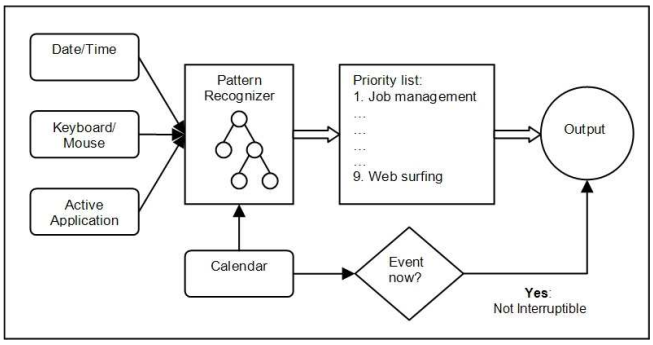

Figure 4. A diagram of one participant's mental model of the experimental system. System inputs are sent to a pattern recognizer that uses a decision tree-like algorithm to identify activities that are ordered by importance.

Self-reported (Remote Control)

One participant held to the belief throughout our study that the display was controlled manually from inside the office by the manager. While one dialog used to collect training data could be that it is a "remote control" to manipulate the display, there was no overlap in the deployment of the training dialog and our door displays.. Therefore, this participant had no visual evidence to support this belief. However, this participant was not confident that a computer system could capture the factors inherent in assessing interruptibility, and therefore adopted the simpler avenue despite having little observational support for it. This mental model proved robust to mismatches between the manager's displayed and actual interruptibility:

> Interviewer: So why would the display say red even when [your manager] is available?
>
> P: Well…maybe she finished whatever she was doing, but for one reason or another, hadn't hit the key (to change the interruptibility level).

Also interesting was the fact that this participant was part of the group that received additional feedback on our door displays about relevant features used for the system's estimate. Rather than interpreting these features as leading to the estimates shown, they were instead interpreted as being there to justify the estimate chosen by the manager and help the potential interrupter with their decision of whether or not to heed the display.

Prioritized cases

Two participants (one from each condition) described models incorporating the following elements:

- A *prioritized list* of activities ranging from most to least important to interruptibility,
- A *recognizer* of sorts to identify which activity is currently occurring from the sensor inputs,
- One or more *conditional rules* used to handle overriding situations or special cases.

One such model as described and confirmed by a participant is shown in Figure 4.

A list of activities might range from web browsing or checking email (low priority) to working on a large report or spreadsheet (high priority). Participants assumed that these priorities were determined during the training phase through either automated means or through interviews/observations of the managers.

In terms of recognition, participants described the system as using available sensor data to test a series of conditions that would establish which activity was occurring. To quote one participant, "…it has to be something like, you know, like a tree…if 'yes' then this, if 'no' then this." This participant went on to describe a decision tree-style process by which the system builds its list of activities from past history and tests the current situation along a number of branching conditions to see if it matches anything in the list.

Lastly, participants included additional rules outside of the main priority/recognition mechanism to account for special cases or observed behaviors. For example, one participant believed the presence of a calendar appointment would override any other recognized activity and immediately designate the manager as uninterruptible. Another participant added a "confused" state to the model to explain situations where one display appeared to fluctuate between several states within a short amount of time.

Similarity to average

Three other participants articulated a model that, while similar structurally to the prioritized cases model described above, also explicitly incorporated the use of simple statistics to determine degrees of interruptibility. In two cases, this constituted a deviation from the mean level of activity sensed in the manager's training data. In the third case, the participant described a more nebulous "calculation" that nonetheless relied on historical statistics to arrive at an estimate. In these cases, participants thought in terms of "levels" of activity rather than whether a particular activity was occurring or not. One participant described it this way:

> P: For example, the number that it looks up tells it they're not busy, but the number that this is coming up with is higher.
>
> Interviewer: Ok, so it has some sort of idea of, uh…
>
> P: A baseline.

The "number" here refers to the average level of several inputs (mouse, keyboard, audio, *etc.*) that are associated in the training data with a certain level of activity.

