# Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models

Todd Kulesza[1], Simone Stumpf[2], Margaret Burnett[1], Sherry Yang[3], Irwin Kwan[1], Weng-Keen Wong[1]

[1]School of EECS
Oregon State University
Corvallis, Oregon, USA

[2]Centre for HCI Design
School of Informatics
City University London, UK

[3]Computer Systems Engineering Technology
Oregon Institute of Technology
Klamath Falls, Oregon, USA

{kuleszto, burnett, kwan, wong}@eecs.oregonstate.edu, Simone.Stumpf.1@city.ac.uk, sherry.yang@oit.edu

***Abstract*—Research is emerging on how end users can correct mistakes their intelligent agents make, but before users can correctly "debug" an intelligent agent, they need some degree of understanding of how it works. In this paper we consider ways intelligent agents should explain themselves to end users, especially focusing on how the soundness and completeness of the explanations impacts the fidelity of end users' mental models. Our findings suggest that completeness is more important than soundness: increasing completeness via certain information types helped participants' mental models and, surprisingly, their perception of the cost/benefit tradeoff of attending to the explanations. We also found that oversimplification, as per many commercial agents, can be a problem: when soundness was very low, participants experienced more mental demand and lost trust in the explanations, thereby reducing the likelihood that users will pay attention to such explanations at all.**

***Keywords—mental models; explanations; end-user debugging; recommender systems; intelligent agents***

## I. Introduction

How should intelligent agents explain themselves to users? The predominant approach in commercial agents is to "keep it simple" (e.g., the music recommender Pandora.com describes its song recommendations via a single sentence). However, such simplicity may prevent users from understanding how the agent makes decisions, erecting a barrier to users' potential ability to help the agent improve; i.e., it may obstruct their ability to effectively *debug* the agent's reasoning.

As with other kinds of end-user debugging, users' *mental models* of how an agent works help them decide exactly what about the agent they need to correct, and how to go about it [11]. In this paper we raise the question of whether simplicity is always the right attribute to prioritize when designing agent explanations. Another possibility is to prioritize explanation completeness; prior work has shown that providing end users with detailed explanations about an intelligent agent's reasoning can increase their understanding of how the system works [11]. However, information comes at the price of attention—a user's time (and interest) is finite, so the solution may not simply be "the more information, the better".

To investigate how intelligent agents should explain themselves to their users, we performed a qualitative study to separately consider two dimensions of explanation *fidelity*: *soundness* (how truthful each element in an explanation is with respect to the underlying system) and *completeness* (the extent to which an explanation describes all of the underlying system). We then investigated how varying soundness and completeness (as in Fig. 1) impacted users' mental models of a music-recommending intelligent agent, what types of information were most helpful in the explanations, how explanation fidelity impacted users' perceptions of the costs and benefits of attending to these explanations, and users' trust in the explanations' veracity. Our research questions were:

*RQ-1*: How do *soundness and completeness* impact end users' mental models?

*RQ-2*: Which *types of information* are most helpful for users' mental models?

*RQ-3*: What *obstacles* do end users encounter when building mental models of an intelligent agent's reasoning?

*RQ-4*: How do users' perceived *costs and benefits* of attending to explanations change with explanation fidelity?

*RQ-5*: How does user *trust* change as explanation soundness and completeness increase?

## II. Background and Related Work

### A. Functional and Structural Mental Models

Mental models are internal representations that people build based on real world experiences. These models allow people to understand, explain, and predict phenomena [9], and to then act accordingly. For example, a mental model of a computer could be that it displays everything typed on the keyboard and remembers these things after the user presses a "save" button. This simple model would help a novice predict that turning off the computer without pressing "save" will result in lost work. Mental models can vary in their fidelity—software developers hold higher fidelity models of computers, for example.

There are two types of mental models: functional models

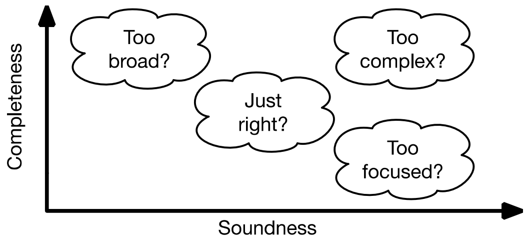


Fig. 1. Our problem space: How sound and complete do explanations need to be to help end users build high-fidelity mental models?

This work was supported by NSF 0803487.

978-1-4799-0369-6/13/$31.00 ©2013 IEEE

imply that the end user knows how to use something, but not how it works in detail, whereas structural models provide a detailed understanding of how and why it works. Norman [14] reported many instances of erroneous mental models leading to behavior with unexpected consequences, suggesting the importance of structural models.

### B. *Building Mental Models of Intelligent Agents*

Our prior work investigated the feasibility of end users building high-fidelity structural mental models of recommender systems [11]; participants grasped many nuances of the recommendation algorithm, and as they developed higher-fidelity mental models, they better controlled the recommender. That study, however, relied on a human instructor instead of automated explanations. Tullio et al. studied user interactions with an intelligent agent that predicted office workers' availability, finding that even when faced with contradictory evidence, users often tried to fit it into their existing mental model of the agent [19]; this suggests that it is important to help end users develop correct models as quickly as possible. Cramer et al. found that providing explanations of why an intelligent agent believes two items are similar helped participants *believe* they understood how the agent worked, but their actual mental models were not measured [4].

Lim and Dey [12] explored *intelligibility types*—different types of information—that intelligent agents can present to end users. Their taxonomy includes categories for the inputs used by an agent, its reasoning process, and concrete explanations of why it made a particular decision, among others. They later showed that one of their intelligibility types can help end users better understand the output of decision trees [13]. Our work uses the Lim/Dey taxonomy of intelligibility types in our investigation of explanation fidelity.

