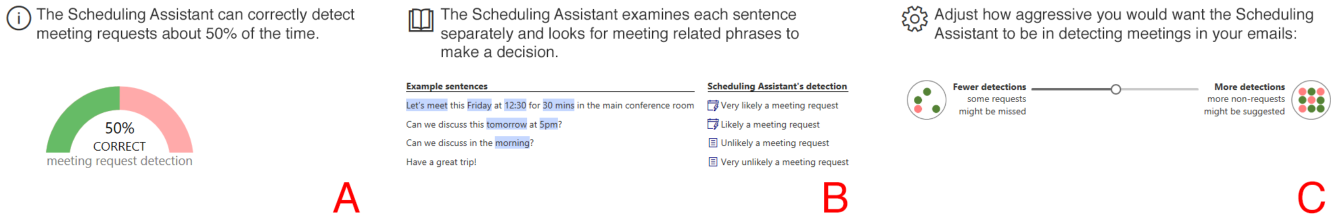
Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems

Rafal Kocielnik
University of Washington
Seattle, USA
rafal.kocielnik@gmail.com

Saleema Amershi
Microsoft Research
Redmond, USA
samershi@microsoft.com

Paul N. Bennett
Microsoft Research
Redmond, USA
paul.n.bennett@microsoft.com


The Scheduling Assistant can correctly detect meeting requests about 50% of the time.

The Scheduling Assistant examines each sentence separately and looks for meeting related phrases to make a decision.

Example sentences	Scheduling Assistant's detection
Let's meet this Friday at 12:30 for 30 mins in the main conference room	Very likely a meeting request
Can we discuss this tomorrow at 5pm?	Likely a meeting request
Can we discuss in the morning?	Unlikely a meeting request
Have a great trip!	Very unlikely a meeting request

Adjust how aggressive you would want the Scheduling Assistant to be in detecting meetings in your emails:

Figure 1: Expectation setting design techniques used prior to interaction with the Scheduling Assistant - an AI system for meeting request detection from free-text of emails. A) Accuracy Indicator - directly communicating to the user the expected accuracy of the AI component, B) Example-based Explanation - helping the user understand the basic principles of how the systems detects meeting requests, C) Control - giving the user control over AI decision making process through detection threshold adjustment.

ABSTRACT

AI technologies have been incorporated into many end-user applications. However, expectations of the capabilities of such systems vary among people. Furthermore, bloated expectations have been identified as negatively affecting perception and acceptance of such systems. Although the intelligibility of ML algorithms has been well studied, there has been little work on methods for setting appropriate expectations before the initial use of an AI-based system. In this work, we use a Scheduling Assistant - an AI system for automated meeting request detection in free-text email - to study the impact of several methods of expectation setting. We explore two versions of this system with the same 50% level of accuracy of the AI component but each designed with a different focus on the types of errors to avoid (avoiding False Positives vs. False Negatives). We show that such different focus can lead to vastly different subjective perceptions of accuracy and acceptance. Further, we design expectation adjustment techniques that prepare users for AI imperfections and result in a significant increase in acceptance.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; *Empirical studies in visualization*; Laboratory experiments;

KEYWORDS

AI infused systems, AI system on-boarding, Shaping AI expectations, Perception and Acceptance of AI

ACM Reference Format:
Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3290605.3300641

1 INTRODUCTION

Expectations impact how accepting end-users are of the technologies they use. For example, inflated expectations about usability and ease of use have been shown to decrease user satisfaction and willingness to use products when those expectations are not met [20, 36]. Artificial intelligence (AI) introduces additional factors impacting user expectations

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05...$15.00
https://doi.org/10.1145/3290605.3300641

and acceptance of modern AI-powered technologies. Specifically, the underlying algorithms driving AI functionalities, such as natural language understanding [12, 14, 30], sensor-based inferences [28, 29], web behavior prediction [27], or object recognition in video or images [25], are probabilistic and almost always operate at less than perfect accuracy. However, most users do not expect their applications to behave inconsistently and imperfectly [15] which can lead to disappointment and potential abandonment of these technologies [41, 51]

Previous work has shown that end-user expectations of technology are impacted by a variety of factors including external information, knowledge and understanding, and first hand experience. For example, Olshavsky & Miller [38] showed that giving a description of a product that emphasizes lack of care in its design can lead to negative expectations, while a description that communicates the opposite can lead to more positive expectations. However, outside of work in marketing and advertising [11, 32, 38], few have explored how end-user expectations can be directly and explicitly shaped.

In this work, we explore techniques for shaping end-user expectations of AI-powered technologies prior to use and study how that shaping impacts user acceptance of those technologies. Each of the proposed techniques is designed to affect different aspects of expectation forming: external information, understanding, first-hand experience through a sense of control. We test these techniques within the context of an email application that includes an AI-powered Scheduling Assistant that automatically detects meeting requests and helps users schedule calendar appointments. We also investigate how these techniques are impacted by different types of AI imperfections (False Positives vs. False Negatives). Testing the impact of different types of errors is important because people may perceive these differently and as noted in prior work [13], there is a lack of work in the literature which investigates the impact of False Positives and False Negatives on the UX – a critical question for experiences which depend on machine learning. In two studies with 150 and 400 participants respectively carried out on an internal crowd-sourcing platform similar to Mechanical Turk, we show that our proposed techniques successfully impact key aspects of user expectations. Further, we show that False Positive errors are more accepted by participants in the context of using the Scheduling Assistant, which we hypothesize is due to lower cost of recovery from this type of error in our application. Finally, we show that our expectation shaping techniques significantly increase user satisfaction and acceptance when the AI is tuned to avoid False Positive errors.

Our contributions are as follows:

(1) We propose three techniques for setting expectations: 1) Accuracy Indicator - that explicitly states the accuracy of the system, 2) Examples based Explanation - that seeks to increase user understanding, and 3) Performance Control - that allows the user to directly adjust the performance of the system.
(2) We demonstrate the effectiveness of our three techniques and show their ability to preserve user satisfaction and acceptance of an imperfect AI-powered Scheduling Assistant we implemented.
(3) We show that an AI-powered system tuned to avoid False Positive errors can lead to lower average perception of accuracy and lower user acceptance than one tuned to avoid False Negatives (even when both versions perform at the same overall accuracy). We discuss the likely role of the cost of error recovery in this scenario in determining the best balance between False Positives and False Negatives.

2 RELATED WORK

Expectations Research in HCI and Marketing

The topic of expectations, albeit not specifically in relation to AI, has been explored in areas of product marketing [3, 8], information systems [49], and recently also HCI [36].

Marketing and psychology studies have focused on expectations related to purchase of physical products [3, 8]. That work focused mostly on immediate pre- and post- purchase behaviors of customers. It explored the impact of initial customer expectations about the product on post-purchase satisfaction and willingness to use the product [11]. The main conclusion is that too high of expectations lead to dissatisfaction, while very low expectations followed by positive product use experience can boost satisfaction. While those works contributed to formulation of various theories related to the impact of expectations, they explored relatively little about feasible mechanisms for explicitly shaping expectations. As the focus was on establishing a link between expectations and customer satisfaction, expectations were set by means of deception (selecting only negative or positive reviews of a product [39], distorting product descriptions using negative or positive keywords [11], or deliberate under or over statements of product features [3]). While effective for one time purchasing decisions, deception can have ethical implications. Moreover, we posited that deception would be less effective in our scenario when ongoing use would eventually reveal the true performance of a system.

