Assessing Human-AI Interaction Early through Factorial Surveys: A Study on the Guidelines for Human-AI Interaction

TIANYI LI, Purdue University
MIHAELA VORVOREANU, DEREK DEBELLIS, and SALEEMA AMERSHI, Microsoft

This work contributes a research protocol for evaluating human-AI interaction in the context of specific AI products. The research protocol enables UX and HCI researchers to assess different human-AI interaction solutions and validate design decisions before investing in engineering. We present a detailed account of the research protocol and demonstrate its use by employing it to study an existing set of human-AI interaction guidelines. We used factorial surveys with a 2 × 2 mixed design to compare user perceptions when a guideline is applied versus violated, under conditions of optimal versus sub-optimal AI performance. The results provided both qualitative and quantitative insights into the UX impact of each guideline. These insights can support creators of user-facing AI systems in their nuanced prioritization and application of the guidelines.

CCS Concepts: • **Human-centered computing** → *HCI design and evaluation methods;*

Additional Key Words and Phrases: Human-AI interaction, AI, UX, design guidelines, evaluation, factorial survey

ACM Reference format:
Tianyi Li, Mihaela Vorvoreanu, Derek DeBellis, and Saleema Amershi. 2023. Assessing Human-AI Interaction Early through Factorial Surveys: A Study on the Guidelines for Human-AI Interaction. *ACM Trans. Comput.-Hum. Interact.* 30, 5, Article 69 (September 2023), 45 pages.
https://doi.org/10.1145/3511605

1 INTRODUCTION

The field of **human-computer interaction** (**HCI**) has a long tradition of creating and using design principles or heuristics intended to improve usability and **user experience** (**UX**). Some design principles are intended to be universally applicable (e.g., [54, 80, 81, 103, 104, 106]). Others are delimited to specific contexts such as games (e.g., [32, 63, 92, 93]) or mobile interfaces (e.g., [18, 69, 98]). Yet another type of principle is targeted towards interaction with specific user groups such as older adults (e.g., [89]) or vision-impaired users (e.g., [67]), to name a few. More recently, guidance has been proposed for **artificial intelligence** (**AI**) design, mostly in white papers by large technology companies such as Google [37], IBM [49], and SAP [99] for all AI systems; and Microsoft [70], Amazon [4], and Facebook [34] for conversational AI. In research scholarship, comprehensive AI

Authors' addresses: T. Li, Purdue University, 610 Purdue Mall, West Lafayette, IN 47907; email: li4251@purdue.edu; M. Vorvoreanu, D. DeBellis, and S. Amershi, Microsoft, One Microsoft Way Redmond, WA 98052–6399; emails: mihaela.vorvoreanu@microsoft.com, samershi@microsoft.com; D. DeBellis, 7595 Technology Way, Suite 400 Denver, CO 80237; email: derek.debellis@gmail.com.

1073-0516/2023/09-ART69 $15.00
https://doi.org/10.1145/3511605

design guidance is not as abundant. In 2019, Amershi et al. [5] synthesized more than 20 years of scholarly and industry research in interaction with AI and mixed-initiative systems into a set of 18 guidelines for human-AI interaction (HAI guidelines).

While evidence exists about the effects on UX of older sets of principles such as Nielsen's 10 heuristics [79], this research contributes such evidence about AI design guidance. Focusing on the HAI guidelines articulated in [5], the goal of this study is to characterize these guidelines' impacts on product preference and UX. To clarify, evidence about each of the HAI guidelines' effectiveness exists in previous literature. For example, Guideline 11 is about explanations, and an extensive body of work documents how and when explanations are effective [1, 39, 74]. The question at hand is not whether the guidelines are effective; it is to understand *how* each guideline impacts product preference, user perceptions, and UX:

RQ1 What are participants' perceptions and reasons for preferring a product that applies or a product that violates each HAI guideline?

RQ2 What is the effect of applying vs. violating each of the guidelines on multiple UX metrics, under conditions of optimal and sub-optimal AI performance?

The first research question aims at identifying how users compare products that apply or violate a guideline but are otherwise identical. It is designed to capture user perceptions about both products and reasons for preferring one of them. With the second research question, we capture specific impacts a guideline might have on UX metrics such as feelings, trust, and perceived product quality. We also examine whether this impact has an interaction effect with AI performance. For example, would applying the guidelines mitigate some of the negative perceptions of a product that does not perform optimally? To address the research questions, we conducted 18 separate studies, one for each of the 18 guidelines by Amershi et al. [5]. We used factorial surveys with a 2×2 mixed design that manipulated guideline compliance (with two levels, application, and violation), and AI performance (also with two levels, optimal and sub-optimal). We represented these variations in vignettes, using fictional products, to avoid choices influenced by brand preference.

A total of 1,300 participants from a crowdsourcing platform participated in the 18 factorial surveys. Two surveys failed because we did not manipulate the independent variable in noticeable ways. The 16 factorial surveys that passed the manipulation checks revealed nuanced user perceptions related to guideline compliance and showed demonstrable impacts on UX. We report the qualitative results of open-ended questions and the quantitative results about each guideline's impact on each UX metric. Since interaction effects between the two factorial variables (guideline compliance and AI performance) were not detected in most studies, we only include results about the interaction effects when they are statistically significant. Thematic analysis of the qualitative results revealed participants' reasons for their choice of preferred products, leading to a number of nuanced insights and design implications about the guidelines' application. Some recurring reasons for preferring guideline application include having more control, product reliability, feeling of productivity, to name a few. Occasionally, participants chose the product that violated the guideline. The most common rationales were concerns about privacy and trust, or sometimes personal preferences. In addition to the qualitative results, the effect size analysis provides more granular insights into the guidelines' impact on different UX metrics. We present the results of each guideline in a chart that shows both the effect sizes and the statistical significance of the impact on all dependent variables.

This work contributes specific findings for 16 of the 18 guidelines in the context of productivity applications. The results can support creators of user-facing AI systems in their nuanced prioritization and application of the guidelines. Furthermore, the study illustrates and reflects on the use of factorial surveys for conducting research about user perceptions of AI systems.

2 RELATED WORK

2.1 AI Guidance

As more and more popular consumer products and services become infused with AI, various organizations have advanced guidance relative to AI systems. Some of this guidance pertains to the ethical, or responsible, use of AI (see [53] for a review of this landscape). Other guidance applies to interaction with AI. Most of it encompasses broader practices in designing for AI, such as understanding how and whether AI systems serve user needs [37], understanding the fundamentals of AI and machine learning [49], or giving specific guidance for how to develop specific types of AI systems such as conversational AI [4, 34, 70]. Research scholarship has examined interaction with AI systems and even proposed a set of principles for interaction with mixed-initiative systems as early as 1999 [43]. However, a lot of guidance on human-AI interaction was scattered across the literature, and unusable by product teams building user-facing AI. In 2019, Amershi et al. synthesized this literature into a set of 18 validated guidelines for human-AI interaction [5]. Previous research also supports these guidelines' validity. For example, some of the guidelines, such as the one about providing explanations, draw upon entire bodies of work (see [1, 39, 74] for overviews of explainable AI). Other work has looked at what aspects of AI influence user perceptions such as trust [36, 120]. Therefore, the goal of this article is not to further validate the guidelines' effectiveness, but to contribute nuanced insights about how they impact user preference, user perceptions, and UX metrics.

In the remainder of this section, we review the HAI guidelines and methods previously used for evaluating design principles and AI systems, which led us to factorial surveys as a suitable method for this research.

2.2 Guidelines for Human-AI Interaction

We provide an overview of the guidelines in [5] with a focus on requirements for evaluating their impact on UX. The 18 guidelines synthesize more than 20 years of research and reflections about interacting with AI-powered systems. As part of creating the guidelines, the authors conducted three rounds of validation to ensure the guidelines apply to AI-infused systems and can be understood by HCI professionals.

The guidelines are roughly grouped into four categories, based on when a user would be exposed to them during interaction with a system: initially, during the interaction, when wrong, and over time.

Initially. There are two guidelines in the "initially" group. They are about setting expectations about what the AI system can do (Guideline 1) and how well the system can do what it can do (Guideline 2).

During interaction. The guidelines in this group are about taking into consideration the user's immediate context (Guidelines 3 and 4) and the larger societal and cultural context (Guidelines 5 and 6). Since AI systems adapt to the user's context, and the extent to which they do so has an impact on UX, evaluation of the guidelines has to take context into consideration either by analyzing a system used in the wild or by simulating the context.

When wrong. Guidelines in this group suggest how AI systems should behave when they inevitably make mistakes. The group includes guidelines that help mitigate common errors such as false positives and false negatives and are intended to alleviate the cost of these errors on the user. They recommend supporting efficient invocation (Guideline 7) and dismissal (Guideline 8), efficient correction of system outputs (Guideline 9), scoping services when the AI system has low confidence (Guideline 10), and providing explanations (Guideline 11). Thus, evaluation of the guidelines has to capture situations when the AI system is wrong. Doing so under regular use conditions

in the wild are difficult because it is hard to predict when errors might occur or how to cause them. Simulating errors is more feasible for a systematic evaluation of the guidelines.

Over time. The fourth category is specific to AI systems that learn and adapt over time. The guidelines suggest not only remembering recent interactions (Guideline 12) and learning from user behavior (Guideline 13) but also updating and adapting cautiously (Guideline 14) in order to avoid disruptive changes. Lastly, guidelines in this group encourage requesting feedback (Guideline 15), conveying the consequences of user actions (Guideline 16), providing global controls for users to customize the AI system's behavior (Guideline 17), and notifying users when the system changes its behavior (Guideline 18), for example as a result of a model update. Therefore, evaluation of the guidelines needs to include the effects of longitudinal interaction, either through longitudinal case studies of interaction with a system or through simulation.

This work builds on the contributions of [5] which outlined bestpractices for human-AI interaction based on a synthesis of previous work and showed evidence of their *understandability* and *applicability* as determined by expert UX practitioners evaluating a wide range of AI-infused systems. The goal of this work is to understand, by various UX metrics, *how much end-user preferences and perceptions are impacted* by applying or violating each guideline. In order to assess end-user perceptions of the guidelines without implementing each in an AI product scenario or running costly controlled experiments with existing AI products, we developed a new research protocol involving factorial surveys and administered it to non-expert participants from a crowdsourcing platform. In the next section, we contrast our protocol with methods used in previous studies to evaluate design principles and heuristics and discuss their appropriateness for evaluating HAI guidelines.