We believe two main factors contributed to this belief. First, participants with these models associated the training phase of the study with a collection of "average" levels of activity for the various inputs. Second, these participants noted the continuous nature of the color scale used to display interruptibility and associated it with a continuous range of possible values. It is possible that the two participants who adopted the prioritized case-based model described previously did not make this association and therefore did not develop a statistics-based model.

Of all the models described by participants, this class was most similar in structure and function to the actual system used. Most notably, this model incorporated the use of history and statistics, as well as some combination of sensor data with varying degrees of influence, to arrive at the final estimate. Interestingly, two of the three participants with this model came from the group that had no additional feedback about the features used by the system. In fact, none of the four types of models listed seemed to be related to whether the participants were given additional information about the features used. We address this finding in the discussion section.

DISCUSSION

Here we review the major findings of our study and discuss their implications for future research and design of intelligent user interfaces.

Similarity to the system model

While none of our participants possessed a mental model that was identical to the system model, three participants developed a statistics-based mental model that used past experience to associate current sensor activity with an estimate of interruptibility. Elements of both the system's training phase and user interface contributed to this understanding. At a basic level, we observed that for systems incorporating machine learning and estimates of current or future state, the key concepts of learning from history and statistical inference from current sensor data must be effectively communicated in the interface.

It was clear from our results that incorporating feedback about relevant features was of only moderate use in helping our participants understand the system's operation. Participants had little success in retaining knowledge of these features from week to week, and perhaps more importantly, were unable to make higher-level structural adjustments to their mental models using this feedback. While all participants had some knowledge of the training phase prior to the deployment of our office door displays, not all of them incorporated this phase into their mental models. In addition, despite our use of a gradient scale to represent the continuous nature of the output, only about half of our participants regarded it in these terms, and fewer made the connection of this output to statistical patterns.

Stability of structure

The fact that participants described basically the same model structure with each successive interview was somewhat surprising to us. We had expected participants to frequently modify their models as more observations were made of the system's behavior and conversations with coworkers exposed them to alternative theories. As it turned out, they rarely discussed the system with one another, with some participants worried it would be considered "cheating", even though we did not discourage it. The most significant changes to mental models came when participants incorporated additional inputs, removed others,

or tacked on conditions to handle special cases that were not covered by the existing model.

In addition, some mental models held up in light of conflicting experiences or clear deficiencies in terms of incorporating known components of the system, such as the training phase. In the case of the participant who saw the system's estimates as the result of manual control, feedback about relevant features was regarded as auxiliary information, and the training application as being repurposed into a remote control. Moray's theory of "cognitive lockup", where operators maintain their beliefs about a system even in the face of contradictory evidence, may hold here, but this theory is intended for expert users with a grasp of the system's normal operating parameters. Given that a number of our participants verbally expressed low confidence in their mental models at the end of the study, it may not be appropriate to regard them as experts after five weeks of experience.

Instead, it may be that the feedback provided was at too low a level to help participants in breaking out of their conceptualizations of the system's mechanics. While the interface provided information about inputs and relevant features (in one condition), and clues to the probabilistic nature of the output by using a gradient scale, it provided no higher-level information on the features' relationships to one another or the process used to relate those features to the training phase.

Retaining knowledge of sensors

In contrast to the stability we observed in model structure, participants showed that they were able to reduce the set of potential sensors they believed were used by our system over the course of the study. Moreover, they were able to achieve modest improvements in the correctness of these beliefs over time. Rather than using the feedback provided by our system, however, participants relied more on observations of the system's behavior and reflective assessments of the feasibility of potential sensors.

For the group that was given additional information on model features, we noticed that with the exception of audio, in nearly every case where a new sensor was learned from the display in a given week, it was not reported again the following week. These included sensors for window focus events, nearby wireless access points, and keyboard/mouse activity. While these items were not used by participants to determine interruptibility prior to the introduction of our system, participants had consistently mentioned other events that also were not part of their existing practices. It is possible that much more exposure to this information is needed before it is retained. It is also possible that more technical aspects of the system such as window focus events are more difficult to remember than everyday or more familiar events/artifacts such as conversation, email, and calendars. Further study is needed to determine whether more technical inputs to intelligent systems require additional explanation or emphasis to learn.