Relatively recent and still limited HCI work adapted and extended some of the findings from these early marketing studies into use of digital systems, such as games [36], websites [20], and mobile payment services [31]. This work has shown that the link between expectations, user satisfaction, and user acceptance can be extended to modern digital services. It has also shown that expectations of different aspects of user experience (i.e., usability and usefulness) can be examined separately and further extended the knowledge about the impact of expectations to long-term use of digital systems [31]. At the same time, these works mainly adopted deception based approaches of setting user expectations, resorting to two classes of manipulations - priming - selectively emphasizing negative or positive aspects of a system or - framing - distorting the information presented to the users. Furthermore, they did not explore expectation of properties crucial and unique to the AI-powered systems, such as the accuracy with which it can operate, the successes and mistakes it can make, the likely reasons for those, and the impact of user actions (e.g., the contents user generates) on AI-powered systems behavior.

Intelligibility and Transparency of AI Algorithms

A significant body of research exists on the topic of intelligibility of AI systems in particular [1, 35, 40, 46]. Lim et al. explored 7 intelligibility types aimed at communicating detailed mechanisms of how a simulated decision tree based machine learning (ML) algorithm in a hypothetical activity recognition system makes its decisions [33–35]. Parallel work explored the topic of comprehensibility of ML algorithms to domain experts (e.g., expert biologists not knowledgeable about ML) [46]. A number of these proposed approaches have been shown to positively impact transparency and trust. In most cases, however, transparency techniques are intended for post fact explanations about the decisions of an AI or ML system [19, 22], rather than for adjusting user expectations of an AI-powered system prior to its use. A few recent works have endeavored to extend the explanation approach to pre-use expectation adjustment. These works automatically select informative examples prior to actual system use [2, 23, 40] to give the user an intuition about the system behavior or attempt to summarize internal system decision rules [21]. They, however, do not directly investigate impact on acceptance and explore only one strategy of shaping expectations, which is by educating the user about system behavior via examples. Most proposed approaches also require significant end-user interest and effort in understanding system behavior, which can be inefficient in many scenarios [7].

Furthermore, the strategies proposed in many of these works, with notable exceptions of [18, 40], rely on a technical ability to generate end-user explanations from the algorithms themselves (i.e., in Lim et al. [35] decision tree path traversal was used to generate free-text description presented to the user). However, the shift to deep learning models in many end-user applications makes it practically infeasible to apply many intelligibility types proposed in earlier work [51]. The techniques we proposed in this work are in principle algorithm agnostic.

Theories Related to Expectations

A number of theories related to expectations have been proposed [6, 37, 39]. One prominent theory that has been tested in various studies, and used in marketing and HCI work described earlier, is the Expectation Confirmation Model (ECM) [6]. This model postulates that user satisfaction and acceptance of a system is directly related to the difference between initial expectations and their actual experience. Specifically a negative dis-confirmation of initial expectations (i.e., when a user expects more than the system can deliver) leads to lower satisfaction and decreased acceptance. Hence we designed our proposed techniques to achieve more accurate expectations of system capabilities.

While the ECM describes a relationship between a change in expectations and user satisfaction/acceptance of a product or system, it provides little guidance as to how user expectations can be shaped. ECM posits only that *"expectations are influenced by many sources - advertisements, brands, word of mouth, product reviews, discussion forums, and exposure to related products"* [39]. It has, however, linked expectations to the general theory of attitudes and beliefs [17] and defined expectations as a sum of beliefs about the level of product or service [6]. In the context of AI-powered systems, these could be beliefs about how well such a system can work. According to this prior work there are three major mechanisms in which user beliefs contributing to expectations are formed: 1) through information from external sources (e.g., being directly told about the specific properties of a system by a third party), 2) reasoning and understanding (e.g., forming expectations as an extension of understanding of how the system works), 3) first hand experience (e.g., forming an action-effect association that comes from direct interaction and experience with a system). These mechanisms inspired the design of our expectation adjustment techniques.

3 RESEARCH QUESTIONS AND HYPOTHESES

An AI system can operate at the same level of overall accuracy (total number of correct predictions over all possible predictions), yet produce different proportions of two types of errors: False Positives or False Negatives [13]. These are typically quantified as precision vs. recall. In many cases, existing systems are optimized for high precision and therefore avoid False Positives (e.g., avoid recommending a movie user may not like). Anecdotal reports from practitioners indicate a general belief that avoiding False Positives is considered better for user experience, and previous work [13] notes that the impact of False Positives vs False Negatives on UX is generally uninvestigated. We therefore explore the following research question and associated hypothesis: **RQ1. What is the impact of an AI system's focus on avoidance of different types of errors on user perception?**

- **H1.1** An AI system focused on High Precision (low False Positives) will result in higher perceptions of accuracy and higher acceptance.

Prior work has shown three major contributors to user expectations: information from external sources, reasoning and understanding, and first hand experience. Hence, our next research question explores design techniques for achieving these mechanisms: **RQ2. What are the design techniques for setting appropriate end-user expectations of AI systems?**

- **H2.1** Directly communicating AI system accuracy will lead to lower discrepancy between system accuracy and user perception of it.
- **H2.2** Providing explanations will lead to higher perceptions of understanding how the AI system works.
- **H2.3** First-hand experience, through direct impact on the system, will lead to higher perceived level of control over system's behavior.

Finally, we expect that more accurate expectations of an AI system's capabilities should result in users being better prepared for AI system imperfections and therefore result in higher satisfaction and acceptance. We therefore explore the following research question and associated hypothesis: **RQ3. What is the impact of expectation-setting intervention techniques on user satisfaction and acceptance of an AI system?**

- **H3.1.** Presence of an intervention prior to use of an imperfect AI system will lead to higher acceptance and satisfaction compared to lack of such intervention.

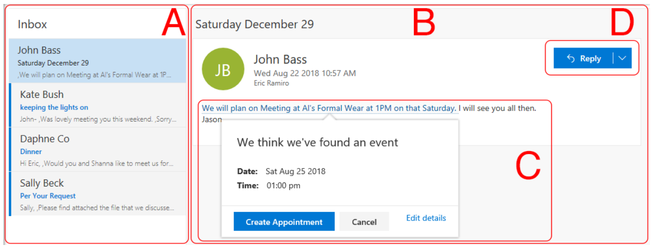
4 AI POWERED SCHEDULING ASSISTANT

The Scheduling Assistant is an AI-powered application that mimics the Inbox of a web version of a popular email client - MS Outlook. We chose to recreate this environment in order to freely control elements of the interface and the underlying free-text meeting detection and highlighting AI functionality. The Scheduling Assistant operates as a website that can be accessed via web browser.

User Interface

The user interface of the Scheduling Assistant mimics the MS Outlook's web client interface as can be seen in Figure 2.

Users can see a list of emails in the Inbox (see A in Figure 2). As in MS Outlook, unread emails are marked with a left side blue bar indicator (e.g., third email from the top). Once the user clicks and views the email, it will no longer be highlighted with a blue bar (e.g., first email from the top). When an email is clicked, its contents along with the subject, name of the sender, the time it was sent, and the contents can be viewed (see B in Figure 2). The AI functionality offers automated highlighting of meeting request sentences in email contents. When a user clicks a highlight, a pop-up dialog offers a shortcut to placing the meeting in the user's calendar (see C in Figure 2). For emails that do not request a meeting or when the AI functionality did not detect a genuine meeting request, a user can use the "Reply" button to either reply with text or manually schedule a meeting (see D in Figure 2). Manually scheduling meetings requires inputting their time and date manually, while these are determined automatically if the AI functionality identifies a meeting request.



John Bass
Saturday December 29
,We will plan on Meeting at Al's Formal Wear at 1P...