2.3 Methods for Evaluating Design Principles

Previous research has evaluated design principles and heuristics in two main ways. One method involves validation of newly proposed heuristics by comparing them with existing sets of design principles. Another method involves evaluation of their UX impact through redesign and comparative usability testing.

One of the main values of design heuristics such as [78] is their ability to identify usability problems in the user interface through inspection methods like heuristic evaluation [80]. When creating heuristics, a common method for validating them is to assess how good they are at identifying usability problems, often as compared to Nielsen's 10 heuristics [78]. This validation method was used, for example, by [51] to validate new heuristics for mobile interfaces, by [92] for video games, [93] for networked multiplayer games, and [75] for virtual worlds.

A somewhat similar validation method was used by [5], who asked HCI experts to identify applications and violations of each proposed guideline in popular AI-infused systems. To motivate system creators to implement design guidance, it is important not only to validate new guidelines but also to understand the impact of applying them on UX. Therefore, we next discuss methods for evaluating design principles' impact on usability and UX.

A common method for evaluating design principles' impact on usability and UX is to redesign an interface based on the design principles under evaluation, then conduct comparative usability testing of the original and redesigned interfaces. For example, [25] and [97] looked at multiple principles derived from human factors research and evaluated if applying those principles improved system usability in healthcare settings. In addition, [58] evaluated a set of principles for augmented reality applications on smartphones by using redesign coupled with comparative usability testing. Similar procedures were used by [113] to assess whether GenderMag, an inclusive design usability inspection method, is able to identify usability issues that, when fixed, improve cognitive inclusiveness. Several other studies such as [83] or [2] discuss case studies of interface redesigns based on

Nielsen's 10 heuristics. Such case studies add to the body of evidence that applying those heuristics improves UX.

While evaluating design principles through redesign is a commonly used method, it is expensive and faces severe feasibility challenges for AI systems. Redesigning systems is complex and costly, as evidenced by several of the studies above. Those studies compared an existing system with a redesigned mockup or prototype, instead of a functional system (e.g., [25, 58, 83, 97, 113]). Redesigning is particularly resource-intensive for AI systems, which are not only expensive to redesign but also cannot be effectively evaluated with mockups and prototypes using Wizard of Oz methods [118]. AI systems' changing nature through learning, adaptation, personalization, and unpredictable errors makes it difficult to identify and cover a representative number of user scenarios to test [42]. Moreover, a redesign study is usually used to evaluate an entire set of guidelines, but has a hard time distinguishing the impact of individual ones. Because of this and the requirements, we identified when reviewing the guidelines (Section 2.2), we found that factorial surveys are well suited for evaluating specific implementations of the HAI guidelines.

2.4 Factorial Surveys

Factorial surveys is a research method that combines classical experiments with survey methodologies. Factorial surveys use short narratives, called vignettes, to represent various levels of independent variables that are too complex or unethical to create and manipulate in real-world or lab situations. Research participants read one or more vignettes and then answer survey questionnaires that measure dependent variables. By supporting systematic combinations of variables, vignette studies make factorial design possible, and thus are able to assess causality [9].

Besides making it possible to study situations that do not yet exist, are complex, are unethical to reproduce, factorial surveys present other research advantages: They are more realistic and rich than traditional surveys [115] and capable of presenting more realistic scenarios than even some experimental situations [9]. They have good validity, as evidenced by studies that found attitudes identified with factorial surveys do predict actual behavior [84, 95].

Factorial surveys have a long history of use in the social sciences [11, 114], where they have been used to investigate respondents' beliefs, attitudes, and judgments on a variety of social, ethical, political, psychological, and sensitive issues (e.g., [11]). More recently, factorial surveys have been used to study computing systems, including AI-powered ones, to assess the acceptability and adoption intention of future scenarios that are not yet in existence [38, 50, 66, 108].

In HCI and technology studies, factorial surveys have been used to examine, for example, attitudes and behavioral responses to phishing (e.g., [33, 88, 90, 101]) and privacy (e.g., [19, 20, 46, 48, 68, 116]). Both topics lend themselves well to factorial surveys because exposing people to real phishing or privacy threats would be unethical, and experimental studies might not simulate a sufficiently rich and complex experience. A different use case for factorial surveys in HCI is the study of IoT smart home technologies that are expensive to create and cannot be represented convincingly with prototypes (e.g., [6, 22, 29, 66, 108]). Factorial surveys have also been used to assess AI systems [38, 50]. Grgic-Hlaca et al. examined AI-assisted decision-making in the judicial arena. They argued that studying causality under ecologically valid situations is challenging because it would require manipulating variables for decision-makers such as judges. They presented legal cases as vignettes to participants on a crowdsourcing platform to assess the effect of AI system advice on decision making [38]. Janbocke et al. used factorial surveys to study automation adoption. The vignettes in their study described a specific work context and a possible interaction with a future AI-based system. They found that vignettes helped respondents envision an AI-powered system that otherwise was difficult to grasp, and recommended the factorial survey method to study AI systems at a pre-deployment stage [50].

In light of the requirements established for studying the HAI guidelines and the affordances presented by factorial surveys, we decided to adopt this method for creating a research protocol that enables understanding each HAI guideline's impact on UX and user perceptions. We explain this protocol next and then showcase the results obtained by using it.

3 METHODS

In this section, we explain the research protocol we used to conduct 18 independent studies, one for each of the HAI guidelines in [5]. In future research, this protocol can also be adapted and used to study different human-AI interaction designs, such as different ways of implementing one HAI guideline or to investigate interaction effects between multiple guidelines.

3.1 Factorial Variables

Each guideline's study uses factorial surveys. We manipulated two variables: *guideline compliance* and *AI performance* in vignettes. Guideline compliance enabled us to assess the difference in the UX impact of applying or violating each guideline. This impact is assessed under two AI performance conditions: optimal and sub-optimal. We treated AI performance as a potential moderator because it added realism to the vignettes and enabled us to inquire into interaction effects. For example, it is possible that applying a guideline might mitigate the negative effects of sub-optimal product performance. We used compliance with the guideline as a *within-subject* factor so we could collect data about how users compare products that apply or violate each guideline. *AI performance* served as a *between-subject* factor. This enabled us to see whether it influences the impact of guideline compliance on user perceptions and the dependent variables (described in Section 3.3).

3.2 Vignette Development

All vignettes described interactions with products from the productivity category: document editors, slide editors, search engines, e-mail applications, and spreadsheet applications. We selected the category of productivity products considering the severity of UX impact for guideline compliance and AI performance. We reasoned that the UX impact might be unnoticeable for low-stakes products such as music recommenders; high-stakes products such as autonomous vehicles and healthcare applications, on the other hand, can lead to extreme harm and should be evaluated by experts, not consumers [5]. Future researchers might choose to test our protocol with other product categories.

We did not constrain the vignettes to one specific product or feature, as the 18 guidelines are not meant to be applied as a whole and at the same time to one product and one feature. Our primary focus here is to assess the UX impact of each individual guideline. Thus, for each guideline, we aimed at selecting the interaction scenario that emphasizes that one specific guideline as much as possible, rather than enforcing the same product, feature, and operation complexity. When possible, the product types and AI features were reused in multiple studies, such as with Guidelines 7, 8, 9, and 13; Guidelines 11 and 15; and Guidelines 17 and 18.

Below we describe our procedure for developing the vignettes for each guideline. We went through two phases of development and each phase had several iterations.

Phase One: Scenario Selection. We went through an iterative brainstorming process. In the diverging stage, two of the authors brainstormed how each guideline could manifest in productivity apps and drafted multiple interaction scenarios. About five to eight interaction scenarios were specified for each HAI guideline. Then we reviewed, rewrote, and when necessary, replaced the scenarios with new ones to be more appropriate for each guideline.

Phase Two: Vignette Composition. As we developed the vignettes, we followed recommended best practices to make them simple, clear, and realistic [10]. We illustrated some with images if

the interaction described in the vignette was not understandable otherwise. We took an iterative approach to develop and pilot the vignettes as recommended by [47]. To mitigate the influence of writing styles, we composed the vignettes using a consistent, three-part structure to describe interaction with an AI-powered feature:

(1) Product and feature introduction. A short statement introducing the fictional product and one of its AI-powered features that the user was described to interact with. Each fictional product was positioned as similar to multiple real products in the same category to help align it with the respondents' experience. We used and recommend using consistent introduction statements, following this format: "You are using a [Productivity Category] app called [Product Name] to do X. It is similar to [example real products A, B, C]. [Product Name] has a feature Y that does ...."

(2) Product behavior description. A statement describing the product's behavior under the guideline application or violation conditions. This statement was used to manipulate the first factorial variable, guideline compliance. The description included a user interaction with the product: "When you do X, the product displays/does...."

(3) Product performance description. A statement about the product's performance was used to manipulate the second factorial variable, AI performance. The statements about AI performance were identical across all vignettes. For optimal AI performance, the statement was: "After using it for a few weeks, you notice that [Product Name] sometimes made mistakes, **but most of the time** worked well." For sub-optimal AI performance, the statement was: "After using it for a few weeks, you notice that [Product Name] sometimes made mistakes, **and sometimes** worked well."

Example Vignettes. To illustrate, the vignettes used to manipulate compliance with G1: Make clear what the system can do, in the optimal AI performance condition, were:

Application Vignette

> You are using a presentation app similar to Microsoft PowerPoint, Google Slides, Apple Keynote to make slides for a presentation. It is called [Product Name]. [Product Name] has a capability called Presenter Coach that gives you feedback on your presentation skills as you practice your presentation in front of your computer.
>
> When you turn on Presenter Coach, it displays information like this:
>
> > As you practice your presentation, we will give you feedback about your presentation style: how fast you speak, use of filler words (such as "um" and "like"), use of inappropriate words (such as "damn").
>
> After using it for a few weeks, you notice that [Product Name] sometimes made mistakes, but most of the time it worked well.