There is an open question of *how much* information agents should present to users—existing research about explanation fidelity has focused on the value of low fidelity explanations versus no explanations. For example, when Herlocker et al. [8] evaluated 21 ways of explaining collaborative filtering systems, they investigated only low fidelity explanations because they wanted to "avoid creating a new kind of information overload by presenting too much or too confusing data". Other researchers have also investigated explanation fidelity in recommender systems (Tintarev and Masthoff provide a comprehensive overview in [18]), but primarily as it relates to trust, acceptance, or satisfaction.

One hypothesis is that more information in an explanation will help users build better mental models. However, very complete or complex explanations require more attention to process, which disincentivizes users to build accurate mental models. Rosson et al., for example, found that the Minimalist explanation model [3]—which minimizes passive learning tasks, such as reading long explanations, favoring instead short explanations coupled with try-it-out exploration—helped programmers understand Smalltalk programs up to two orders of magnitude faster than traditional instruction techniques [15].

On the other hand, the Attention Investment Model predicts that people will still use high-cost explanations *if* they perceive the benefits will outweigh the costs (e.g., time) and risks (e.g., no reward) of doing so [1]. Thus, rather than simplifying explanations to one or two salient points (as contemporary agents do), an alternative may be to identify the most helpful information for the end user (as in the Minimalist explanation model), and then communicate the benefits of paying attention to it. One successful way to communicate the benefits of invested attention is the Surprise-Explain-Reward method [20], which leverages curiosity by surprising users (e.g., showing odd values to test spreadsheet formulas), then explains the benefits of the behavior it is trying to encourage (e.g., fewer formula errors), which is the user's reward for investing their attention. In the case of intelligent agents, benefits may take the form of an enhanced ability to control the system [11], or a more appropriate level of trust in the system.

Without clear benefits, however, users may ignore explanations altogether. For example, Bunt et al. [2] found that when users had no direct control over an agent's reasoning, user interest in *any* type of explanation was very low.

## III. Explanation Soundness and Completeness

We tease apart *soundness* and *completeness* because agent system designers can make choices independently in each as to the fidelity of their agents' explanations. The terms soundness and completeness are borrowed from the field of formal logic, in which a deductive system is *sound* if all of the statements it can create evaluate to true, and *complete* if its compositional rules allow it to generate every true statement. We apply these terms to explanations in an analogous manner:

**Soundness** ("nothing but the truth"): the extent to which *each component of an explanation's content* is truthful in describing the underlying system.

**Completeness** ("the whole truth"): the extent to which *all of the underlying system* is described by the explanation.

For example, an agent that explains its reasoning with a simpler model than it actually uses (e.g., a set of rules instead of additive feature weights) is reducing soundness, whereas an agent that explains only some of its reasoning (e.g., only a subset of a user neighborhood) is reducing completeness.

## IV. Methodology

To investigate our research questions, we presented 17 participants with up to eight music recommendations made by a functional prototype. Each recommendation came with various kinds of explanations of the system's reasons for choosing that song, and participants were asked why they thought the system made that recommendation.

### A. *Prototype Recommender System*

We developed a prototype music recommender to make personalized song recommendations for each participant. Our prototype used a hybrid recommendation approach, as such approaches have been shown to out-perform more traditional types of recommenders [17] and provide more "moving parts" to explain. Specifically, our prototype employed user-based collaborative filtering to find artists, and a content-based approach for selecting songs by those artists.

To train our recommender, we collected the listening habits of about 200,000 Last.fm listeners between July 2011 and July

2012 via the Last.fm API[1]. We identified the 50 most-played artists for each of these listeners during this time period, and then used the Mahout framework [2] to build a *k*-nearest-neighborhood (*k*=15), where distance between Last.fm users was based on overlap in the artists they listened to (calculated via the log-likelihood metric [6]).

Prior to the study, we asked each participant to imagine a situation where they would want a playlist of music, and to tell us five artists they would like to hear on it. Our prototype took these artists and, using the technique described above, recommended 20 *artists* for the given participant (Fig. 2, top). To select specific *songs*, our prototype used a bagged decision tree based on Weka's J48 implementation [7] (the bagging ensemble consisted of 100 decision trees). This classifier was independently trained for each participant using a set of positive training instances (the top 1,500 songs played by Last.fm listeners in the participant's user neighborhood) and a set of negative training instances (the top 1,500 songs played by Last.fm listeners who did *not* listen to any artists that neighborhood members listened to). This resulted in a classifier able to predict whether a given user would or would not like a particular song, along with a certainty score (Fig. 2, bottom). The song features (a *feature* is a piece of information a classifier can use to discriminate between output classes) came from The Echonest's[3] database, which includes information such as tempo, energy, and key.

To determine which songs to recommend to a participant, our prototype collected the 25 most popular songs by each recommended artist (a 500 song set). We used these songs' feature vectors as input to our classifier, which predicted whether or not the participant would like each song. The positive results were sorted by decreasing certainty, with the top eight used as song recommendations for the participant.

### B. Treatments and Explanations

Even though this was a qualitative investigation, we explored four treatments, which are shown in Table I: HH (high-soundness, high-completeness), MM (medium-soundness, medium-completeness), HSLC (high-soundness, low-completeness), and LSHC (low-soundness, high-completeness). Fig. 1 visualizes this design space, with HH in the top right, HSLC in the bottom right, LSHC in the top left, and MM in the middle. We used multiple treatments to gather data on a variety of explanation configurations, but restricted ourselves to four for feasibility.