Kate Bush
keeping the lights on
John- ,Was lovely meeting you this weekend. ,Sorry...

Daphne Co
Dinner
Hi Eric, ,Would you and Shanna like to meet us for...

Sally Beck
Per Your Request
Sally, ,Please find attached the file that we discusse...

John Bass
Wed Aug 22 2018 10:57 AM
Eric Ramiro

We will plan on Meeting at Al's Formal Wear at 1PM on that Saturday. I will see you all then.
Jason

We think we've found an event

Date: Sat Aug 25 2018
Time: 01:00 pm

Figure 2: Screenshot of the Scheduling Assistant interface mimicking the inbox part of a web interface of a popular email client - Microsoft (MS) Outlook. A) list of emails in the inbox, B) content of the selected emails, C) the AI functionality - detection and highlighting of email requests from free-text, D) reply button allowing user to either reply with text or schedule a meeting manually

Implementation

The Scheduling Assistant system has been implemented as a web application using Node.JS [1] framework with use of Express for managing the server part as well as ReactJS [2] and FabricUI [3] for handling GUI interactions and look & feel respectively. Using a WebKit [4] package the whole application can be rendered into a single web page that can be opened in a local browser. For the purpose of the subsequent Study, the application has also been augmented with instrumentation tracking user activities with emails, such as email selection and handling actions.

Email Dataset and Meeting Request Classification

In order to control the information users interact with in the subsequent Study, the Scheduling Assistant was fixed

[1] https://nodejs.org/en/ - a JavaScript based framework for implementation of back-end

[2] https://reactjs.org/ - a JavaScript library for building library

[3] https://developer.microsoft.com/en-us/fabric - the official front-end framework for building Office and Office 365 web interface.

[4] https://webkit.org/ - an open source web browser engine

to operate on a pre-determined subset of 28 email messages from the *Enron* email corpus available online. [5] As the Enron corpus contains 0.5M messages from 150 users, we selected a subset of messages using a meeting request detection ML model trained using an interactive concept learning tool for training machine learning models similar to [44]. To label each sampled email as either containing or not containing a meeting request, we referred to the following definition of a meeting request: *"Implies the sender has a meeting intent (e.g., proposing to meet at a specific time, setting up a meeting, updating an existing meeting time)"*. Following this definition we labeled a total of 85 emails with 37 of them labeled as containing meeting request. The model features included keywords related to meeting terms, availability terms, time related terms and group reference terms. The model achieved an accuracy of 93.83% on this training set. In order to select a set of emails to be used in the Study with Scheduling Assistant, we used a built-in functionality of our ML tool to select predicted positive detections, predicted negative detections and borderline detections. We further coded these messages among two independent coders and selected only the emails for which both coders agreed on the label.

Determining AI component's Accuracy

Given our **RQ2** and **RQ3**, we wanted an AI component that will perform below expectations of most end-users. AI performing at a level disappointing for most users enables testing the effectiveness of the expectation adjustment techniques. To arrive at the most appropriate accuracy for the system, we ran pretests on the internal quality-controlled crowd-sourcing platform we used which helped us determine that users come in with a nominal expectation of accuracy at a level of 75% (SD= 10%). This is in response to a question: *"How well do you feel the Scheduling Assistant works?"* with answers provided on a scale from *"0% (Never correctly detects meetings)"* to *"100% (Always correctly detects meeting)"* with 10% increments. We therefore decided to set the accuracy of the Scheduling Assistant to 50%.

Preparing the High Recall and High Precision Versions

A system at the same level of accuracy can still vary in the types of errors it makes. Given an email contains a meeting request, the Scheduling Assistant can correctly determine that it indeed contains a meeting request (*True Positive - TP*) or it can incorrectly determine that it does not contain a meeting request (*False Negative - FN*). Similarly, given an email that does not contain a meeting request, the system can correctly determine that indeed it does not (*True Negative - TN*) or incorrectly determine that the email does contain a meeting request (*False Positive - FP*). A summary of these possible classifications outcomes along with concrete examples can be found in Table 1.

Table 1: A summary of possible correct and erroneous classifications that the Scheduling Assistant's AI component can make

Type	Predicted label	Example
(TP)	Request	Let's meet 3:30pm on Friday
(TN)	No request	We appreciate all the help
(FP)	Request	Yesterday's meeting was good
(FN)	No request	How about lunch? Maybe 1:30?

In order to test our **RQ1** related to user perceptions of an AI-powered system of the same accuracy, but focused on avoidance of different types of errors, we manipulated the composition of correct classifications and the types of errors the system makes to arrive at two versions. The *High Precision* system minimizes FP types of errors. The *High Recall* system, on the other hand, minimizes FN types of errors. To achieve these versions, we manipulated the classification of 20 email messages obtained from the Enron corpus. Both *High Recall* and *High Precision* versions of the system had 5 TP classifications as well as 5 TN classifications. The *High Recall* system, however, made 8 FP detection errors and only 2 FN errors. This system achieved an accuracy[6] of 50%, a recall score[7] of 71.4% and a precision score [8] score of 38.5%. The *High Precision* system on the other hand made only 2 FP types of errors, but 8 FN ones. This system as well achieved an accuracy of 50%, but a recall score of 38.5% and a precision score of 71.4%.

5 DESIGNS FOR ADJUSTING END-USER EXPECTATIONS

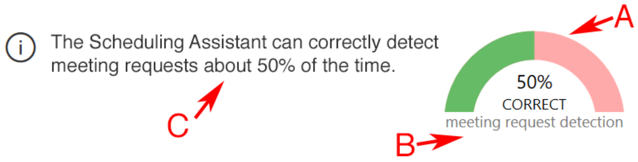
Prior work [17] has identified that expectations can be formed in three principle ways: 1) through information from external sources (e.g., being directly told about the specific properties of a system), 2) reasoning and understanding (e.g., forming expectations as an extension of understanding of how the system works), 3) first hand experience (e.g., forming an action-effect association that comes from direct interaction and experience with a system). These mechanisms inspired our three design techniques for adjusting expectations, respectively: 1) Accuracy Indicator, 2) Example-based Explanation, and 3) Control Slider.

[5] https://www.cs.cmu.edu/ ./enron/

[6] accuracy - the proportion of true results (both TP and TN) among the total number of cases examined (TP+TN+FP+FN)

[7] recall - the proportion of correctly identified positives, TP, to the total number of all positives (TP+FN)

[8] precision - the proportion of correctly identified positives TP to the total number of predicted positives (TP+FP)



The Scheduling Assistant can correctly detect meeting requests about 50% of the time.

Figure 3: The Accuracy Indicator design. A) a solid gauge chart to visually communicate accuracy, B) the accuracy expressed as number to compliment the chart, C) An associated textual description

In this section we discuss the general principles we followed in each of our techniques as well as the designs of the techniques themselves.

General Design Principles

Combining visualization and text. Prior work has compared the use of text vs. visualizations for communicating various statistical aspects of algorithms [4, 16, 43, 50]. Recent work by Fernandez et al. [16] indicated that visualization can offer better support for user decisions. Visionalizations have also been found to be generally more effective at grabbing user attention. At the same time, different users have been found to process either text or visual content with more ease. Therefore, in our designs we decided to combine limited text with simple visualization elements.