Violation Vignette

> You are using a presentation app similar to Microsoft PowerPoint, Google Slides, Apple Keynote to make slides for a presentation. It is called [Product Name]. [Product Name] has a capability called Presenter Coach that gives you feedback on your presentation skills as you practice your presentation in front of your computer.
>
> When you turn on Presenter Coach, it displays information like this:
>
> > We will help you improve your presentation style.
>
> After using it for a few weeks, you notice that [Product Name] sometimes made mistakes, but most of the time it worked well.

3.3 Dependent Variables

For each guideline, our research protocol collected qualitative data about participants' product preferences and quantitative user perceptions regarding selected UX metrics. The list of dependent variables are adapted from prior research and streamlined based on the results of several pilot studies (Section 3.4).

To understand user preference between [Product A] and [Product V], we asked about the respondent's preference and reasons. "Which product would you prefer to use? Please briefly explain why." We did not want to assume that people would always prefer [Product A]. In fact, we learned a lot from participants who explained why they preferred the violation product after comparing it with the application one.

We assessed quantitatively UX metrics that fall into three major categories: feelings, trust, and perceived product quality. All metrics are measured with 7-point Likert scales and the order of the questions was randomized for each vignette. The protocol can be modified to assess other metrics, as needed.

Feelings. Following the approach to assess feelings in UX [16], we measured how guideline compliance influenced feeling in control, inadequate (reverse-coded), productive, secure, and uncertain (reverse-coded).

Trust. Trust plays an important factor in how much people accept and use smart systems [102]. Jiun-Yin et al. identified 12 factors of trust between human and automated systems [52]. Considering the nature of our vignettes and the feedback from the seven HCI researchers in the pilot study (Section 3.4), we adopted four trust-related items from [52]): trust, reliability, suspicion, and expectation of harm. The latter two were reverse-coded. These items are independent; they do not constitute a trust scale.

Perceived product quality. We also collected metrics related to perceived product quality and user acceptance: **perceived usefulness** (**PU**), perception of product performance, behavioral intention to use the product, and **Net Promoter Score** (**NPS**) [94]. These metrics are known to be related to the acceptance and use of AI-infused systems. The **technology acceptance model** (**TAM**) [30] suggests that when users are presented with a new technology, factors such as PU and **perceived ease-of-use** (**PEOU**) influence their decision about how and when they will use it. Their general perception of the technology [40] influences **behavioral intent** (**BI**), which is a predictor of behavior [3]. We excluded PEOU from our dependent variables because we found in the pilot studies that it could not be assessed from reading a vignette.

In addition to the survey items assessing the dependent variables, we also included three types of gate-keeping questions for each vignette's survey:

Attention check. We asked an attention check question to verify if the respondent read the vignette carefully before answering the dependent variable survey questions. The attention check asked about AI performance and also served as a manipulation check [7] for this variable. This attention check question repeated the product performance description in the vignette word by word. If a participant selected the disagree options from the Likert scale, this was an indicator of not paying attention, and therefore not passing the attention check. If a participant did not pass the attention check, their data was excluded from the analysis.

Vignette comprehension. We asked two questions to check if the participant was able to understand the vignette and whether the consequences of the product's behavior were indeed perceived as medium-stakes.

Manipulation check. We asked two manipulation check [7] questions, one closed, and one open-ended, to verify that the respondent perceived the manipulation of guideline compliance. The closed-ended manipulation check questions took the form of statements mirroring the text of

the guidelines themselves. The open-ended manipulation check question asked participants to describe the difference between [Product A] and [Product V]: "Please briefly describe the differences between Kelso and Ione." (Kelso and Ione are fictional names randomly assigned to [Product A] and [Product V].) If guideline compliance was successfully manipulated, we expected the participants to select the agree options from the Likert scale for [Product A] and the disagree options for [Product V]. More importantly, their answers to the open-ended question will describe the differences between [Product A] and [Product V] that are related to the guideline. We discarded data from two studies that failed to manipulate the independent variable: those for Guidelines 2 and 16.

3.4 Pilot Studies and Results

We conducted two rounds of piloting for vignette development and dependent variable surveys. In the first round, we got qualitative feedback from seven HCI researchers not familiar with the project on the vignettes and the survey. We clarified the vignettes through several rounds of editing with the team, aiming for conciseness, consistency, realism, and clear manipulation of the variables. Based on the first pilot results, we also shortened the dependent variable survey to eliminate items that did not apply, and decided to use fictional product names to facilitate the connection between the product described in the vignettes and the survey questions. We used fictional, gender-neutral names (Kelso and Ione) to refer to the products in the vignettes in order to avoid the influence of brand loyalty, to make the vignettes realistic, and to make it easy to recall the products in the survey. The fictional names were randomly assigned to the products in the two vignettes for each participant, to avoid product preference due to names. We suggest that others using this protocol also pilot their vignettes and identify any issues with wording.

We then conducted a second pilot study with the updated vignettes and survey questions on Amazon **Mechanical Turk** (**MTurk**). We assessed the efficacy of the variable manipulation and the clarity of the vignettes' writing style. The manipulation check questions are described in Section 3.3. The writing style question was "The writing style for the scenario (vignette) was easy to understand." (A 7-point Likert scale question from strongly disagree to strongly agree.) Each vignette was piloted with five participants. That is, five participants * four conditions * 18 guidelines, 360 participants in total. Of those, 352 (97.7%) participants found the vignettes easy to understand. Also, 349 participants (97%) passed the attention check, and 316 (87.7%) participants passed the manipulation check. Due to the nature of crowd work and the small sample size (N = 5), we did not attempt to tailor the vignettes for a small group of crowd workers and achieve 100% pass rate for all gate-keeping questions. This step complements the first pilot to verify that most crowd workers also find the vignettes easy to understand and perceive the difference between [Product A] and [Product V], and the two levels of AI performance.

3.5 Participants

For the actual studies, we determined the sample size with the assumption of 80% statistical power. That is, should any effect exist, there is an 80% chance of detecting a small-medium effect (f = .18). As a result, each factorial survey required at least 65 responses.

We recruited respondents from a general-purpose, paid crowdsourcing platform, Amazon MTurk. Crowdsourcing platforms have the advantage of accessing a wide range of end-users within a reasonable budget. To control for quality, we limited recruitment to crowd workers with acceptance rates above 95% and at least 100 approved **human intelligence tasks** (**HITs**), and who were located in the United States and at least 18 years old. Considering that not all participants would pass the attention check, we over-recruited for each factorial survey (72–74 participants).

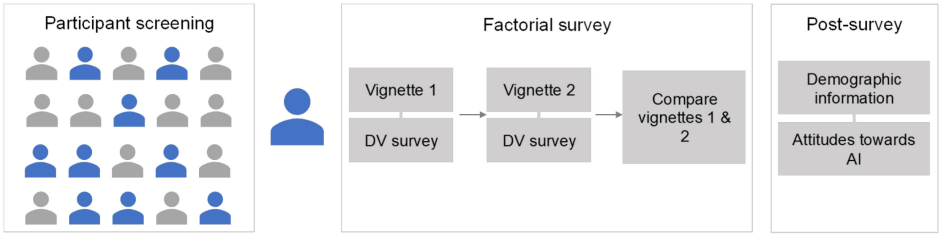

Fig. 1. The study procedure is shown from left to right. A screening survey is presented to the crowd workers to make sure they have some experience with the type of productivity software to be described in the vignettes. Participants who passed the screening test (marked as blue profile icons) will be directed to the factorial survey task. The order of all survey questions is randomized. Vignettes 1 and 2 are randomly assigned to be about [Product A] and [Product V]. After completing the factorial survey task, the crowd workers will fill out the post-survey and submit the HIT.

3.6 Procedure

To facilitate randomization in participation and fair pay for the screening questions, we built a web application to manage participant recruitment on MTurk. Crowd workers were recruited through a screening survey HIT worth $0.20. Only crowd workers who met the recruitment requirements were able to preview and accept the task. Below we describe the study procedure (Figure 1).

Participant Screening. After workers accepted the HIT, they were presented with our institution's IRB consent form. Upon electronic agreement with the consent form, the worker was asked three screening questions. The first two questions asked about their experience with a type of productivity software described in the vignettes. The third question asked the worker's age, to confirm participants were at least 18 years old, per IRB regulations. All crowd workers who completed the screening were paid $0.20. We screened out workers with little or no experience with the type of product (e.g., used the product for less than one year, and less than once every few months). Qualified workers who completed the entire survey were compensated with an additional $5 through MTurk's bonus system.

Factorial Survey. Participants who qualified for the study were then shown the factorial survey. Each participant was first shown a vignette and a set of survey questions assessing the dependent variables. After that, the participants are directed to the second vignette and the same set of survey questions. The order of the vignettes (one about [Product A], one about [Product V]) and the survey questions for each vignette were randomized. After finishing the survey questions for both vignettes, participants were directed to the third page where the same two vignettes were displayed side by side. The participants were asked to describe the differences between [Product A] and [Product V], select which product they preferred, and explain why.

Post-survey. Finally, they were asked to fill out a questionnaire about their demographic information and attitudes towards AI. This information about participants enabled us to monitor whether extreme attitudes towards AI might have influenced their answers. We used the attitudes towards AI items from [119].

3.7 Data Analysis

One of the authors analyzed the short answers to the two open-ended questions in the factorial surveys using thematic analysis [21]. The short answers to the question that asked participants to describe the difference between [Product A] and [Product V] were coded to check whether each participant perceived the independent variable manipulation (Yes/No). For example, the following is a comment that described the difference between the two products for G1: Make clear what

THE SYSTEM CAN DO: "[Product V] doesn't specifically state what features it has. It simply says that [Product V] would help me with my presentation. However, [Product A] does state that it will help me improve by giving me feedback on my use of filler words or profanity." This comment captured the difference between the products and was coded as passing the manipulation check.

The other open-ended question asked the participants to provide reasons for their preference for the product from the two vignettes, [Product A], or [Product V]. These answers were also analyzed using thematic analysis. One of the authors read the open-ended answers provided in each factorial survey repeatedly until codes began to emerge. Then, the codes were recorded and each comment was coded. Codes were mutually exclusive, so a comment counted towards only one code. Once codes were identified and all comments were coded, the same author merged codes into themes.