To objectively manipulate completeness, our treatments used a varying number of the *intelligibility types* identified by Lim and Dey [12]: *inputs* (features the system is aware of), *model* (an overview of the agent's decision making process), *why* (the reasons underlying a specific decision), and *certainty* (the agent's confidence in each decision). We also increased completeness by exposing more information in the *why* (artist) intelligibility type. All treatments included explanations of the song selection process (Fig. 3), five members of the user's "neighborhood" of similar Last.fm listeners (Fig. 4), and the features the recommender could use (Fig. 5). The treatments with more completeness (MM, HH, and LSHC) added the *certainty* intelligibility type (Fig. 3, bottom left) and showed 10 members of the participant's user neighborhood. The high-completeness treatments (HH and LSHC) also added a high-level description of the recommender's algorithm (the *model* intelligibility type, Fig. 6) and showed all 15 members of the participant's user neighborhood.

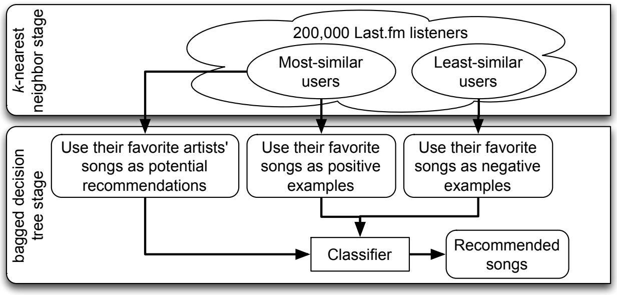

*k*-nearest neighbor stage
200,000 Last.fm listeners
Most-similar users
Least-similar users
bagged decision tree stage
Use their favorite artists' songs as potential recommendations
Use their favorite songs as positive examples
Use their favorite songs as negative examples
Classifier
Recommended songs

Fig. 2. Our prototype used a *k*-nearest neighbor stage to identify similar and dissimilar users (top), and a bagged decision tree stage to predict which songs the participant would most enjoy (bottom).

To objectively manipulate soundness, our treatments used a range of simplified models of the recommender's reasons for each song selection. The explanation used in the high-soundness treatments (HH and HSLC) described the bagged decision tree (the actual algorithm used to produce the playlist). For the medium-soundness treatment (MM), we trained a simpler model (a single J48 decision tree) using the bagged classifier's predicted labels for all of the training instances, and explained this derived model (a variation of the technique in [5]). For the low-soundness treatment (LSHC), we used the same approach to train an even simpler model (a one-feature decision tree, or *decision stump*) to explain (Fig. 3, bottom right). Because the low-soundness model only explained one highly discriminative feature, we considered it a functional analog for contemporary agent explanations (e.g., a movie recommender that explains its selections by their genres).

### C. Participants and Study Task

We recruited 17 participants (10 females, 7 males) from the local community via flyers and announcements to university mailing lists. Participants' ages ranged from 19 to 34, none had a background in computer science, and each was randomly assigned to one of the four treatments.

During the study, participants listened to their recommended playlist while a researcher provided participants with the paper explanations described in 4.B. After each song, a researcher asked the participant why they thought it had been recommended. At the end of the study we measured participants' mental models via a combination of short-answer and Likert scale questions. Each session was videotaped and later transcribed.

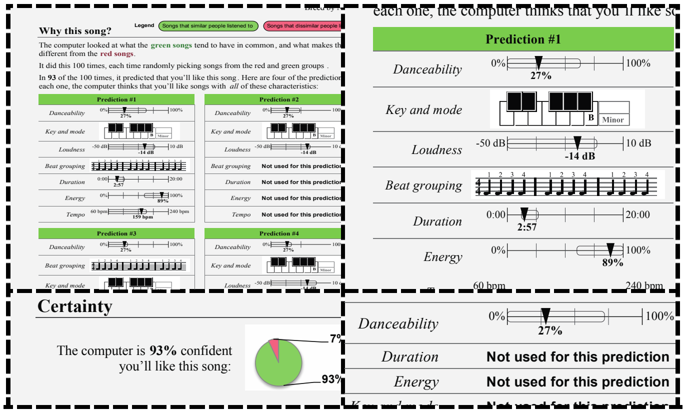
### D. Data Analysis

To qualitatively analyze the data, we developed a code set based upon how well participants understood the operation of the recommender system, plus additional codes for their knowledge gaps, produced using grounded theory methods [16]. The resulting code set is presented in Table II.

[1] http://www.last.fm/api

[2] http://mahout.apache.org

[3] http://developer.echonest.com



**Why this song?**

Legend 🟩 Songs that similar people listened to 🟥 Songs that dissimilar people l

The computer looked at what the **green songs** tend to have in common, and what makes th different from the **red songs**.

It did this 100 times, each time randomly picking songs from the red and green groups .

In **93** of the 100 times, it predicted that you'll like this song. Here are four of the prediction each one, the computer thinks that you'll like songs with *all* of these characteristics:

| Prediction #1 | |
|---|---|
| *Danceability* | 0% 27% 100% |
| *Key and mode* | B Minor |
| *Loudness* | -50 dB -14 dB 10 dB |
| *Beat grouping* | |
| *Duration* | 0:00 2:57 20:00 |
| *Energy* | 0% 89% 100% |
| *Tempo* | 60 bpm 159 bpm 240 bpm |

| Prediction #2 | |
|---|---|
| *Danceability* | 0% 27% 100 |
| *Key and mode* | B Minor |
| *Loudness* | -50 dB -14 dB 10 d |
| *Beat grouping* | Not used for this predictio |
| *Duration* | Not used for this predictio |
| *Energy* | Not used for this predictio |
| *Tempo* | Not used for this predictio |

| Prediction #3 | |
|---|---|
| *Danceability* | 0% 27% 100% |
| *Beat grouping* | |
| *Key and mode* | |

| Prediction #4 | |
|---|---|
| *Danceability* | 0% 27% 100 |
| *Key and mode* | B Minor |
| *Loudness* | -50 dB -14 dB 10 d |

each one, the computer thinks that you'll like s

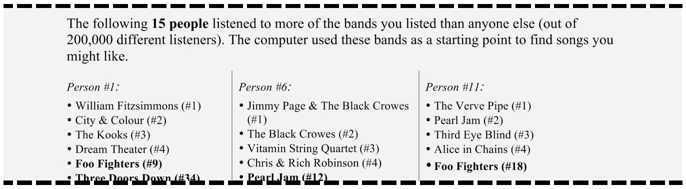
| Prediction #1 | |
|---|---|
| *Danceability* | 0% 27% 100% |
| *Key and mode* | B Minor |
| *Loudness* | -50 dB -14 dB 10 dB |
| *Beat grouping* | |
| *Duration* | 0:00 2:57 20:00 |
| *Energy* | 0% 89% 100% |
| | 60 bpm 240 bpm |