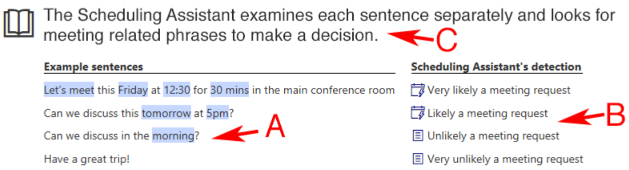
Striving for simplicity. While presence of visualization has been found beneficial, past work has also indicated that too complex visualizations (e.g., probability distributions) are too difficult for the everyday user to interpret [24, 26]. Also, too lengthy or technical textual descriptions have been shown to discourage end-users from wanting to invest time in them [45]. Given these indications, we strove for simplicity and clarity in our designs and for use of visualization elements that can be easily understood by general public.

Design Process

Our design process involved a number of iterations tested among our research team. We further performed two informal qualitative feedback session with 4 external users. These helped drive decisions regarding competing designs choices and offered early and richer insight into users' perceptions. Finally, we also performed a number of limited deployments on our internal quality-controlled crowd-sourcing platform to check the general understandability of the designs.

Designing the Accuracy Indicator

Goals. The main goal of this design is to improve user's ability to correctly estimate the percentage accuracy with which the AI-powered system performs.



The Scheduling Assistant examines each sentence separately and looks for meeting related phrases to make a decision.

Example sentences	Scheduling Assistant's detection
Let's meet this Friday at 12:30 for 30 mins in the main conference room	Very likely a meeting request
Can we discuss this tomorrow at 5pm?	Likely a meeting request
Can we discuss in the morning?	Unlikely a meeting request
Have a great trip!	Very unlikely a meeting request

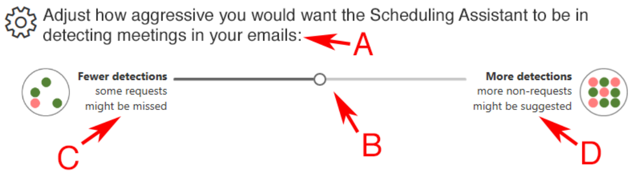
Figure 4: The Example-based Explanation design. A) examples of email sentences ordered from most unambiguous meeting request to not a meeting request, B) the decision the system made about each sentence as being or not being a meeting request, C) an associated textual description

Design. The design of our Accuracy Indicator is composed of three basic elements as shown in Figure 3. The visualization element indicated in Figure 3 A is a solid gauge chart visually depicting the 50% accuracy of the system (half-way filled with green). This visual encoding is reinforced with an explicit number expressing accuracy along with clarification that this number relates to the percentage of correct meeting request detections (see Figure 3 B). We further included a purely textual description of accuracy (see Figure 3 C).

Designing the Example-based Explanation

Goals. This design is meant to increase user understanding of how the AI component operates. Specifically, it communicates that: 1) the system examines email contents on a sentence level, 2) the presence of specific meeting related terms increases the chance the system will consider the sentence to be a meeting request. The design is also meant to implicitly communicate that the system can make mistakes. By increasing understanding of how the AI works, users can update their expectation of how well and in which situations the system is likely to work.

Design. The design is composed of two elements: a textual description (Figure 4 C) and a table with four examples (Figure 4 A and B). The textual description communicates that each sentence is examined separately and that the meeting related phrases help the system make a decision. The table shows a variety of example sentences the Scheduling Assistant may encounter and the Scheduling Assistant's decisions about whether those sentences are likely a meeting request or not. Example sentences are ordered such that the top sentence represents the most complete and unambiguous meeting request, while the bottom example is likely not a meeting request. In each example, the key phrases that are commonly associated with meeting requests, such as time, duration, location, date, and an invitation phrase are highlighted in blue. The "Scheduling Assistant's decision" column shows the decision along with the system's confidence in its



Adjust how aggressive you would want the Scheduling Assistant to be in detecting meetings in your emails:

Figure 5: The Control Slider design. A) text information about the functionality offered by the slider, B) an interactive control slider that allows the user to control False Positive vs False Negative rate, C) highest precision extreme of the slider setting, D) highest recall extreme setting

decision indicating that the system relies on some probabilistic reasoning, e.g., *"Very likely meeting request* or *"Unlikely a meeting request"*. In the example shown in Figure 4, the system arguably makes a mistake classifying the third sentence: *"Can we discuss in the morning?"* as not representing a meeting, likely because meeting request related keywords are sparse or lacking.

Designing the Control Slider

Goals. This design aims at two goals. First, it allows the user to experiment with controlling the rate of False Positive and False Negative mistakes. Second, it allows the user to set the system's decision threshold. Prior work has shown that letting users contribute to a system's behavior may make them more accepting of the system's mistakes [47].

Design. The Control Slider design is composed of two main elements: a textual description (see Figure 5 A) and a user controllable interactive slider (see Figure 5 B,C,D). The textual description offers instruction to the use of the slider, letting the user know that the slider controls the system's threshold for detecting meeting requests. By controlling the slider users can adjust the system's behavior to balance two extremes. The extreme left, sets the system to work in a High Precision mode, this is communicated to the user as *"Fewer detection - some requests might be missed"*. Setting the slider to the extreme right sets the system in a High Recall mode, this is labeled for the user as *"More detections - more non-requests might be suggested"*. Both ends of the slider also feature images meant to visually reinforce the nature of high and low precision. In these images, dots indicate detections with green dots representing correct detections while red dots represent errors.

6 STUDY 1 - IMPACT ON EXPECTATIONS

The purpose of our first Study was to verify if our proposed designs impact user expectations as intended, specifically as outlined by our three hypotheses for **RQ2**. We designed the Study as an incomplete factorial setup with 6 separate conditions. Two pure conditions were Accuracy Indicator and example based Explanation design elements by themselves. The other four conditions were a combination of these with addition of the Control Slider element. The Control Slider element was not present by itself as a separate condition as it was deemed not useful without a feedback mechanism provided by the Accuracy Indicator or Explanation elements. We deployed the Study on an internal crowd-sourcing platform similar to Mechanical Turk. A total number of 150 participants from US only aged 18+ was recruited (25 per condition). We used a standard recruitment supported by our crowdsourcing platform posting short information about the task, the expected time required, and the payment. Each participant was allowed to complete the Study only once. The Study has been approved by internal IRB and took on average 5:21 min (*SD* : 3.45 min). Participants were compensated $1.35 per task.

Procedure

The Study procedure involved participants completing 6 steps, each on a separate page. First they were shown information about the Study along with consent. After expressing their consent, on the second page they were shown a short description of the Scheduling Assistant. This introduction informed them that the system operates by automatically analyzing text in emails to detect meeting requests. Once such a request is detected it is highlighted. This was accompanied by a screenshot showing an example email with a meeting request sentence highlighted similar to Figure 2. The next page involved a short interactive demo of the Scheduling Assistant in which participants were required to appropriately respond to 4 example emails before moving forward. While short, we believe such an introduction represents a realistic amount of information users may be willing to attend to in a real-life setting (users are notorious for skipping long tutorials).

After the demo, participants filled in a survey indicating their initial expectations of the Scheduling Assistant's accuracy and answering questions about their tech-savviness, as well as their familiarity and frequency of use of AI-powered systems as detailed in the Measures section. On the next page they viewed 1 one the 6 conditions they were randomly assigned to. There was no minimal time requirement for viewing the condition. After experiencing their assigned condition, participants were asked an attention check question: *"What did you see on the previous page? (Check all that apply)"*. The 4 multiple answers described each possible expectation adjustment design elements. On the final page, participants were again asked questions about their accuracy expectations, as well as their understanding and feeling of control as detailed in the Measures section.