For the impacts on UX metrics, we measured statistical significance with adjusted p-values [17] and we also investigated effect sizes measured by the index generalized eta-squared [85]. The **American Statistical Association** (**ASA**) released a statement [8] that explicitly advised against drawing scientific conclusions based on p-values alone. First, to mitigate the risk of inflating the false positive rate due to multiple testing, we adjusted the p-values with the methods proposed by Benjamini and Hochberg [17]. This method controls the false discovery rate, which is the expected proportion of false positives among all positives that rejected the null hypothesis and not among all the tests undertaken. To further quantify the impact, we investigated effect sizes measured by generalized eta-squared [85], which is a useful estimate of the percentage of variance accounted for by a variable—in our case, compliance with a guideline. Generalized eta-squared provides comparability across between-subjects and within-subjects designs [12]. Given the lack of proximal research to determine the standards for effect sizes, we followed the recommendation by Cohen [27], which categorizes the effect sizes into four levels: unsubstantiated ($\eta_G^2 < 0.02$), small ($0.02 \leq \eta_G^2 < 0.13$), medium ($0.13 \leq \eta_G^2 < 0.26$), and large ($\eta_G^2 \geq 0.26$).

Because the factorial surveys for each guideline are independent of each other and used different vignettes, we have to consider each guideline's study in isolation and cannot perform comparisons across guidelines.

4 RESULTS

We first describe background information about the participants and the manipulation check results, then we present the details of the qualitative and quantitative analyses.

We conducted thematic analysis on the open-ended responses about product preference and effect size analysis for each UX metric. Since the interaction effects between guideline compliance and AI performance were not significant for most studies, we report the interaction effects between guideline compliance and AI performance only when they are detected. The thematic analysis revealed participants' perceptions about the products, providing context about the guideline's impact on metrics and also pointing to pitfalls in applying and implementing each guideline.

The manipulation check results showed that two of the 18 factorial surveys failed to manipulate the independent variable. We reflect on the issues with the vignettes of the two guidelines, and present the results for each of the 16 remaining studies.

It is important to keep in mind that the results do not support comparisons across guidelines and are discussed independently. Each factorial survey used different products and features in the vignettes, therefore, the 16 studies were independent. Future users of this protocol might choose to collect data only about a few relevant guidelines as needed for their specific product or feature.

4.1 Participants' Background

In total, we collected 1,300 responses from MTurk. As we will describe in Section 4.2, the studies for Guidelines 2 and 16 failed the manipulation check. Therefore, the total number of participants in

the successful studies is 1,155. Of those, we eliminated the responses from participants who failed the attention checks. As a result, 1,043 participants were included in the data analysis. While some of the factorial surveys ended up having less than 65 (the targeted sample size) valid responses (see Table 1), the fact that most medium and large effects are also statistically significant (see Figures 2–20) suggests we had a good amount of power. Therefore, the reason for not seeing effects for some of the dependent variables in some of the factorial surveys is not because of low statistical power.

We present the aggregated background data about the participants who passed the attention checks in Appendix A. This includes a breakdown of participant demographics (Tables 2 and 3) and attitudes towards AI (Tables 4 and 5) by experimental group. There was no statistically significant difference between the participants in the optimal and sub-optimal AI performance conditions in each guideline.

Among the 1,043 responses, 539 (52%) participants identified as female, 488 (47%) as male, 13 (1.2%) as nonbinary or gender nonconforming, two preferred not to answer, one did not respond. Participants skewed young. Their distribution across age groups was as follows: 128 (12%) 18–24 years old; 437 (42%) 25–34 years old; 266 (26%) 35–44 years old; 139 (13%) 45–54 years old; 58 (6%) 55–64 years old; 14 (1%) 65–74 years old, and one participant did not answer.

Most participants had positive prior experiences with the product type in the vignettes they were exposed to: 953 (91%) considered products in the same category useful, and 950 (91%), reliable.

Most participants had some familiarity with computer science/technology either through college-level coursework, degrees, and/or programming experience. Of the 1,043 respondents, 357 (34%) had no such experience.

Participants' attitudes towards AI tended to be positive. Most participants stated they would support the development of AI (852, 82%), 86 (8%) would oppose, 101 (9.7%) were neutral, and four did not know or did not answer. We also asked participants to indicate their feelings toward progress in AI. The most common feelings were curiosity (728, 70%), excitement (528, 51%), and optimism (505, 48%). Negative-leaning feelings such as concern (318, 30%), apprehension (255, 24%), and unease (177, 17%) were not as widespread among respondents. When asked whether society will become better or worse because of increased automation and AI, most participants (736, 71%) indicated better, 166 (16%) thought it would become worse, and 140 (13%) thought it would not change. Overall, our respondents' attitudes towards AI were more favorable than those of the American public [119].

Most of the respondents (1,005, 96%) found the vignettes easy to understand and 24 (2%) were neutral. The product scenarios in the vignettes were perceived to have a medium-stakes impact, confirming that participant perceptions (971, 93%) aligned with our intent to study medium-stakes products.

4.2 Manipulation Check Results

4.2.1 Qualitative Manipulation Checks Were More Effective. When analyzing the manipulation check results, we found severe discrepancies between the quantitative and qualitative ones. Several participants' quantitative answers to multiple experiments indicated the independent variable manipulation had failed. However, their open-ended answers described the difference between the two products in the vignettes accurately and unambiguously. For example, in the study of G6: Mitigate social biases, only 23 (34%) of the 68 participants passed the quantitative manipulation check. However, in the open-ended responses, 66 (97%) of the participants perceived the manipulated differences. A total of 39 (57%) participants recognized that [Product A] was diverse or inclusive, while [Product V] was biased or discriminatory: "[Product V] is biased and not inclusive. [Product A] makes mistakes, but is not biased." An additional 27 (40%) participants pointed

Table 1. Number of Recruited and Included Respondents for the Successful Studies (Excluding Guidelines 2 and 16), and the Corresponding Preference for Products that Applied [Product A] or Violated [Product V]

Guideline	Respondent Inclusion Criteria		Preference	
	Initial/All	Pass Attention Check	Product A	Product V
1	72, 100%	64, 89%	55, 86%	9, 14%
3	72, 100%	68, 94%	55, 81%	13, 19%
4	72, 100%	66, 92%	40, 61%	26, 39%
5	72, 100%	69, 96%	51, 74%	18, 26%
6	72, 100%	68, 94%	64, 94%	4, 6%
7	72, 100%	60, 83%	58, 97%	2, 3%
8	72, 100%	61, 85%	56, 92%	5, 8%
9	72, 100%	60, 83%	56, 93%	4, 7%
10	72, 100%	66, 92%	62, 94%	4, 6%
11	73, 100%	65, 89%	64, 98%	1, 2%
12	72, 100%	69, 96%	60, 87%	9, 13%
13	72, 100%	60, 83%	49, 82%	11, 18%
14	72, 100%	68, 94%	55, 81%	13, 19%
15	74, 100%	66, 89%	60, 91%	6, 9%
17	72, 100%	65, 90%	58, 89%	7, 11%
18	72, 100%	68, 94%	61, 90%	7, 10%
Total	1,155, 100%	1,043, 90%	904, 87%	139, 13%

out that [Product A] showed women and different races, and [Product V] did not: "[Product V] only shows images of white males, while [Product A] shows people of all races and genders."

We attribute this discrepancy to the wording of the quantitative manipulation check items, which mirrors the original scholarly language of the guidelines rather than using a more descriptive wording similar to the vignettes. For example, the manipulation check question for the G6 surveys was phrased as: "[Product Name] mitigates undesirable and unfair stereotypes and biases." As a result, we decided to drop the quantitative results for manipulation checks and only use the qualitative results.

4.2.2 Guidelines 2 and 16 Failed to Manipulate Guideline Compliance. Based on the qualitative analysis, two of the 18 studies (those for Guidelines 2 and 16) failed to manipulate guideline compliance. In the factorial survey for G2: Make clear how well the system can do what it can do, guideline compliance was manipulated by changing one word. When recommending ideas for slide designs, [Product A] said "Here are designs you **might** like" while [Product V] said, "Here are designs you **will** like." Our intention was to indicate that [Product A] communicates that it might make mistakes and [Product V] is overly confident. Out of the 60 participants who passed attention check, 22 (37%) were not able to distinguish between the two vignettes: "The descriptions and images are the exact same. Are you trolling me? I'm sorry. But I see no difference." 20 (33%) participants were able to spot the different wording but did not perceive it as an indicator of how well the product might perform: "[Product V] and [Product A] are essentially the same software both offering me design help. However, the one difference is in the wording of the language, [Product A] says "Designs you might like" as to where [Product V] says "Designs you will like.""

In the factorial survey for G16: Convey the consequences of user actions, we manipulated guideline compliance by changing the product's behavior upon the user clicking Like/Dislike buttons. [Product A] displayed a message after clicking the buttons, while in [Product V], the buttons

briefly changed color after being clicked, without displaying a message. In retrospect, it became clear that [Product V] also conveyed some consequences of user actions with its color change, as 47 (78%) out of 62 participants pointed out: "[Product A] offers verbal feedback on your choices where as [sic] [Product V] uses colors instead." Actually, 17 (27%) participants liked [Product V]'s implicit communication and expressed dislike for [Product A]'s message: "I do not like interfaces that gives [sic] too much info. I would choose [Product A] if it gave the message the very first time, but changed to the method that [Product V] uses afterwards."

The issues with the vignettes for G2 and G16 were not revealed in the pilot studies with crowd workers. Our interpretation is that people can have very different reactions to AI-powered features, and thus, involving enough participants (such as a sample size estimated by power analysis) is critical to reveal issues with vignette design. However, others using this research protocol should also be mindful about over-engineering vignettes and surveys for a specific group of participants.