**Certainty**

The computer is **93%** confident you'll like this song:

| | |
|---|---|
| *Danceability* | 0% 27% 100% |
| *Duration* | **Not used for this prediction** |
| *Energy* | **Not used for this prediction** |

Fig. 3. Excerpts from the Why this Song explanation (*why* intelligibility type). (Top left): The high-soundness sheet showed a random sample of decision trees from the bagging ensemble.
(Top right): Each tree was represented as a set of ordered features with allowed ranges of values. The medium soundness sheet was similar, but only showed one derived decision tree that approximated the bagging ensemble's reasoning.
(Bottom right): The low soundness sheet was also similar, but only showed one derived decision stump (single-featured tree).
(Bottom left): For the HH, LSHC, and MM treatments, this sheet also included the *certainty* intelligibility type.



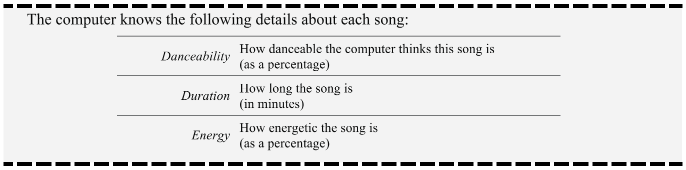
The following **15 people** listened to more of the bands you listed than anyone else (out of 200,000 different listeners). The computer used these bands as a starting point to find songs you might like.

*Person #1:*
- William Fitzsimmons (#1)
- City & Colour (#2)
- The Kooks (#3)
- Dream Theater (#4)
- **Foo Fighters (#9)**
- **Three Doors Down (#34)**

*Person #6:*
- Jimmy Page & The Black Crowes (#1)
- The Black Crowes (#2)
- Vitamin String Quartet (#3)
- Chris & Rich Robinson (#4)
- **Pearl Jam (#12)**

*Person #11:*
- The Verve Pipe (#1)
- Pearl Jam (#2)
- Third Eye Blind (#3)
- Alice in Chains (#4)
- **Foo Fighters (#18)**

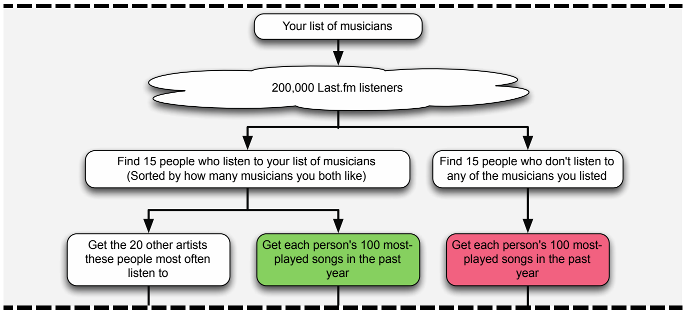
Fig. 4. Excerpt from Why this Artist (*why* intelligibility type), which showed the artists selected by their user neighborhood. All participants received this explanation, but with different neighborhood sizes (see Table I).



The computer knows the following details about each song:

| | |
|---|---|
| *Danceability* | How danceable the computer thinks this song is (as a percentage) |
| *Duration* | How long the song is (in minutes) |
| *Energy* | How energetic the song is (as a percentage) |

Fig. 5. Excerpt from What the Computer Knows (*input* intelligibility type), which showed a comprehensive list of features that the recommender used. All participants received this explanation.



Your list of musicians
200,000 Last.fm listeners
Find 15 people who listen to your list of musicians (Sorted by how many musicians you both like)
Find 15 people who don't listen to any of the musicians you listed
Get the 20 other artists these people most often listen to
Get each person's 100 most-played songs in the past year
Get each person's 100 most-played songs in the past year

Fig. 6. Excerpt from How it All Works Together (*model* intelligibility type), which showed how the participants' artists list was used to make song recommendations. (Positive and negative training sets were color-coded throughout the flow-chart.) Only HH and LSHC participants received this explanation.

We transcribed participant utterances during each song and applied the codes to these utterances (each code could be applied, at most, once per song). Two researchers independently coded a small portion of the transcripts and then discussed areas of disagreement. Once the researchers agreed on the coding of these transcripts, they independently coded two complete transcripts (12% of the data)—their agreement, as calculated by the Jaccard index (the intersection of all applied codes over the union of all applied codes), was 83%. Given this acceptable level of agreement, a single researcher coded the remaining transcripts and post-study questionnaires.