Measures

Perceptions of system accuracy were measured through two questions adopted from the Expectations Confirmation Model [39]: pre-intervention *“How well do you expect the Scheduling Assistant to work”* and post-intervention *“How well do you feel the Scheduling Assistant works”*. Both were answerable on an 11-point scale from *"0% (Never correctly detects meetings)"* to *"100% (Always correctly detects meeting)"* with 10% increments. Understanding of the AI component was measured through two subjective report questions adapted from [42]. One question asked about understanding how the system makes positive detections: *“I feel like I have a good understanding of how the Scheduling Assistant decides whether an email contains a meeting request”*, while the other asked about understanding what kind of mistakes the system can make: *“I feel like I understand what kind of mistakes the Scheduling Assistant is likely to make”*. Hence the questions aimed at covering detections of true positives and true negatives (the two items were moderately correlated $r_s = .55$). Subjective perception of control over the system's behavior was measured by a question adapted from [47]: *“I feel like I have some control over the Scheduling Assistant's behavior”*. Answers were given on a 7-point Likert scale from *"Strongly disagree"* to *"Strongly agree"*. Additionally we asked questions to determine tech-savviness, familiarity and frequency of use of AI and prior experience with the particular AI functionality offered by the Scheduling Assistant. These were used as an additional level of control for the analysis.

Analysis

We removed 34 (23%) of participants that failed the attention check, which resulted in a final participant count of 116. To test each hypothesis we used an independent t-test to compare individual conditions separated into two groups, with and without a particular design element (e.g., with and without Accuracy Indicator). We additionally checked pre-intervention balance in participant reported measures.

Results

Testing H2.1. Comparing conditions with ($N = 70$) and without ($N = 46$), the Accuracy Indicator revealed a significant difference ($p < 0.01$) in expectations of accuracy (see **H2.1** in Figure 6), with lower expectations for participants who saw the Accuracy Indicator ($M = 6.77$, $SD = 1.912$) than for those who did not ($M = 7.92$, $SD = 1.978$). A check on pre-intervention expectations of accuracy revealed a balanced sample ($M = 7.66$, $SD = 1.832$) vs. ($M = 7.38$, $SD = 1.947$). We therefore consider **H2.1 supported** as the Accuracy Indicator brought participants' expectations of accuracy significantly closer to the system's true accuracy of 50%.

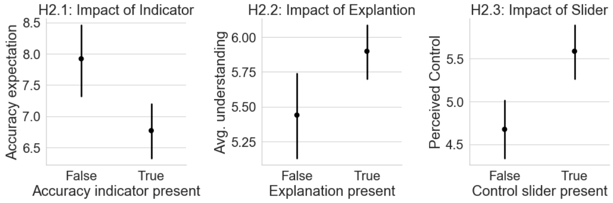


Figure 6: Impact of different design techniques on key aspects of expectations. From the left: 1) impact of Accuracy Indicator on expectations of accuracy, 2) impact of Explanation on average understanding, 3) impact of Control Slider on feeling of control

Testing H2.2. Comparing conditions with ($N = 85$) and without ($N = 35$) Explanation revealed a significant ($p < 0.05$) positive impact (see H2.2 in Figure 6), with higher average level of perceived understanding (mean of both understanding questions) for participants who saw the Explanation ($M = 5.90$, $SD = 0.99$) than for those who did not ($M = 5.44$, $SD = 1.06$). A check on pre-intervention differences in tech-savviness revealed a balanced sample ($M = 6.33$, $SD = 0.95$) vs. ($M = 6.32$, $SD = 1.10$). Similarly no significant imbalance was present in prior frequency of use of AI systems ($M = 3.80$, $SD = 1.62$) vs. ($M = 3.52$, $SD = 1.66$). We therefore consider **H2.2 supported** as Explanation intervention significantly increased participants' perceived level of understanding of how the Scheduling Assistant system works.

Testing H2.3. Comparing conditions with ($N = 58$) and without ($N = 62$) Control Slider revealed a significant ($p < 0.001$) positive impact (see H2.3 in Figure 6), with higher feeling of control for participants that saw the Control Slider ($M = 5.62$, $SD = 1.40$) than for those that did not ($M = 4.67$, $SD = 1.54$). We also note that 87% of the participants moved the slider from its default position. A check on pre-intervention differences in tech-savviness revealed balanced sample ($M = 6.31, SD = 0.93$) vs. ($M = 6.38, SD = 1.03$). Similarly no significant imbalance was present in prior frequency of use of AI systems ($M = 3.59$, $SD = 1.59$) vs. ($M = 3.84$, $SD = 1.67$). We therefore consider **H2.3 supported** as the Control Slider intervention significantly increased participants' perceived level of control over the Scheduling Assistant's behavior.

Additionally we found a significant ($p = 0.048$) negative effect of Accuracy Indicator on feeling of control: ($M = 4.96$, $SD = 1.54$) for conditions with Indicator and ($M = 5.49$, $SD = 1.47$) for conditions without it.

Summary of results

The results from this Study indicate that our expectation adjustment designs significantly affected the desired aspects of expectations in the hypothesized directions. All three hypotheses have been supported.

7 STUDY 2 - TASK-BASED EVALUATION

The purpose of the second Study was two-fold: (1) explore possible differences in perception of accuracy and the acceptance of two versions (High Recall and High Precision) of the Scheduling Assistant system with the same nominal 50% level of accuracy - **H1.1**; (2) evaluate the effectiveness of the expectation adjustment techniques in increasing user acceptance and satisfaction with the system after using the Scheduling Assistant for completing an actual task - **H3.1**.

The Study was designed as a full factorial setup with 16 separate conditions. The same 8 conditions were tested in High Recall and High Precision versions of the system. For each of these versions we tested a baseline condition (going directly to the task without being exposed to any expectation adjustment technique). The 3 pure conditions involved the Accuracy Indicator, the examples-based Explanation, and the Control Slider design elements by themselves (we added Control Slider by itself as a separate condition, a change compared to Study 1). An additional 4 conditions were combinations of these base design elements (e.g., Accuracy Indicator and Control Slider together).

The Study was approved by internal IRB and deployed on our internal crowd-sourcing platform. A total number of 400 participants (25 per condition) were recruited. Recruitment procedures were the same as in Study 1. Each participant could complete the Study only once and participants from Study 1 were not reused. Each task took on average 10:35 min (SD : 6.22 min). Participants were compensated $2.45 per task.

Procedure

The steps were similar to Study 1 with a few modifications. First, in addition to asking questions about expected accuracy, we asked questions about initial acceptance as detailed in the Measures section. Second, questions about understanding and control were dropped and no questions were asked directly after the interventions (except for one attention check question). The questions matching those before intervention were asked after participants completed the task with the system. Third, after seeing the intervention (or without seeing one as in the baseline condition), participants were directed to complete a task involving correct handling of 20 email messages with support from the Scheduling Assistant performing according to the AI system accuracy rate the participant was randomly assigned to (i.e., High Recall or High Precision as detailed in Section 4).

Measures

Questions about accuracy expectations as well as tech-savviness, familiarity and frequency of use of AI and prior experience with the particular AI functionality offered by the Scheduling Assistant were the same as in Study 1. Post task questions evaluating satisfaction were adapted from [6]: *"I am satisfied with how well the Scheduling Assistant worked"*. We also included 5 questions related to various aspects of acceptance adapted from TAM [48] and from [39]. Specifically we asked about future use: *"I would use the Scheduling Assistant if it was available"*, recommendation to others: *"I would recommend the Scheduling Assistant to my friends and colleagues"*, helpfulness: *"I found the Scheduling Assistant to be helpful"*, productivity: *"I found the Scheduling Assistant to be able to improve my productivity"*, and annoyance: *"I found the Scheduling Assistant to be annoying or distracting"*.