Understanding machine learning concepts

One of the more interesting findings of our study was that participants were able to ascribe basic machine learning concepts to our system despite being unfamiliar with the field. One participant described, in lay terms, the basic decision tree algorithm, while several others described the use of nested yes/no questions that demonstrated the use of decision trees without going into the details of their construction. In addition, over half of our participants described the use of "patterns" or "averages" gleaned from historical data as a means of predicting current or future states. While some of these patterns were described in terms of the decision tree-type structures just mentioned, others were described in terms of simple statistics such as deviation from the mean.

These results are encouraging in that, if demonstrated across a broader cross-section of the population, designers can potentially know what to expect in terms of user sophistication with respect to machine learning systems. In cases where the underlying algorithm makes little difference in terms of functionality, designers may consider using one that is closer to those described by our participants. If a more complex algorithm is needed, a simplified version that is closer to concepts readily understood by users could be presented in the interface.

Reliability of interviews

An issue with our data collection method is the nature of mental models and the reliability with which they can be elicited. Norman claims that simply asking participants is less reliable than collecting data in the context of activities or problem solving [22]. The fact that our study was conducted in the field made it difficult to capture such instances reliably, as interactions at the doors happened at random points throughout the day, and the display itself was noticed peripherally regardless of whether one intended to meet with a manager. Knowing this might be the case, we used the survey data collected at the displays to help participants in re-creating their experiences with them, and also to assist in "running' their mental models to explain how those experiences influenced them. In doing so, we follow the mental model elicitation technique favored by Morgan et al. who, as in our work, focus on less well-constrained domains [19]. The stability of model structure over time lends reliability to our data, given that we made participants describe the entire model at each interview. Overall, our interviewing method was similar to controlled studies of mental models mentioned earlier [14, 25].

CONCLUSIONS AND FUTURE WORK

A natural next step for this work will be to use our findings in the design of user interfaces that express a correct conceptual model to users. While several of our participants were on the right track in terms of understanding the use of history, statistics, and continuous evaluation of features to arrive at estimates, others were not aware of these key concepts and required more than knowledge of the features being used. Any design must illustrate these higher-level

concepts to assist users in breaking out of incomplete or incorrect mental models, such as those based on simple conditional rules. We hope to generate our designs for the interruptibility displays used in this study and evaluate their effects in a lab setting before returning to the field.

Our study is limited in terms of the domain of our application and our participants' demographics. Additional work is needed in other application areas with different machine learning algorithms to establish a broader pattern of how users come to understand intelligent systems.

Lastly, mainstream intelligent applications such as spam filters and recommenders incorporate adaptive user models that improve over time. By using static models, our study assumed a stable system and ignored any kind of learning phase. An interesting direction for this work would be to examine how user and system models co-evolve during such a learning phase.

In this paper we have described a field study of how users' mental models develop around an intelligent system. We have shown that users are capable of attributing concepts of machine intelligence to our system. Designers can use these concepts to incorporate higher-level feedback in the user interface that could correct stable but flawed mental models. We have also shown that simple feedback about features used by our system was not enough to improve the rate at which users learned these features, and that more work is needed to ensure that users retain this knowledge.

ACKNOWLEDGMENTS

We would like to thank Eric Horvitz for early discussions that inspired this work. Thank you also to Jen Mankoff, Elaine Huang, and Daniel Avrahami for their help on the paper and analysis. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010, and by the National Science Foundation under grants IIS-0121560, IIS-0325351, and IIS-0205644.