TABLE I. THE "WHY (SONG)" INTELLIGIBILITY TYPE WAS AVAILABLE IN ALL TREATMENTS, BUT ITS SOUNDNESS VARIED. THE OTHER INTELLIGIBILITY TYPES WERE USED TO VARY COMPLETENESS.

| | **Treatment** | HH | MM | HSLC | LSHC |
|---|---|---|---|---|---|
| | **Relative Soundness** | High | Medium | High | Low |
| | **Relative Completeness** | High | Medium | Low | High |
| **Intelligibility types** | **Why (song)** | Bagged decision tree | Decision tree | Bagged decision tree | Decision stump |
| | **Why (artist)** | Nearest neighbor (*k*=15) | Nearest neighbor (*k*=10) | Nearest neighbor (*k*=5) | Nearest neighbor (*k*=15) |
| | **Certainty** | Yes | Yes | No | Yes |
| | **Model** | Yes | No | No | Yes |
| | **Input** | Yes | Yes | Yes | Yes |

Participants' mental model "scores" were the number of correct minus the number of incorrect statements participants made during the experiment and on the post-study questionnaire, translated to a 0-to-10 (lowest-to-highest) scale. Table II shows which types of verbalizations/responses were correct vs. incorrect. Participants' verbalizations during the study and post-study questionnaire responses were weighted equally.

## V. RESULTS

### A. RQ-1 and RQ-2: Soundness, Completeness, and Types

As Fig. 7 shows, HH participants achieved three of the top four scores. In contrast, all but one of the participants in the other treatments clustered around lower scores. This surprised us because we had expected the HH treatment to overload participants to the point where they would not attend to so much complex information. Instead, we expected the MM treatment to be a "sweet spot" in the trade-off between informativeness and simplicity—but most of the MM participants clustered around the *lowest* scores.

Further, HH participants' mental model scores were consistently high across features and processes, as Fig. 8's results from the post-task questionnaire show. In fact, HH participants were the only ones to correctly describe the song selection process (third column of Fig. 8, coded as per Table II), and only one HH participant made any incorrect post-task observations at all (right half of Fig. 8). (Note from Table II that participants in *any* of the treatments could potentially get credit for process descriptions that had correct process concepts, e.g., using combinations of features.)

#### 1) Completeness and Intelligibility Types

Two of the intelligibility types, *why* and *input*, relate to features, and participants tended to do better at understanding features than process (Fig. 8). However, a closer look at *which*

TABLE II. CODE SET USED TO ASSESS PARTICIPANTS' MENTAL MODELS.

| Category | Code | Participant discussed/said… |
|---|---|---|
| *Correct:* the participant correctly discussed an aspect of the recommender | Valid artist process | the artist was chosen via collaborative filtering |
| | Valid song feature | specific song features used by the recommender |
| | Valid song process | a combination of features were responsible for the recommendation |
| *Incorrect:* the participant incorrectly discussed an aspect of the recommender | Invalid feature | specific features not used by the recommender |
| | Invalid process | the computer's reasoning involved a single path through a decision tree or another incorrect description of the artist/song selection process. |
| *Knowledge gaps:* the participant expressed uncertainty about their knowledge of the recommender | Don't know | not knowing how the recommender works |
| | Uncertain | uncertainty regarding their answer of how the recommender works |
| | More explanation details | needing more details about the explanations |
| | More recommender details | needing more details about the recommender |

participants did better suggests that their understanding of features aligned with completeness. For example, participants in the high-completeness groups (HH and LSHC) averaged 5.5 valid feature codes per participant, versus the other treatments' average of 4.3. The invalid features added more evidence consistent with this, with high-completeness participants averaging 4.6 invalid features versus other participants' 6.3 invalid features.

Completeness may also have helped participants understand the recommendation process. As Fig. 9 shows, participants' understanding (as per Table II codes) of the *artist* recommendation process (explained through the *model* and *why-artist* intelligibility types) tended to increase with the completeness of their treatment. In particular, the *model* explanation was referenced by half of the participants who correctly discussed the artist recommendation process (Fig. 10). Completeness showed no evidence of impacting participant understanding of the *song* recommendation process; however, this was primarily explained via the Why this Song explanation, and this explanation did not vary in the completeness dimension across treatments.

Recall that we also increased completeness by adding the *certainty* intelligibility type, but this type did not seem to interest participants: only two participants mentioned certainty at all, and each did so only once. Although research has shown that certainty is a useful intelligibility type to users assessing an intelligent agent's *reliability* [10], other researchers have found that certainty does not help users' *perceived understanding* of how a recommender operates [4]. Our work suggests that this finding extends to *actual understanding*.

These results suggest that increasing completeness was beneficial to participants' mental models, and that some effective ways to increase completeness included the *model* intelligibility type and the completeness of the *why* type. However, we found no evidence that increasing completeness via *certainty* improved participants' mental models.

### 2) *Soundness and Intelligibility Types*

Although HH participants' performance may at first glance suggest that high soundness was also helpful, looking at soundness in isolation suggests a different story. High-soundness participants (HH and HSLC) showed almost no differences from the other participants in their mentions of valid vs. invalid features or processes. Instead, the clearest pattern was one of *decreased* understanding of the *artist* recommendation process as soundness increased (Fig. 9).

One hypothesis is that HH and HSLC participants spent most of their attention on their complex Why this Song explanations, causing them to ignore other explanations. Indeed, participants in these high soundness treatments viewed the How it All Works explanation only about half as often as participants in the low-soundness treatment (mean 0.8 vs. 1.4 views per person). Instead, they focused on their complex Why this Song explanations: they viewed these during more songs than participants in the low-soundness treatment (mean of 7.6 vs. 6.3 songs) and often even reviewed prior Why this Song explanations (during an average of 1.9 songs vs. 0.7). P9-HH

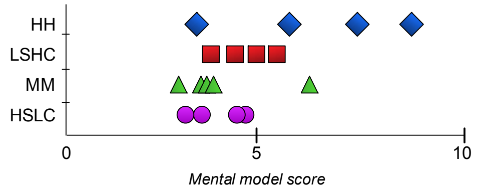


Fig. 7. Participants' mental model fidelity scores. Each mark is one participant's score. (Note: MM had one more participant than the others.) The highest scores were mostly those of HH participants.