Analysis

For this analysis we removed 75 participants (19%) that failed the attention checks, which resulted in a final participant count of 325. As this could have resulted in imbalance of initial measures, we check for it and in case such imbalance has been detected we perform and report additional analyses using change in measure rather than post measure directly. Also as acceptance items showed high correlation (Cronbach's Alpha: 0.81), meaning they can be treated as aspects of the same constructs, unless otherwise stated, we report acceptance as a combined measure formed by averaging these 5 items (with inversion of annoyance scale).

Participant Characteristics

Participants reported frequently interacting with AI based applications (47% multiple times a day and only 18% once a month of less frequently) and having relatively high general perception of their performance ($M = 71.5\%$, $SD = 16.6$). Their self reported tech-savviness was also high ($M = 6.37$, $SD = 0.90$), however, familiarity with particular AI functionality offered by the Scheduling Assistant (i.e., automatic highlighting in text) was moderate ($M = 4.32$, $SD = 1.73$).

Results

Testing H1.1. The first part of this hypothesis relates to the impact of AI versions on perceptions of accuracy. Comparison of the AI component versions focused on High Recall (low False Negative rate) ($N = 158$) and on High Precision (low False Positive rate) ($N = 167$)) revealed a significant impact on accuracy perceptions ($p < 0.001$) with High Recall version leading to higher post-use perceptions of system accuracy ($M = 7.09$, $SD = 1.92$), than the High Precision version ($M = 5.75$, $SD = 2.41$) as can also be seen in Figure 7. As the evaluation was made using percentage scale with

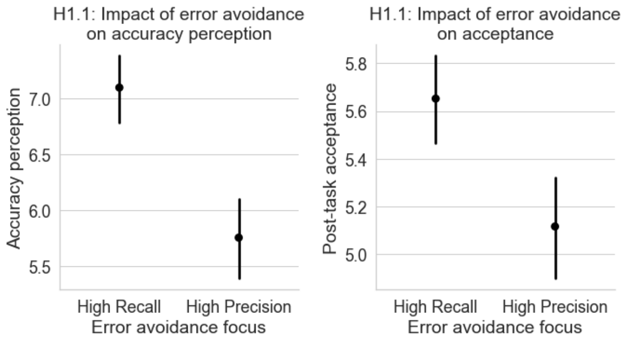


Figure 7: Impact of different focus of AI component on error avoidance on accuracy perceptions. High Recall - low False Negatives rate, High Precision - low False Positives rate

10% increments, this represents a mean difference of 13.4% in terms of how the system's accuracy in correctly detecting meeting requests in emails is perceived by the user. The underlying system in both cases was operating at 50% accuracy. A check on pre-exposure balance of accuracy expectation revealed no significant difference: High Recall ($M = 7.61$, $SD = 1.85$), High Precision ($M = 7.50$, $SD = 1.88$).

The second part of this hypothesis relates to the impact on acceptance. This analysis also revealed a significantly ($p < 0.001$) higher post-use acceptance of the High Recall version of the systems ($M = 5.65$, $SD = 1.21$) as compared to the High Precision version ($M = 5.12$, $SD = 1.37$) as can also be seen in Figure 7. The mean difference is half a point on a seven point Likert scale. A check on pre-exposure balance in acceptance again showed no significant differences: High Recall ($M = 5.81$, $SD = 0.97$) vs. High Precision ($M = 5.68$, $SD = 1.16$).

Given the significant result in the opposite direction to the one initially hypothesized, we consider **H1.1 rejected** with High Recall version of the system, at least in the context of the Scheduling Assistant and the task given, resulting in higher subjective perceptions of accuracy and triggering higher level of acceptance.

Testing H3.1. Initial explorations of the data revealed significant differences in pre-study acceptance for Mixed-techniques (e.g., Accuracy Indicator and Explanation together) as compared to Baseline ($t_{151} = -1.829, p = 0.07$) and Pure-techniques ($t_{279} = 2.179, p < 0.05$) (e.g., Accuracy Indicator by itself). We were therefore not able to draw meaningful conclusions on the impact of these Mixed-techniques and decided to exclude them from further analysis.

As we hypothesized mainly about impact of Pure-techniques, we continue with the analysis focused around these. Comparison of post-use acceptance scores for Baseline ($N = 44$) and Pure-techniques ($N = 119$) revealed a non-significant difference: Baseline ($M = 5.25$, $SD = 0.907$) vs. Pure-techniques ($M = 5.40$, $SD = 1.108$). As we have earlier observed significant differences in perceptions of High Recall and High Precision versions of the system, we further examined the impact of techniques in both versions separately.

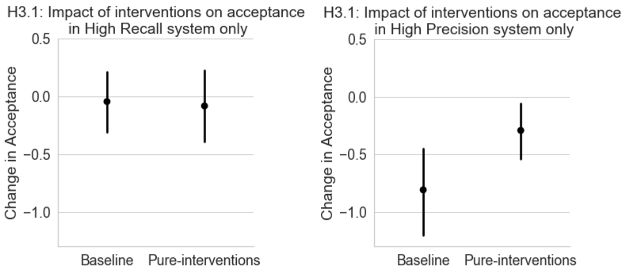


Figure 8: Comparison of impact of Baseline vs Pure-techniques conditions for High Recall (left) and High Precision (right) systems separately

Looking at the High Recall version of the system revealed a non significant difference in post-use acceptance ($p = 0.16$) between the Baseline ($N = 21$, $M = 6.02$, $SD = 1.52$) and Pure-techniques ($N = 61$, $M = 5.57$, $SD = 1.28$). As the check on pre-use acceptance rating revealed a slight initial difference between Baseline ($M = 6.07$, $SD = 1.43$) and the Pure-techniques ($M = 5.66$, $SD = 1.12$), albeit still not significant ($p = 0.13$), we examined change in acceptance, which revealed a difference far from significance ($p = 0.90$) as shown on the left in Figure 8.

Looking, on the other hand, only at the High Precision version of the system revealed a significant ($p < 0.05$) difference in post-use acceptance between the Baseline ($N = 23$, $M = 4.56$,$SD = 1.59$) and Pure-techniques ($N = 58$, $M = 5.23$, $SD = 1.15$). A check on pre-use acceptance revealed no significant differences - Baseline: ($M = 5.38$, $SD = 1.21$), Pure-techniques: ($M = 5.52$, $SD = 1.10$). For consistency with the analysis for High Recall condition, we also checked the delta change in acceptance, which was also significant ($p < 0.05$) as shown on the right in Figure 8.

Similar results have been found with respect to satisfaction. Participants in the High Recall version of the system reported no significant difference in satisfaction between the Baseline ($M = 5.85$, $SD = 1.24$) and Pure-techniques ($M = 5.52$, $SD = 1.58$). At the same time, in the High Precision version, participants exposed to Pure-techniques reported significantly ($t_{163} = -1.93, p = 0.05$) higher satisfaction ($M = 4.43$, $SD = 1.70$) than those in the Baseline ($M = 3.86$, $SD = 1.91$).

Given that our preparation techniques were effective only when the system was perceived as significantly below user expectations - High Precision, but not when the system was perceived as being on pair with user expectations - High Recall, we consider **H3.1 partially supported**.