4.3 Factorial Survey Results

Therefore, we present the results of the 16 factorial surveys that successfully manipulated the independent variable. For each study, we include information pertaining to each research question: the number of participants who preferred [Product A] and [Product V], along with the reasons participants provided for their preferences, which provide insights into their perceptions. For each study, we also present the effect sizes (measured in generalized eta squared, noted as η^2_G) on the dependent variables and, where applicable, interaction effects. Despite η^2_G values ranging from 0 to 1, to help interpret the directionality of the effects, we use the additive inverse of η^2_G to indicate when [Product V] received more positive ratings with the dependent variables. The factorial surveys of most guidelines did not show statistically significant interaction effects between guideline compliance and AI performance on UX metrics (see Table 6 in Appendix A), except for Guidelines 11, 13, and 15 (see Figures 12, 15, 18). Each of these guidelines demonstrated interaction effects on several UX metrics, but only the interaction effect for Guideline 13 had substantiated effect sizes. This also suggests that AI performance did not significantly influence the impact of guideline compliance on user perception and the different aspects of UX we tested. Therefore, we combine the results for both levels of AI performance when reporting the results for each guideline.

When discussing the qualitative results, in an attempt to keep the summaries concise, we focus on dominant themes. For example, if only one or two participants preferred [Product V], their comments are not summarized unless they provide insights into nuances of applying a guideline.

G1: Make Clear what the System Can Do. The vignettes described a coaching feature in a presentation application. [Product A] informed users of the presentation coach's capabilities with a detailed list of features, while [Product V] used a generic statement: "We will help you improve your presentation style."

We saw a strong preference for [Product A] (55, 86%). The qualitative results show that 19 participants preferred [Product A] because of the feature's specific description: "Both "coaches" aim to tweak presentations, but [Product A] explicitly states how it functions. Based off the narrative, I don't know much about how [Product V] specifically aims to improve presentation skills." These participants' comments are consistent with the substantiated effects on feeling less uncertain and more secure. Another 12 participants expressed their understanding that [Product A] would provide more detailed feedback than [Product V]: "Again, it seems more useful. I don't need general feedback, I need specific information in order to improve and [Product A] has that." Even though the intention of the vignette was to be careful and not overstate what the presentation coach can do, some respondents interpreted [Product A] as being able to do more, which might explain the substantiated effect on perceived performance ($\eta^2_G = 0.02$). The rest of the participants had various

Fig. 2. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 1. Applying Guideline 1 has substantiated small effects on *feeling more secure*, *feeling less uncertain*, and *perceived performance*. None of the adjusted p-values for the dependent variables suggest statistic significance (all are greater than 0.05).

reasons for preferring [Product A], such as finding it useful or innovative (6), liking what it did (4), wanting to improve their presentations (3), and other (7). Among the participants who preferred [Product V], three did not feel comfortable being recorded by the product, and two wrote statements favorable to [Product A], suggesting they might have selected preference for [Product V] by mistake.

While the qualitative analysis focused on the fundamental themes of the user perception, several comments do indicate specific impact by the coaching feature. The three participants who did not feel comfortable being recorded might prefer [Product A] in a different product where no recording is needed. However, those who preferred [Product A] expressed a preference for specificity in feature description, which is consistent with accepted best practices in writing for user interface [72]. In addition, previous works on rehearsal support and feedback systems [100, 109, 111] primarily focused on the quality improvement when evaluating the efficacy of the systems. This study complements the previous evaluation strategies and reveals additional user concerns about being recorded. A unique finding and important consideration when applying Guideline 1 is to find a balance between providing a clear, specific description of the AI system with not over-stating system capabilities.

G3: Time Services Based on Context. The vignettes described an email application that would stop notifications when it senses that the user is busy [Product A] or pops up notifications as messages came in [Product V].

[Product A] was preferred by 55 (81%) participants. Those who preferred [Product A] appreciated that it protected the user's focus (25) and that it knew when it is appropriate to show notifications (24): "I like the fact that [Product A] can predict when I am too busy to deal with notifications popping up on my screen and disturbing my concentration"; "[Product A] seems to have a better understanding of when notifications are appropriate and inappropriate." Consistent with the qualitative results, the quantitative results showed a medium effect on feeling more productive ($\eta^2_G = 0.15$). Interestingly, even though it was the product that controlled notifications, [Product A] also made participants feel more in control ($\eta^2_G = 0.16$). However, the desire for control was one of

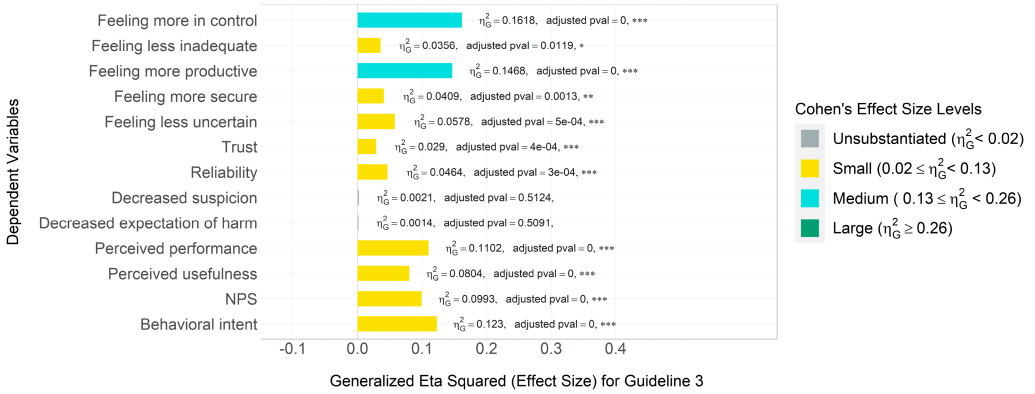

Fig. 3. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 3. Applying Guideline 3 has substantiated effects on most dependent variables, except *decreased suspicion* and *decreased expectation of harm*. All substantiated effects are also statistically significant (asterisks next to the bars indicate that the adjusted p-value for the corresponding dependent variable is smaller than 0.05, indicating statistic significance).

the reasons five participants stated for preferring [Product V]. Another five participants preferred [Product V] because they wanted to see notifications as they came in, and three respondents had privacy concerns: "[Product A] is tracking me somehow because it holds the notifications, so I prefer the less invasive [Product V]."

The results reflect an interesting tension about control: some participants felt more in control when the AI protected their focus; others felt more in control when the product showed notifications as new messages arrived. The findings suggest that when AI systems make decisions for users, these decisions can increase convenience but may also decrease the perception of control. This tension can happen in other products that apply G3, not only e-mail applications. Prior research on notifications has also pointed out challenges such as user stress and feelings of hindrance [59] and costs of interruptions [44, 45]. This study extends previous findings by showing factors for successfully timing the services based on context, such as providing user with sufficient control, making clear what information is tracked by the app, and clarifying privacy concerns.

G4: Show Contextually Relevant Information. The vignettes described a document editing application with a feature that provided definitions of acronyms. [Product A] showed definitions of acronyms specific to the user's workplace and relevant to the user's document. [Product V] always used a standard list of definitions from a popular dictionary.

Respondents were quite split in terms of their preferences, with 26 (39%) actually preferring [Product V]. This is aligned with the lack of substantiated effects in the quantitative results. Those who preferred [Product A] (40) did mention reasons such as it being more tailored to their work. However, the participants who preferred [Product V] raised various concerns about [Product A]: Some (11) perceived [Product A] as being too limiting: "I would prefer to see all of the possible definitions as opposed to having the software narrow the options for me." Six participants raised concerns of trust or possible errors: "I would rather use [Product V] because it gives me the choice of choosing which definition I would want to go by. [Product A] would be easier, but if [Product A] were to make a mistake on me, I would have a hard time trusting it because I did not make any part of the decision." Four participants were concerned about privacy: "I would prefer to use

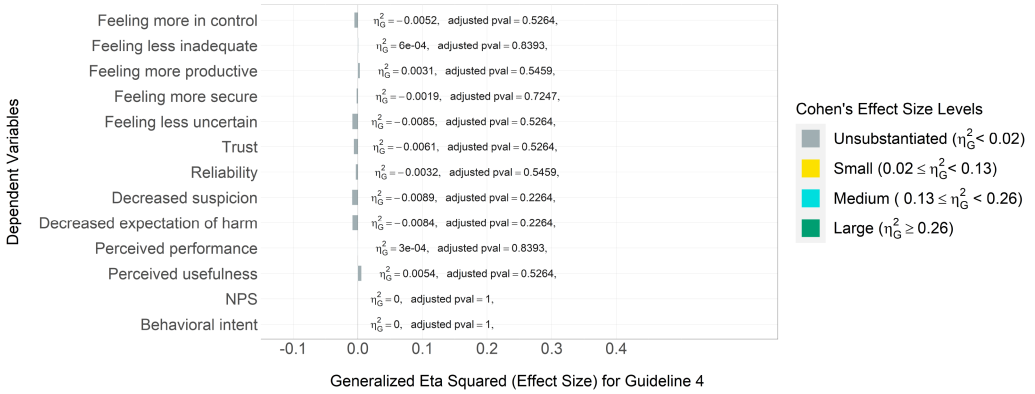

Fig. 4. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 4. Applying Guideline 4 had unsubstantiated effects on the dependent variables, and in some cases, [Product V] received more positive ratings with some UX metrics (the additive inverse of η^2_G was used to indicate directionality). None of the adjusted p-values for the dependent variables suggest statistic significance (all are greater than 0.05).

[Product A] because [Product V] feels a bit more intrusive. I would be nervous that it is pulling data from things like my other software and my browsing history. This would be unacceptable since I work with sensitive PII [personally identifiable information]."

The choice of the acronym feature revealed a common dilemma in personalization: how to customize content without creating a filter bubble [87]. The results also suggest that modern users of personalized features are exceptionally aware of these issues and desire more control over their information exposure. This is aligned with the findings of more extensive studies with deployed recommendation systems [41, 77]. It is also possible that the findings for this study were overly influenced by the phrasing of the vignette, which led to [Product A] being perceived as more limiting than we had intended. However, some of the results point to a well-known tension between privacy and personalization [60].

G5: Match Relevant Social Norms. The vignettes described a document editing application. [Product A] introduced suggestions for improving writing style with the statement, "Consider using..."; however, [Product V] used a different tone: "You made a mistake. Replace with...".