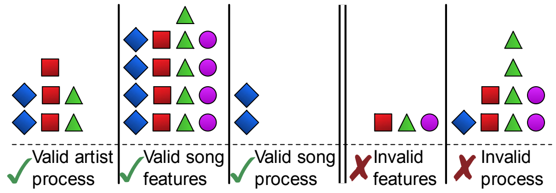


Fig. 8. Post-task questionnaire results. Each mark is one participant, represented as in Fig. 7. Only HH participants described all the valid aspects of the recommender (left), and only one made an invalid description (right).

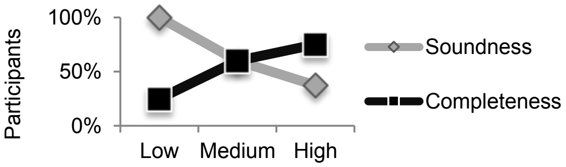


Fig. 9. Each dot is the percentage of participants who correctly understood the artist recommendation process (as per Table II's codes). More participants understood it as completeness (black) increased, but fewer participants understood it as soundness (light) increased.

explained why she kept reviewing prior explanations:

*P9-HH: "The [high-soundness Why this Song] sheet is a little bit hard to look at [flips through prior Why this Song sheets], but I'm just looking for things that I'm seeing, from one song to the next, that are similar, and that it says it's using for matching and making the predictions."*

Further, the high-soundness explanations were associated with over twice as many Information Gap codes, which indicate that as participants viewed these explanations, they had additional questions and expressed more uncertainty as they described why they thought each song had been recommended (mean 7.0 codes per participant) than other treatments (mean 3.3 codes per participant).

Thus, increasing soundness may risk over-complexity, but increasing completeness alongside soundness may mitigate this effect. While HH participants (those with high soundness *and* high completeness) ended our study with the best mental models, the HSLC (high-soundness and low-completeness) participants' models were among the worst (Fig. 7).

### B. RQ-3: Barriers to High-Fidelity Mental Models

No participant's understanding of the recommender was perfect: the highest mental model score was 8.4 out of 10 (recall Fig. 7). We found evidence of two barriers to building high-fidelity mental models; these barriers were shared among all participants, regardless of treatment.

First was participants' incorrect assumptions about the explanations' completeness. Every participant, at some point during their task, incorrectly assumed that the recommender used information that it did not have access to (e.g., the tone of the singer's voice)—even though the *input* explanation (What the Computer Knows) was complete across all treatments. For example, participant P6-HSLC had read the What the Computer Knows explanation multiple times before asking:

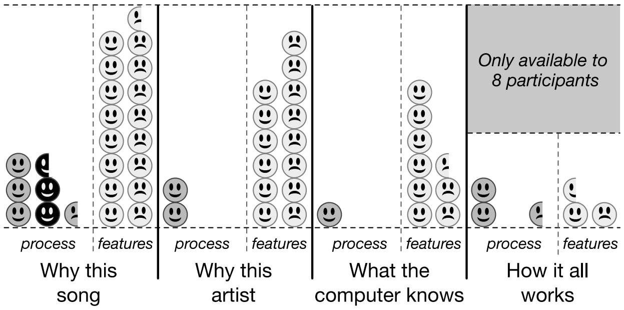


Fig. 10. Participants giving correct (smiles) and incorrect (frowns) descriptions upon referencing an explanation. Each face = 2 participants. (Light): song features. (Gray): artist recommendation process. (Black): song recommendation process. Both *why* explanations were popular, but What the Computer Knows produced fewer invalid *features*, and How it All Works had the highest percentage of participants correctly describing the *process*.

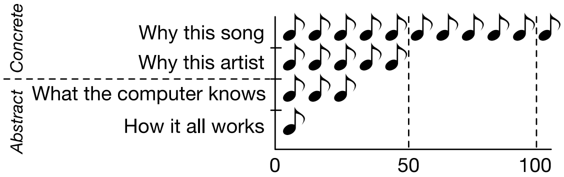


Fig. 11. Number of times participants referenced each explanation: each music note = 10 references. Participants referenced the Why this Song explanation during almost every recommended song.

*P6-HSLC: "So I guess, does a computer have access to lyrics for a song, does it take that into consideration?"*

*[Facilitator refuses to answer, and participant re-reads the What the Computer Knows sheet yet again.]*

*P6-HSLC: "Oh right, so probably not then."*

The counts from the post-session questionnaire results were consistent with this phenomenon. In responding to a question asking if the explanations included every important detail about why a song was recommended, the average response was only 13.0 (21 indicating "always", 0 indicating "never"). HH participants, however, responded more positively (mean of 18.0), suggesting that high soundness and high completeness together can help convince users that the explanations *do* discuss everything relevant to the agent's reasoning.

The second barrier was lack of knowledge of the *process* of how recommendations were made. Participants rarely discussed process, focusing much more heavily on features, as Fig. 10 illustrates. Some participants even described a single feature as the sole reason for a recommendation:

*P2-HH: "Yeah, see, it's all the way at the bottom of the loudness [feature]. So… that's why [it was recommended]."*

Features may have been easier for participants to understand because they were explained concretely (i.e., in the context of specific examples). Fig. 11 shows that participants used the concrete Why this Song and Why this Artist explanations much more than the abstract (i.e., no specific examples) How it All Works and What the Computer Knows explanations.

Note, however, that although our abstract How it All Works explanation was infrequently used, when participants *did* use it, a larger percentage (50%) correctly discussed the recommendation process than with any other explanation (Fig. 10). Similarly, participants who used the abstract What the Computer Knows explanation discussed more valid features than invalid features. Perhaps abstract explanations may be best made available on demand via a layering approach, in which users can "drill up" from a concrete explanation to view more abstract details.