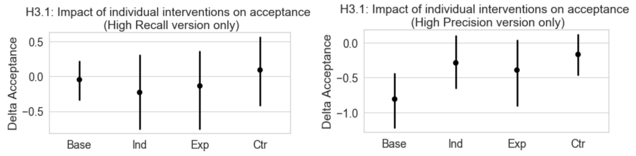


Figure 9: Comparison of impact of baseline and individual techniques on change in acceptance for both High Recall (left) and High Precision (right) versions of the Scheduling Assistant. Base - denotes baseline condition, Ind - Accuracy Indicator, Exp - Explanation, and Ctr - Control Slider

Additional analysis. We performed additional tests to compare completion time and accuracy on the task for Baseline and Pure-techniques. We found no significant differences for task accuracy ($M = 0.91$, $SD = 1.00$) for Baseline and ($M = 0.88$, $SD = 0.95$) for the Pure-techniques; task completion time for Baseline ($M = 5.38$, $SD = 3.22$) and Pure-techniques ($M = 4.85$, $SD = 2.52$). this indicates that differences in expectations had no strong impact on objective user performance, which is consistent with prior work [5].

Although we did not hypothesize specific differences between individual techniques, we looked into these as well. Figure 9 shows the impact on change in acceptance of Baseline and individual techniques separately for High Recall (left) and High Precision (right) version of the system. None of the techniques in the High Recall version had a significant impact on acceptance as compared to Baseline. In the High Precision version, compared to change in acceptance in the Baseline ($N = 23$, $M = -0.809$, $SD = 0.973$), we found a significant ($t_{39} = -2.54, p < 0.05$) positive impact of Control Slider ($N = 18$, $M = -0.167$, $SD = 0.637$) and a weakly significant ($t_{39} = -1.80, p < 0.1$) positive impact of Accuracy Indicator ($N = 16$, $M = -.287$, $SD = 0.826$). The impact of Explanation was not significant compared to the Baseline. Comparisons between the techniques did not reveal any significant differences either.

Summary of results

Results from this Study indicate that the High Recall version of the Scheduling Assistant results in significantly higher perceptions of accuracy and significantly higher acceptance as compared to the High Precision version. This is surprising as we had originally hypothesized an opposite direction of change. Furthermore the expectation adjustment techniques have been shown effective in significantly increasing user satisfaction and acceptance, however only in the High Precision version of the system. This is the version of the system that had been perceived as performing worse by our users and our techniques appear to mitigate this effect.

8 LIMITATIONS

Unfortunately due to the imbalance in initial user acceptance after the necessary data cleaning we were unable to investigate the impact of mixed-techniques. Furthermore, the Scheduling Assistant represents a type of system that is passive/assistive and casual, where the impact of AI imperfections is arguably less critical than in e.g., critical support medical systems. Our findings, should therefore, be considered with this in mind.

9 DISCUSSION

Hypotheses **H2.1, H2.2, and H2.3** have been supported in our results showing that our expectation adjustment techniques successfully impacted the intended aspects of expectations. Furthermore, **H3.1** has also been partially supported showing that the techniques are successful in increasing user satisfaction and acceptance with an imperfect AI system - Scheduling Assistant. Finally, the rejection of **H1.1** shows, contrary to expectations from anecdotal reports from practitioners, that user satisfaction and acceptance of a system optimized for High Recall (i.e., system that make more False Positive mistakes) can be significantly higher than for a system optimized for High Precision. We hypothesize that since users can easily recover from a False Positive in our interface (highlighting can be ignored) than from a False Negative (no highlighting and therefore more careful reading as well as manual scheduling of the meeting required) that the optimal balance of precision and recall is likely in part a function of the cost of recovery from each of these error types. This also informs the call by [13] to understand how precision-recall impacts UX in systems with machine learning.

Our work provides insights into feasible preparation techniques for end-users interacting with imperfect AI-powered systems. This is especially valuable as our techniques are very simple and light-weight to process. In the Study users spent just 15.45 seconds (SD=34.3) on average looking at the interventions, compared to the time of 5.54 minutes (SD=6.4) spent on the task and overall Study time of 10.5 minutes (SD=7.7). Despite such short exposure, the techniques managed to significantly improve user satisfaction and acceptance.

We believe our techniques can offer a substantial contribution. We show that user satisfaction and acceptance can be improved not only through deception as used in marketing [8] or in-depth involved understanding shaping user mental models as used in intelligible AI works [51], but also through fairly simple expectation adjustment techniques. This addresses an important gap in existing research on preparing end-users for imperfect AI-powered systems. Being light-weight, they also address the issue of end-users not willing to engage in complex understanding of underlying algorithms to be able to accept an AI-powered system [51].

Different Impact of Interventions in High Recall and High Precision System Versions

We found that our expectation adjustment techniques did not offer significant improvement in acceptance or satisfaction in the High Recall version of the system, in which user satisfaction and acceptance dropped only slightly in the Baseline. However, they proved effective in the High Precision version, in which users experienced much higher "disappointment". We believe this shows that expectation adjustment works as intended. We designed the Study to expose users to the imperfections of AI in order to check if preparation through expectation adjustment can be an effective approach. The fact that one version of our system was perceived as on par with user expectations, rendered the expectation adjustment unnecessary to mitigate any negative effects, (i.e., preparing users for AI imperfections when the user expectations are met in actual use will not result in any difference). Another explanation could be that users randomly assigned to interact with the High Recall version of the system simply came with lower expectations and acceptance in the first place. We, however, verified this was not the case as reported in section 7.

Differences in Perception of the Two System Versions

Current practitioner belief assumes that focus on High Precision, hence avoidance of False Positives is the most appropriate choice for AI-powered systems while [13] points out the impact on UX is unstudied. Practitioner rationale is that if mistakes are hidden from the user, the imperfections will be less distracting and annoying. At the same time, missed opportunities to support the user are less visible. This can certainly be the case in e.g., a movie recommender, where failing to recommend a good movie (False Negative) may remain unnoticed, while recommending a movie a user does not like (False Positive) will make system imperfections very visible. In our Study, we show, however, that the task in which the AI functionality is supposed to assist the user should be carefully analyzed as well, especially in relation to the impact of different types of AI mistakes. In particular a number of aspects should be considered, such as *workload*, both *mental* (e.g., ignoring incorrect suggestions, scanning multiple suggestions) and *physical* (e.g., having to execute a task manually or reverse an incorrect system action), as well as *criticality of consequences* of following an incorrect AI suggestion (e.g., not scheduling meeting, making an unsolicited purchase). In the case of the Scheduling Assistant, the task of meeting scheduling, without system assistance requires the user to perform more manual and mental work (hence a High Precision system might be perceived by the user as not offering much support). At the same time, if the system highlights a sentence incorrectly (High Recall focus), the effort for the user to recover from such mistake is fairly minimal as the user can just examine the sentence and ignore it (the cost is mostly additional perceptual load, but not manual effort). Such systems may be perceived as offering more assistance, even if they make more mistakes.

Generalizability of Findings

Scheduling Assistant represents a type of system that is passive/assistive and casual (similar to "passive context-awareness" from [9]). While this is only one class of AI-powered systems, we believe this class represents many current efforts of integrating AI into end-user applications (e.g., chat/email response suggestion, sharing past memory in social network, assistance in content analysis [10], etc.). Furthermore we believe our findings should generalize to other systems and tasks on a number of levels: 1) Our expectation adjustment techniques are task agnostic as they are informed by the high-level theory of attitudes and the general mechanisms in which people learn new information. 2) We provide concrete empirical evidence for the need of careful analysis of costs associated with different types of AI errors. This finding generalizes to a number of systems and tasks. 3) The specific suggestion to optimize for High Recall, hence avoid False Negatives, is much more specific. The importance of avoiding different types of errors depends on domain and system design. This is indeed a complex issue and we are not suggesting that avoiding False Negatives should always be preferred. Having said so we believe that this finding generalizes to a class of passive systems (i.e., user makes final decision) in which the relative ratio of workload for False Positives and False Negatives is low (e.g., meeting scheduling highlight, email/chat reply suggestion, autocorrect, autocomplete in search). In high-cost critical systems, it is more important to analyze severity of consequences of different errors rather than workload (e.g., cancer screening, suspicious behavior detection).