Most respondents (51, 74%) preferred [Product A]. Participants were able to perceive the difference in tone as a matter of politeness, and 49 participants referred to this in their open-ended comments: "I consider the "You made a mistake." rather obnoxious. If you are saying something like that, you need to be absolutely perfect, no "*sometimes* made mistakes" allowed." However, 13 out of the 18 participants who preferred [Product V] actually liked its tone: "I like the way it is blunt." They interpreted it as a sign of confidence: "[Product V] sounds more confident, which makes me trust it more than I trust myself. [Product A] would just be a nuisance to me." It is not surprising that the polite tone had an effect on feeling less inadequate. Applying the guideline also had a small effect on feeling more secure ($\eta^2_G = 0.02$).

The results indicate some disagreement among participants about what might be acceptable social norms. For some participants, the more blunt tone was acceptable, too. It is possible that had the vignette used offending language, the results might have been different. Even so, the results point out the importance of conducting user research to understand social norms and what is

Fig. 5. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 5. Applying Guideline 5 has substantiated small effects on *feeling less inadequate* and *feeling more secure*. There is also an unsubstantiated negative effect on *reliability* (the additive inverse of η^2_G was used to indicate directionality). The impact on *feeling more secure* and *perceived performance* showed statistic significance (adjusted p-values smaller than 0.05, see asterisks in the chart).

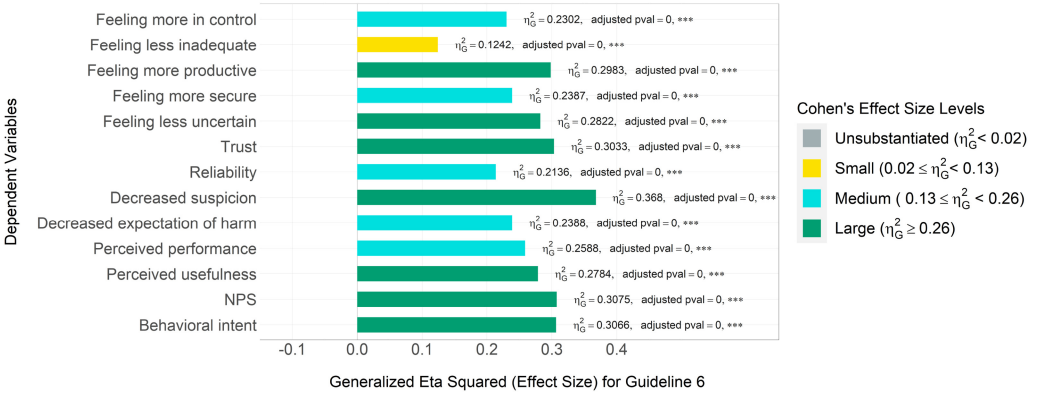

Fig. 6. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 6. Applying Guideline 6 has substantiated effects on all dependent variables with statistic significance (adjusted p-values smaller than 0.05, see asterisks in the chart).

acceptable to different user groups. The results also indicate an interaction between G5 and G2, as the more blunt tone was interpreted as a sign of confidence and better product performance. It is important to take G2 into consideration when applying Guideline 5 to describe system behaviors or capabilities, so as to avoid creating unrealistic expectations about the AI system.

G6: Mitigate Social Biases. The vignettes described an online search engine. When searching for images of CEOs and doctors, [Product A] showed people of different genders and skin tones while [Product V] did not show images of women or people of color.

[Product A] was preferred by 64 (94%) respondents and resulted in substantiated effects on all UX metrics, one being small and the others being either medium or large. Most (34) participants'

reasons for preferring [Product A] converged around the theme that [Product A] did not have bias, or an agenda with bad intentions: "Although it makes mistakes sometimes as well, it seems more well-intentioned than [Product V]. Or rather, [Product V]'s creators." This is consistent with the large effects on UX metrics related to trust ($\eta^2_G = 0.30$). Another 14 participants referred to benefits of diverse search results: "Diversity of results (all relevant to the search) allows me easy access to a variety of content, and gives me a degree of control over what content is available for me to use." Yet another reason for preferring [Product A] was the perception that it reflected reality, mentioned by 9 participants: "I work in the medical field and understand that many positions have more women than men filling the roles *and* to have a search result that excludes them, as well as individuals of different races, is an inaccurate depiction of the occupation." Together, these reasons align with effects on the UX metrics related to productivity ($\eta^2_G = 0.30$) and perceived performance ($\eta^2_G = 0.26$). Among the four participants who preferred [Product V], two indicated that [Product A] had an agenda: "Most CEOs and doctors are white males. I feel like the first one is forcing its agenda of being inclusive." One participant seems to have clicked [Product V] by mistake, as their open-ended comment actually describes [Product A]:

> "I believe the [Product V] search engine mitigates undesirable and unfair stereotypes and biases. Every individual deserves to be treated equally. I understand that we all have preference in choosing which person we feel we are comfortable to communicate with, but through [Product V] system, every individual may have the chance to excel better than before."

Due to increased media coverage (e.g., [31, 71, 117]), the general public is sensitive to issues of social bias in AI systems [82], and in U.S. society at large. As in society at large, feelings about social biases were intense, as suggested by the strong language in participant quotes. Mitigating social biases is a complex issue, and even though it is difficult to achieve, there is convergence in academia and industry about the importance of doing so (see [82] for a survey of this topic).

G7: Support Efficient Invocation. The vignettes were about a feature in a presentation application that suggested alternative slide layouts. [Product A] had a button to invoke the feature with layout design ideas if it did not trigger automatically, while [Product V] did not have a button for manual invocation.

[Product A] was preferred by 58 (97%) participants. The main reason was that [Product A] provides the option to request design help manually, which was mentioned by 33 participants. A total of 17 participants found [Product A] more user-friendly and efficient, and 3 participants felt it offered more freedom and control: "I would prefer being able to get help easily if I need it. [Product A] seems more useful;" "[Product A] gives me more control over the program and allows me to tell when I want to apply suggestions about formatting." Participants' comments are consistent with the effects on UX metrics about feeling in control ($\eta^2_G = 0.21$), productive ($\eta^2_G = 0.23$), and those related to perceived product quality.

The results of G7 point to the fundamental user need to have easy access and control to invoke the available features in an app, which is not limited to slide editors or layout recommendation features and is consistent with existing design heuristics about user control (e.g., [79, 105]). Of course, with other types of interfaces, the specific interaction for efficient invocation would be different—e.g., gestures, voice commands [35, 64], but the finding about G7 supporting user control would still apply.

G8: Support Efficient Dismissal. The vignettes used the same features as for G7, but manipulated guideline compliance through the presence or absence of a button to dismiss slide design suggestions when they were not needed.

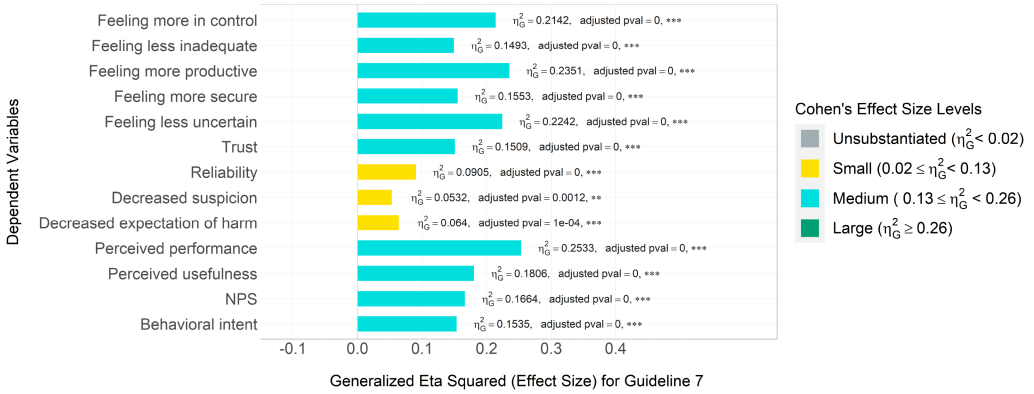

Fig. 7. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 7. Applying Guideline 7 has substantiated effects on all dependent variables with statistic significance (adjusted p-values smaller than 0.05, see asterisks in the chart).

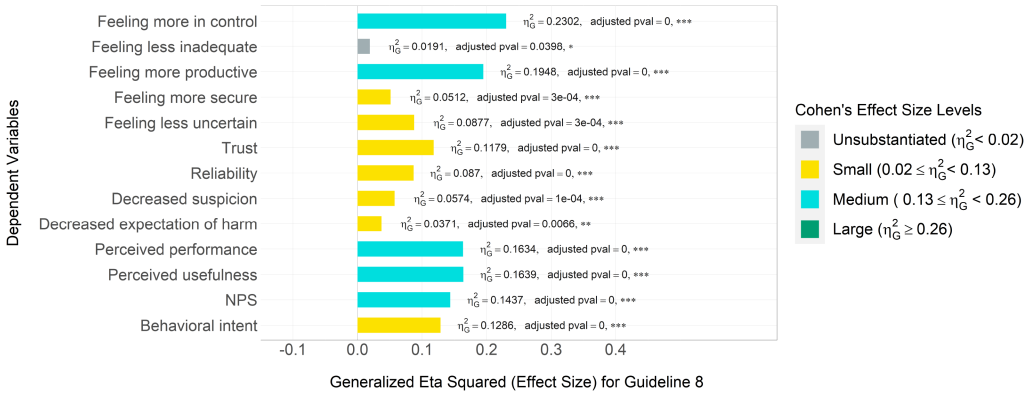

Fig. 8. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 8. Applying Guideline 8 has substantiated effects on most dependent variables, except *feeling less inadequate*. The impact on all dependent variables is statistically significant (adjusted p-values smaller than 0.05, see asterisks in the chart).

Results showed that 56 (92%) participants prefer [Product A]. The presence of the dismissal option was the main reason for preferring [Product A], mentioned by 28 participants. The other 27 participants commented on various UX aspects that differentiated [Product A] from [Product V], such as: **choice and control:** "[Product A] seems to give me a little more control over my user experience;" **frustration:** "Because I would get frustrated not being able to get the help off the screen"; **productivity:** "Being able to remove/hide a tool that is not needed at the time will allow for more productivity and less frustration"; **user-friendliness:** "It's more intuitive and has a more well-designed interface." Feeling in control ($\eta^2_G = 0.23$) and more productive ($\eta^2_G = 0.19$) were also apparent in the quantitative effects on UX metrics, as were three metrics related to perceived product quality: usefulness ($\eta^2_G = 0.16$), performance ($\eta^2_G = 0.16$), and NPS ($\eta^2_G = 0.14$).