REFERENCES

1. Antifakos, S. Kern, N., Schiele, B., Schwaninger, A., (2005) Towards improving trust in context-aware systems by displaying system confidence. In *Proc. MobileHCI 2005*, pp. 9-14.
2. Bellotti, V. and Edwards, W.K. (2001) Intelligibility and Accountability: Human Considerations in Context-Aware Systems, *Human-Computer Interaction*, 16(2-4):193-212.
3. Birnbaum, L., Horvitz, E., Kurlander, D., Lieberman, H., Marks, J., and Roth, S. (1997). Compelling intelligent user interfaces—how much AI? In *Proc. of IUI'97*, pp. 173-175.
4. Borgman, C. (1986) The User's Mental Model of an Information Retrieval System. *'Int'l Journal of Man-Machine Studies* 24(1):47-64.
5. Cannon-Bowers, J.E., Salas, E., and Converse, S. (1993) Shared Mental Models in Expert Team Decision-Making," In *Individual and Group Decision-Making: Current Issues*, J. Castellan, (ed.), Hillsdale, NJ: Erlbaum.
6. Chi, M.T.H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in Instructional Psychology*, Hillsdale, NJ:Erlbaum. pp. 161-238.
7. Dourish, P. (1995). Accounting for System Behaviour: Representation, Reflection and Resourceful Action. In *Proc. of Conference on Computers in Context CIC'95*, pp. 145-170.
8. Dzindolet, M., Peterson, S., Pomranky, S. Pierce, L. and Beck, H. (2003) The role of trust in automation reliance, *'Int'l Journal of Human-Computer Studies*, 58(6):697-718.
9. Fogarty, J., Hudson, S.E., and Lai, J. (2004). Examining the Robustness of Sensor-Based Statistical Models of Human Interruptibility. In *Proc. of CHI 2004*, pp. 207-214.
10. Fogarty, J. and Hudson, S.E. (2007) Toolkit Support for Developing and Deploying Sensor-Based Statistical Models of Human Situations. *To Appear, CHI 2007*.
11. Herlocker , J., Konstan , and J., Riedl, J. (2000) Explaining collaborative filtering recommendations, In *Proc. of CSCW 2000*, pp.241-250.
12. Johnson, H. and Johnson, P. (1993) Explanation facilities and interactive systems. In Proc. of IUI'93, pp. 159-166.
13. Johnson-Laird, P. N. (1983) *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard Press.
14. Kempton, W. (1987) Two theories of home heat control. In N. Quinn & D. Holland (Eds.) Cultural Models in Language and Thought, Cambridge University Press.
15. Kendon, A. and Ferber, A. (1973) A description of some human greetings. In R. Michael and J. Crook (Eds.), *Comparative Ecology and Behavior of Primates*, pp. 591-668. New York: Academic Press.
16. Kohavi, R. and John, G.H. (1997) Wrappers for Feature Subset Selection, *Artificial Intelligence* 97(1-2):273-324.
17. Maes, P. (1994) Agents that Reduce Work and Information Overload. Communications of the ACM, 37(7):31-40.
18. Moray, N. (1987) Intelligent Aids, Mental Models, and the Theory of Machines. *'Int'l Journal of Man-Machine Studies*, 27 (5):619-629.
19. Morgan, M.G., Fischhoff, B., Bostrom, A. and Atman, C.J. (2002) *Risk Communication: A Mental Models Approach*. Cambridge, UK: Cambridge University Press.
20. Muir, B. (1994) Trust in automation Part I: Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37(11):1905-1922.
21. Muramatsu, J. and Pratt, W. (2001) Transparent Queries: Investigating Users' Mental Models of Search Engines, In *Proc. of SIGIR 2001*, pp. 217-224.
22. Norman, D.A. (1983). Some observations on mental models. In D. Gentner & A.Stevens (Eds.) *Mental Models*, pp. 7-15. Hillsdale, NJ: Erlbaum.
23. Suermondt, J. and Cooper, G. (1992) An Evaluation of Explanations of Probabilistic Inference. In *Proc. Computer Applications in Medical Care*, pp. 579-585.
24. Tversky, A. and Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases, *Science* 185(4157):1124-1131.
25. Williams, M.D., Hollan, J.D., and Stevens, A.L. (1983) Human Reasoning about a Simple Physical System. In D. Gentner & A.Stevens (Eds.) *Mental Models*, pp. 131-154. Hillsdale, NJ: Erlbaum.