10 CONCLUSION

In this work, we designed three expectation adjustment techniques and experimentally showed their effectiveness in improving user satisfaction and acceptance of an imperfect AI-powered system, an email Scheduling Assistant. We also showed that focus on High Precision rather than High Recall of a system performing at the same level of accuracy can lead to much lower perceptions of accuracy and decreased acceptance. Our findings open the way to shaping expectations as an effective way of improving user acceptance of AI technologies.

11 ACKNOWLEDGMENTS

We thank Bongshin Lee for valuable feedback on visualization ideas as well as Jina Suh, Adam Fourney and Paul Koch for help with using the crowd-sourcing platform. We would also like to thank all our study participants and all the reviewers for helping to improve this work.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 582.

[2] Dan Amir and Ofra Amir. 2018. HIGHLIGHTS: Summarizing Agent Behavior to People. In *Proc. of the 17th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[3] Rolph E Anderson. 1973. Consumer dissatisfaction: The effect of disconfirmed expectancy on perceived product performance. *Journal of marketing research* (1973), 38–44.

[4] Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. ACM, 9–14.

[5] T Bentley. 2000. Biasing Web Site User Evaluations: a study. In *Proceedings of the Conference of the Computer Human Interaction Special Interest Group of the Ergonomics Society of Australia (OzCHI2000)*. 130–134.

[6] Anol Bhattacherjee. 2001. Understanding information systems continuance: an expectation-confirmation model. *MIS quarterly* (2001), 351–370.

[7] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important?: a study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM, 169–178.

[8] Richard N Cardozo. 1965. An experimental study of customer effort, expectation, and satisfaction. *Journal of marketing research* (1965), 244–249.

[9] Guanling Chen, David Kotz, et al. 2000. *A survey of context-aware mobile computing research*. Technical Report. Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College.

[10] NAN-CHEN CHEN, MARGARET DROUHARD, RAFAL KOCIELNIK, JINA SUH, and CECILIA R ARAGON. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting The Focus to Ambiguity. (2018).

[11] Gilbert A Churchill Jr and Carol Surprenant. 1982. An investigation into the determinants of customer satisfaction. *Journal of marketing research* (1982), 491–504.

[12] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2382–2393.

[13] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 278–288.

[14] Margaret Drouhard, Nan-Chen Chen, Jina Suh, Rafal Kocielnik, Vanessa Pena-Araya, Keting Cen, Xiangyi Zheng, and Cecilia R Aragon. 2017. Aeonium: Visual analytics to support collaborative qualitative coding. In *Pacific Visualization Symposium (PacificVis), 2017 IEEE*. IEEE, 220–229.

[15] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

[16] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 144.

[17] Martin Fishbein and Icek Ajzen. 1975. *Belief, attitude, intention and behavior: An introduction to theory and research.*

[18] Francisco J Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explain Yourself: A Natural Language Interface for Scrutable Autonomous Robots. *arXiv preprint arXiv:1803.02088* (2018).

[19] Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.

[20] Jan Hartmann, Antonella De Angeli, and Alistair Sutcliffe. 2008. Framing the user experience: information biases on website quality judgement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 855–864.

[21] Bradley Hayes and Julie A Shah. 2017. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. ACM, 303–312.

[22] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[23] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. 2017. Enabling robots to communicate their objectives. *Autonomous Robots* (2017), 1–18.

[24] Harald Ibrekk and M Granger Morgan. 1987. Graphical communication of uncertain quantities to nontechnical people. *Risk analysis* 7, 4 (1987), 519–529.

[25] Julia Jarkiewicz, Rafał Kocielnik, and Krzysztof Marasek. 2009. Anthropometric Facial Emotion Recognition. In *International Conference on Human-Computer Interaction*. Springer, 188–197.

[26] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5092–5103.

[27] Rafal Kocielnik, Os Keyes, Jonathan T Morgan, Dario Taraborelli, David W McDonald, and Gary Hsieh. 2018. Reciprocity and Donation: How Article Topic, Quality and Dwell Time Predict Banner Donation on Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 91.

[28] Rafal Kocielnik, Fabrizio Maria Maggi, and Natalia Sidorova. 2013. Enabling self-reflection with LifelogExplorer: Generating simple views from complex data. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*. IEEE, 184–191.

[29] Rafal Kocielnik, Mykola Pechenizkiy, and Natalia Sidorova. 2012. Stress analytics in education. In *Educational Data Mining 2012*.

[30] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 70.

[31] Sari Kujala, Ruth Mugge, and Talya Miron-Shatz. 2017. The role of expectations in service evaluation: A longitudinal study of a proximity mobile payment service. *International Journal of Human-Computer Studies* 98 (2017), 51–61.

[32] Irwin P Levin and Gary J Gaeth. 1988. How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of consumer research* 15, 3 (1988), 374–378.

[33] Brian Y Lim and Anind K Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*. ACM, 195–204.

[34] Brian Y Lim and Anind K Dey. 2010. Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 13–22.

[35] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2119–2128.

[36] Jaroslav Michalco, Jakob Grue Simonsen, and Kasper Hornbæk. 2015. An exploration of the relation between expectations and user experience. *International Journal of Human-Computer Interaction* 31, 9 (2015), 603–617.

[37] Richard L Oliver. 1977. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of applied psychology* 62, 4 (1977), 480.

[38] Richard W Olshavsky and John A Miller. 1972. Consumer expectations, product performance, and perceived product quality. *Journal of marketing research* (1972), 19–21.

[39] Eeva Raita and Antti Oulasvirta. 2011. Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with Computers* 23, 4 (2011), 363–371.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[41] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 101–108.

[42] Andrew B Rosenkrantz and Eric R Flagg. 2015. Survey-based assessment of patientsâĂŹ understanding of their own imaging examinations. *Journal of the American College of Radiology* 12, 6 (2015), 549–555.

[43] Enrico Rukzio, John Hamard, Chie Noda, and Alexander De Luca. 2006. Visualization of uncertainty in context aware mobile applications. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*. ACM, 247–250.

[44] Patrice Simard, David Chickering, Aparna Lakshmiratan, Denis Charles, Léon Bottou, Carlos Garcia Jurado Suarez, David Grangier, Saleema Amershi, Johan Verwey, and Jina Suh. 2014. ICE: enabling non-experts to build models interactively for large-scale lopsided problems. *arXiv preprint arXiv:1409.4814* (2014).

[45] S Stumpf, A Bussone, and D OâĂŹSullivan. 2016. Explanations considered harmful? user interactions with machine learning systems. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

[46] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies* 67, 8 (2009), 639–662.

[47] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The Illusion of Control: Placebo Effects of Control Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 16.

[48] Viswanath Venkatesh and Fred D Davis. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science* 46, 2 (2000), 186–204.

[49] Viswanath Venkatesh and Sandeep Goyal. 2010. Expectation disconfirmation and technology adoption: polynomial modeling and response surface analysis. *MIS quarterly* (2010), 281–303.

[50] Jhim Kiel M Verame, Enrico Costanza, and Sarvapali D Ramchurn. 2016. The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. In *Proceedings of the 2016 chi conference on human factors in computing systems*. ACM, 4908–4920.

[51] Daniel S Weld and Gagan Bansal. 2018. Intelligible Artificial Intelligence. *arXiv preprint arXiv:1803.04263* (2018).