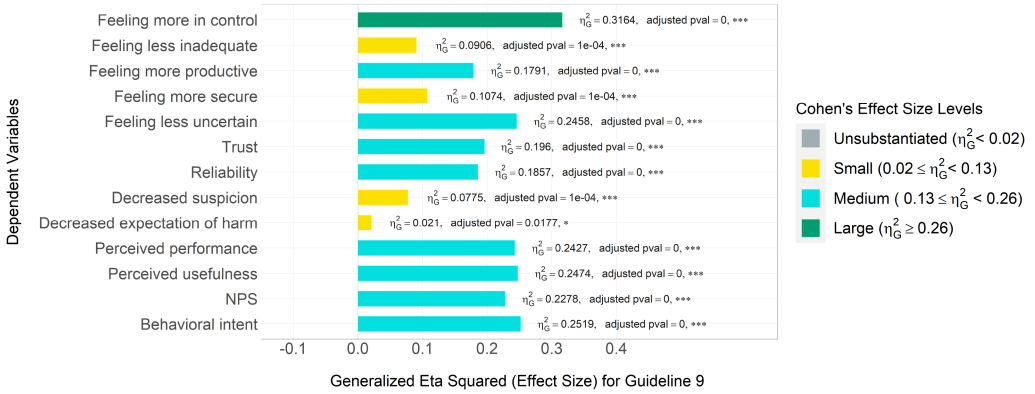

Fig. 9. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 9. Applying Guideline 9 has substantiated effects on all dependent variables, all with statistic significance (adjusted p-values smaller than 0.05, see asterisks in the chart).

Similar to G7, efficient dismissal supports user control and might be implemented differently in other interface types. We recommend AI practitioners follow UX research best practices and assess the effectiveness of a planned interaction, potentially using the same research protocol presented here.

G9: Support Efficient Correction. The vignettes described the same layout suggestions feature as in the studies for G7 and G8. Upon selecting one of the design recommendations, [Product A] allowed the user to make further changes to the layout, while [Product V] did not allow such changes.

The results for G9 show that 56 (93%) respondents preferred [Product A] and all effects on UX metrics were substantiated. Besides stating that they liked the ability to modify the suggested layout (32 respondents), 25 participants' open-ended comments expressed feelings consistent with the quantitative effects on the UX metrics: **feeling in control** ($\eta^2_G = 0.32$): "I do not like feeling out of control. I like to be able to alter the layout to suit my individual needs"; **feeling productive** ($\eta^2_G = 0.18$): "[Product A] allows you to edit things such as size and reposition things which would make the job easier to do"; **reliability** ($\eta^2_G = 0.19$): "[Product A] is more reliable as it allows myself, the user, to fix mistakes that the software will make from time-to-time"; **trust** ($\eta^2_G = 0.20$): "I am happy for the supported help of the Design Helper, but I don't fully trust it, and I want to feel like I am in control of my slide deck software. Therefore, [Product A] seems like a better fit for me." Participants also expressed feelings not captured in the quantitative UX metrics: **freedom and flexibility:** "I like having the freedom of being able to adjust the suggested layout if needed"; **avoiding frustration:** "The ability to customize and make adjustments on [Product A] is a significant improvement over [Product V]. [Product V] would make me very frustrated and unhappy if I couldn't make things just right."

The results about G9 also point to fundamental user needs to have control over AI systems and be able to edit their outputs.

G10: Scope Services when in Doubt. The vignettes described an auto-complete feature in a document editing application. [Product A] provided a list of options when it was not sure which word

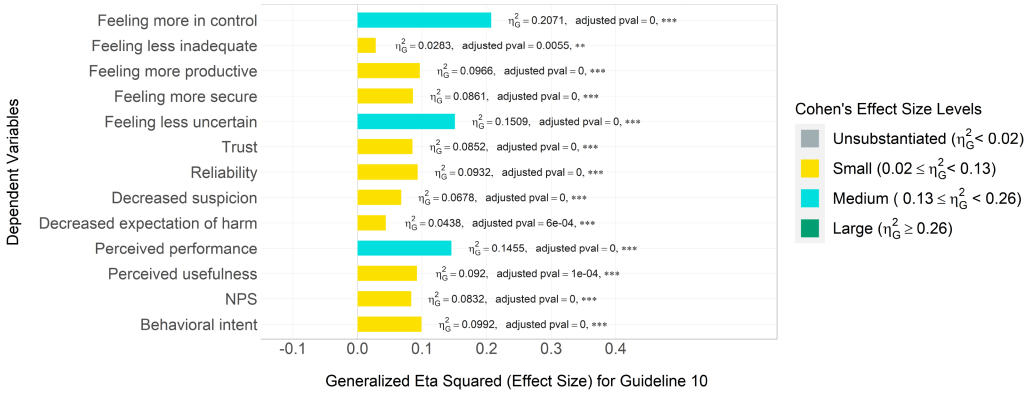

Fig. 10. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 10. Applying Guideline 10 has substantiated effects on all dependent variables with statistic significance (adjusted p-values smaller than 0.05, see asterisks in the chart).

the user is trying to type. [Product V], on the other hand, automatically completed a word after the user typed the first few letters with its best bet.

[Product A] was preferred by 62 (94%) participants. In the open-ended comments, 31 respondents perceived [Product A] as able to reduce the likelihood of errors, which is consistent with the medium effect on perceived performance ($\eta^2_G = 0.15$) and reducing uncertainty ($\eta^2_G = 0.15$). Another 19 participants simply stated they preferred it because of the way it worked. Participants mentioned [Product A] made them **feel in control:** "I would feel more in control, because I would have a choice to make, instead of one being made for me," which is also reflected by a medium effect in quantitative results ($\eta^2_G = 0.20$); **more efficient:** "I would prefer the speed of which I can click on the word to avoid having to type the whole word myself. [Product V] would confuse me and make me sad"; and **less trusting of [Product V]:** "I don't trust [Product V]'s automatic features," echoing a small effect on trust ($\eta^2_G = 0.09$).

The results about G10 indicate that handing over the control from a less confident AI system to a human user is important for maintaining a positive UX. In this case, the cost of engaging the user in disambiguation was lower than the cost of making a mistake. There might be scenarios, where engaging in disambiguation dialogues can be distracting to the user (e.g., while driving). When applying Guideline 10, it is important to consider the relative costs of engaging in disambiguation versus just degrading services when the system is in doubt. These insights align with previous works using simulation [65] or implemented systems [107].

G11: Make Clear why the System Did what it Did. The vignettes used a spreadsheet application that generated insights and recommended charts. [Product A] made available an explanation for these recommendations, whereas [Product V] did not.

All but one participant (64, 98%) preferred [Product A]. Besides liking that [Product A] provided an explanation, mentioned by 29 respondents, participants stated [Product A] helped users understand and check for errors: "If the software is making mistakes then an explanation is 100% needed to avoid frustration and inaccuracy," reinforcing the importance of making available an explanation especially when the AI might be wrong [1, 5]. In fact, this guideline is one of the few that showed a statistically significant interaction effect with AI performance. Applying the guideline

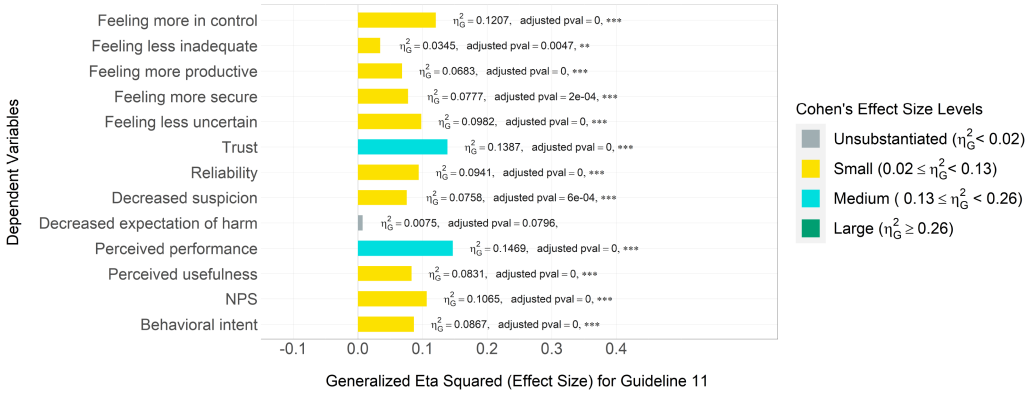

Fig. 11. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 11. Applying Guideline 11 has substantiated effects on most dependent variables, except *the decreased expectation of harm*. All substantiated effects are statistically significant (adjusted p-values smaller than 0.05, see asterisks in the chart).

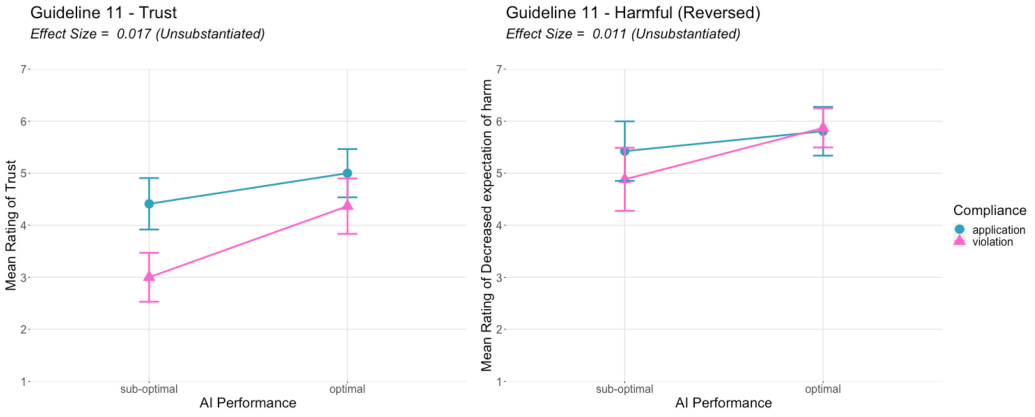

Fig. 12. Interaction effects were detected between compliance of Guideline 11 and AI performance on two dependent variables: *Trust* and *Expectation of Harm*. While the adjusted p-values indicate that the interaction effects are statistically significant, the effect sizes for both interaction effects are unsubstantiated.

made a bigger difference for trust (p = 0.04) and decreased expectation of harm (p = 0.03) when AI performance was sub-optimal (Figure 12). Not surprisingly, applying Guideline 11 also resulted in a medium effect on perceived product performance ($\eta^2_G = 0.15$). Additionally, 6 participants mentioned explanations were useful or valuable, and 6 participants saw [Product A] as more reliable or trustworthy, consistent with the quantitative UX effect on trust ($\eta^2_G = 0.14$) and reliability ($\eta^2_G = 0.14$): "I would trust [Product A] more seeing that it gives you more access to important information."