Alternatively, participants may have paid the most attention to the Why this Song explanations because it was the only explanation that changed during the experiment. The other explanation types were presented at the beginning of the study and may have attracted less participant attention because they were never updated. Dynamically updating explanations may be one presentation option to draw a user's attention to the full range of explanations in a highly complete system, but this is an open question that requires further investigation.

### C. RQ-4: Is It Worth It?

The Attention Investment Model [1] predicts that users will use high-cost explanations *only if* they think the benefits will outweigh the costs. Thus, we investigated participants' perceived benefits (given the perceived costs) using the questions "If the recommendations improve, do you think it is worth the time and effort you spent during this study to give feedback to the recommender?" and "Would you take a similar amount of time as this study to learn similar things about other recommenders you use?" (Each study session lasted less than two hours.) We used the summation of these questions to estimate perceived benefits, and the summation of the NASA-TLX questions about mental demand, effort expended, and frustration/annoyance to estimate costs (each question had a 21-point scale).

As Fig. 12 shows, the LSHC participants were surprisingly positive about the benefits vs. costs of referring to the explanations—more than three times as positive as participants viewing less complete but more sound explanations (MM and HSLC). We had expected the MM treatment to best balance costs vs. benefits—these participants received explanations that seemed likely to be the easiest to understand at a reasonable cost. However, our results showed that instead, high completeness seemed to be important to our participants. To summarize Fig. 12, participants in the two high-completeness treatments perceived working with the explanations to be a better cost/benefit proposition than the other treatments' participants did. In contrast, soundness did not seem to be an asset to participants' perception of cost-benefit. This may come back to the lower understanding associated with higher soundness (recall Fig. 9). P6-HSLC reinforced this point, remarking that the high-soundness explanations *could* have been useful, but she was unable to make much sense of them during the study:

*P6-HSLC: Probably should have looked at [the Why this Song sheet] more.*

*Facilitator: Do you think this could have been useful?*

*P6-HSLC: Yeah… I guess I'm still trying to grasp and understand this whole thing here (points at Why this Song sheet).*

### *D. RQ-5: In Explanations We Trust?*

To some low-soundness participants, the decision stump in their explanations seemed clearly wrong. For example:

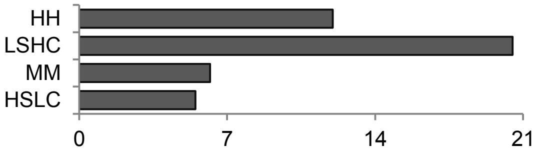


Fig. 12. Perceived benefit vs, cost scores (*benefit score – cost score*), averaged by treatment. The high-completeness participants (top two rows) perceived relatively high benefits vs. costs of the explanations.

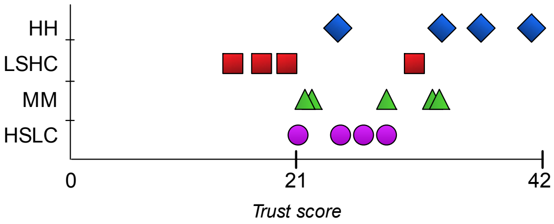


Fig. 13. Trust scores for each participant. The LSHC treatment's scores were relatively low: these participants accurately rated their explanations as unsound, but also *inaccurately* rated them as incomplete.

*P13-LSHC: "It says loudness again, I'm really not understanding why it keeps going back to that and not using energy, or like, anything else."*

To understand participants' perceptions of whether the explanations they viewed were sound and complete, we asked them "Do you think the explanations are accurate about why the recommender chose each song?" (perceived soundness), and "Do you think the explanations are including all of the important information about why the recommender chose each song?" (perceived completeness). We asked about soundness and completeness separately to determine whether participants could discern whether explanations were sound, complete, or both. For example, we hypothesized LSHC participants would rate their explanations as more complete than sound, while HSLC participants would consider their explanations more sound than complete. However, our results suggest that participants did not differentiate explanations in this way: the average difference between the two scores was only 1.5 on a 21-point scale, and both LSHC and HSLC participants rated their explanations as slightly more sound than complete.

Because the perceived soundness and completeness scores together form a holistic assessment of trust, we summed them to yield a single trust score. The results, plotted for each participant, are shown in Fig. 13. The LSHC participants had the three lowest trust ratings, while most HH participants accurately gauged their explanations to be the most sound and most complete. This suggests there is some danger to simplifying explanations by reducing soundness—users may perceive that such explanations do not accurately represent the system's reasoning, and so may distrust (and disregard) them.

## VI. Discussion

Our results suggest that the most sound and most complete explanations (HH) were the most successful at helping participants understand how the agent worked, and did so with a surprisingly good cost/benefit ratio. Further, HH participants trusted their explanations more than participants in other treatments, particularly LSHC. Indeed, the main problem we identified with HH was that participants were at risk of focusing on a single complex explanation to the exclusion of other information.

The story was different when only soundness *or* completeness was at our highest level. High completeness alone (LSHC) provided participants with the best perceived cost/benefit ratio of attending to the explanations, the second-highest average mental model score, and the best understanding of the artist recommendation process. However, these participants placed the least trust in the explanations. High soundness alone (HSLC) did result in more trust, but was also associated with higher perceived costs, lower perceived benefits, and flawed mental models.

Overall, we found that presenting explanations in a sound and complete manner is a surprisingly good design choice, even for relatively low-benefit agents such as media/product recommendation, when they go wrong. (Indeed, we saw a slightly *negative* relationship between mental model score and user satisfaction with the recommendations, suggesting that the hope of improving even such low-benefit agents may be

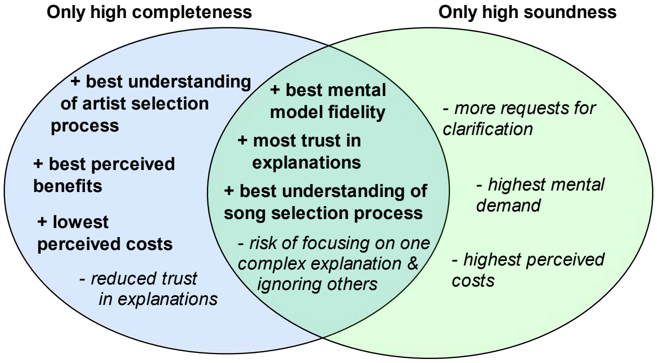

Only high completeness
Only high soundness
+ best understanding of artist selection process
+ best perceived benefits
+ lowest perceived costs
- reduced trust in explanations
+ best mental model fidelity
+ most trust in explanations
+ best understanding of song selection process
- risk of focusing on one complex explanation & ignoring others
- more requests for clarification
- highest mental demand
- highest perceived costs

Fig. 14. The benefits (bold) and costs (italics) of our highest completeness and highest soundness treatments. All of the benefits required high completeness (left and middle), but many of the costs were only observed when soundness was high but completeness was low (right).

sufficient motivation for users to learn more about the agent.) However, if a designer's user testing of an agent system reveals that its target audience believes such explanations are not worth attending to, our findings suggest that reducing soundness while preserving completeness will improve the cost/benefit ratio of attending to explanations. Fig. 14 summarizes what tool designers may expect to see when presenting end users (like our participants) with explanations that are very sound, very complete, or both.

## VII. Conclusion

Part of enabling end users to "debug" their intelligent agents is explaining these agents to users well enough for them to build useful mental models. In this paper we considered two dimensions of explanations—soundness and completeness—and explored how each impacts end users' mental model fidelity, their perceptions of the cost/benefit trade-off of attending to these explanations, and their trust in the explanations. Among our findings were:

*RQ-1 (Soundness and Completeness):* Our most complete explanations (HH and LSHC) were associated with the best mental models; reduced completeness was the shared feature of the two worst-performing treatments (HSLC and MM).

*RQ-2 and RQ-3 (Explanations and Obstacles):* Participants had more difficulty understanding the agent's reasoning process than the features it used, but *abstract* explanations of the *model* intelligibility type helped overcome this obstacle. However, participants appeared to prefer more *concrete* explanations (recall Fig. 11).

*RQ-4 (Costs and Benefits):* Our most complete explanations were associated with the highest perceived benefits and lowest perceived costs of learning about the system; completeness even helped moderate the cost of very sound explanations (as in HH).

*RQ-5 (Trust):* Participants correctly perceived that the LSHC explanations were unsound, but also refused to trust that these explanations were complete. Participants placed the most trust in HH explanations.

These findings suggest that many popular intelligent agents offer explanations that are too low in fidelity to enable users to understand how they work, and show how different intelligibility types (e.g., *why*, *model,* etc.) can increase explanation fidelity, and with it user's mental models. Further, our cost/benefit results show that users *want* to learn more about these systems *if* their effort is rewarded with the ability to improve their intelligent agents. Thus, increasing explanation fidelity can be a win/win for end users—motivated users can learn how their agents operate, and then employ that knowledge to personalize their agents' reasoning.

## Acknowledgments

We thank our study participants for their time, and Romina Rodriguez and Amber Hovarth for evaluating our prototype.

## References

[1] Blackwell, A. (2002). First steps in programming: a rationale for attention investment models. In *Proc. HCC,* 2–10.

[2] Bunt, A., Lount, M., & Lauzon, C. (2012). Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In *Proc. IUI*, 169–178.

[3] Carroll, J., & Rosson, M. (1987). Paradox of the active user. In J. M. Carroll (Ed.), *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction* (pp. 80–111). The MIT Press.

[4] Cramer, H., Evers, V., Ramlal, S., Someren, M., Rutledge, L., Stash, N., et al. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496.

[5] Craven, M., & Shavlik, J. (1997). Using neural networks for data mining. *Future Generation Computer Systems*, *13*(2-3), 211–229.

[6] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

[7] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

[8] Herlocker, J., Konstan, J., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proc. CSCW*, 241–250.

[9] Johnson-Laird, P. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press.

[10] Kulesza, T., Burnett, M., Stumpf, S., Wong, W., Das, S., Groce, A., Shinsel, A., Bice, F., & McIntosh, K. (2011). Where are my intelligent assistant's mistakes? A systematic testing approach. In *Proc. IS-EUD*, 171–186.

[11] Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012). Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proc. CHI*, 1–10.

[12] Lim, B., & Dey, A. (2009). Assessing demand for intelligibility in context-aware applications. In *Proc. Ubicomp*, 195–204.

[13] Lim, B., Dey, A., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proc. CHI*, 2119–2128.

[14] Norman, D. (1983). Some Observations on Mental Models. In D. Gentner & A. Stevens (Eds.), *Mental Models*. Psychology Press, 7–14.

[15] Rosson, M., Carrol, J., & Bellamy, R. (1990). Smalltalk scaffolding: a case study of minimalist instruction. In *Proc. CHI*, 423–430.

[16] Strauss, A., & Corbin, J. (1994). Grounded theory methodology. *Handbook of Qualitative Research*, 273–285.

[17] Su, X., & Khoshgoftaar, T. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, *2009*(12), 1–19.

[18] Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. *Recommender Systems Handbook*, 479–510.

[19] Tullio, J., Dey, A., Chalecki, J., & Fogarty, J. (2007). How it works: a field study of non-technical users interacting with an intelligent system. In *Proc. CHI*, 31–40.

[20] Wilson, A., Burnett, M., Beckwith, L., Granatir, O., Casburn, L., Cook, C., Durham, M., & Rothermel, G. (2003). Harnessing curiosity to increase correctness in end-user programming. In *Proc. CHI*, 305–312.