The results point to the need for explanations of AI system behavior. This principle is not limited to a specific product or feature. The results echo previous findings that the mere presence of an explanation increases user trust in an AI system [56, 96], and are just as concerning. Careful considerations are needed about how to design, implement, and deliver explanations to users in

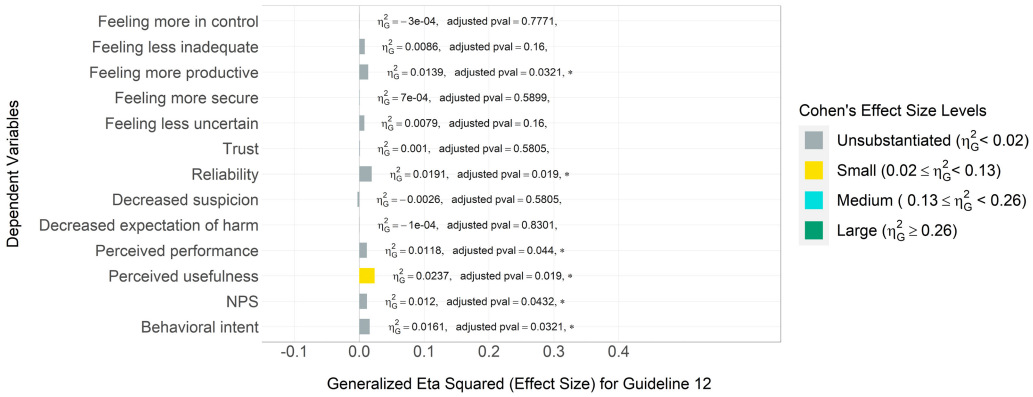

Fig. 13. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 12. Applying Guideline 12 has a small effect on *PU*, which is also statistically significant. The effects on other dependent variables are all unsubstantiated, including a negative effect on *decreased suspicion* (the additive inverse of η^2_G was used to indicate directionality).

different scenarios. For example, recent work has found that explanations can lead to inappropriate trust and over-reliance on AI systems [24, 86], or even poor performance in human-AI collaboration [14].

G12: Remember Recent Interactions. The vignettes described an e-mail application. When attaching a file, [Product A] showed a list of recent files to choose from, whereas [Product B] opened a standard file explorer window for the user to navigate to files.

[Product A] was preferred by 60 (89%) participants. In the open-ended answers, 30 participants made comments about various aspects of [Product A]'s better UX, which are consistent with the substantiated effect on PU ($\eta^2_G = 0.02$). Some aspects mentioned by the participants were not included among the UX metrics we measured: **convenience:** "I like how it shows my recent files because its more convenient"; **ease of use:** "I would want it to remember recently used files to make my work easier"; **user-friendliness:** "I am more likely to reopen a file that I have recently looked at, so I think [Product A]would be more user-friendly for my purposes." The aspect of **productivity** was measured in the UX metrics and two participants mentioned it: "I certainly like the option of a program allowing me to simply click and attach a recent file. This for me, is incredibly productive without having to constantly find where a file may or may not be located to attach it." Surprisingly, this is inconsistent with the unsubstantiated effect ($\eta^2_G = 0.01$) in quantitative results. This could be explained by quotes from the nine participants who preferred [Product V], who thought the features were not needed: "Don't ever use when a program shows the recent files. Not needed"; or were concerned about privacy: "I don't really want my email application looking at what I'm working on. I just want it to be neutral and allow me to choose what I want to send."

The results point to a general privacy concern about how much information is acceptable for the system to track. This concern is much more general and not bound to an email app or a file explorer, and can be especially prevalent in the **Internet of Things** (**IoT**) and AI-infused cyber-physical systems [112].

G13: Learn from user Behavior. The vignettes described a presentation application with the same layout helper feature used in G7 through G9. [Product A] personalized its design

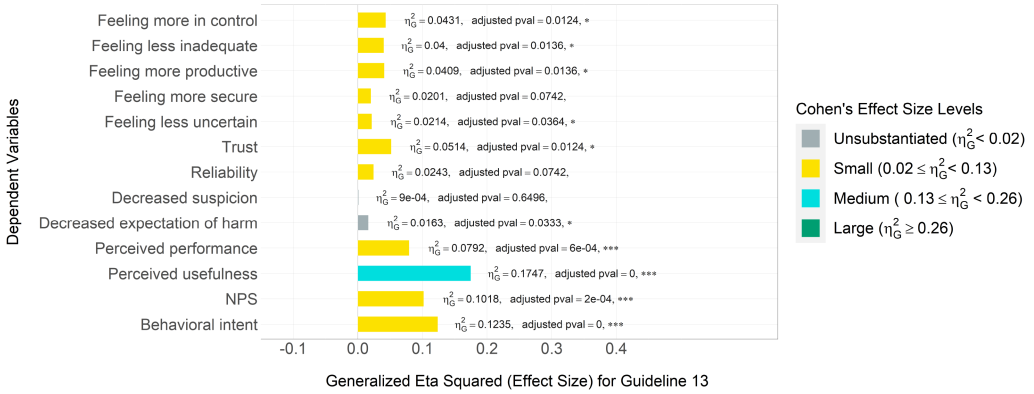

Fig. 14. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 13. Applying Guideline 13 has substantiated effect sizes on most dependent variables except *decreased suspicion* and *decreased expectation of harm*. The impact on most dependent variables is statistically significant (adjusted p-values smaller than 0.05, see asterisks in the chart), except *decreased suspicion*.

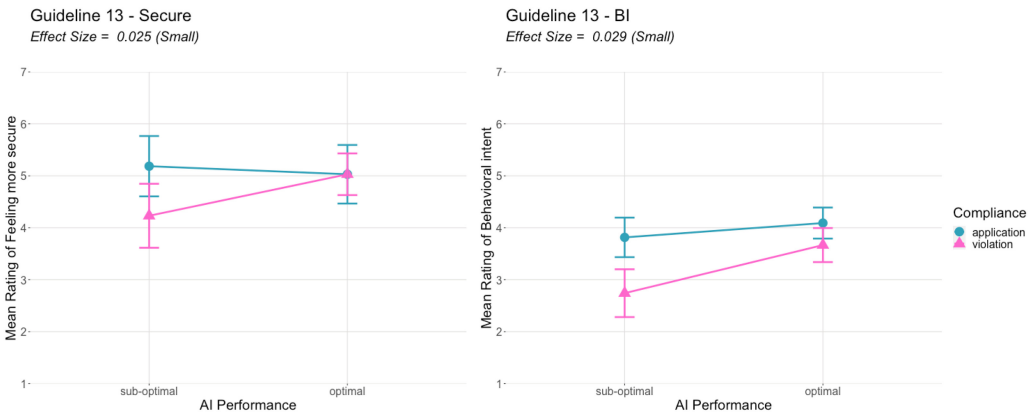

Fig. 15. Interaction effects were detected between compliance of Guideline 13 and AI performance on two dependent variables: feeling of security and BI. Both interactions have small substantiated effect sizes. In other words, applying Guideline 13 could help mitigate the impact of sub-optimal AI performance on users' feeling of security and their BI.

suggestions based on previous user behavior, while [Product V] always showed the default recommendations.

[Product A] was preferred by 49 (82%) participants. Participants who preferred it not only liked that it learned user preferences (23), but also found it more efficient (18), which is consistent with the medium effect on PU ($\eta^2_G = 0.17$). We also observed statistically significant and substantiated interaction effects on two UX metrics between guideline compliance and AI performance. In the sub-optimal AI performance condition, applying Guideline 13 resulted in improved effects on BI ($p = 0.02, \eta^2_G = 0.03$) and feeling secure ($p = 0.04, \eta^2_G = 0.03$). Participants' open-ended answers such as the one below suggest users might be more likely to tolerate sub-optimal AI performance

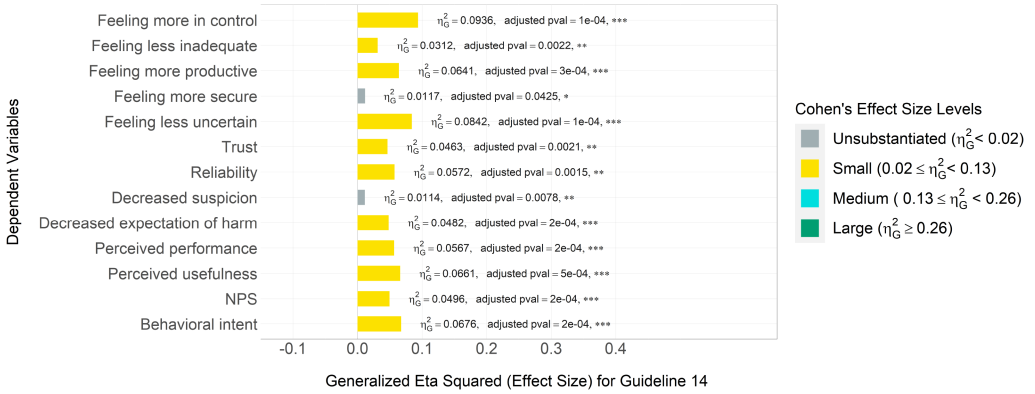

Fig. 16. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 14. Applying Guideline 14 has substantiated small effects on most dependent variables except *feeling more secure* and *reliability*. The impact on all dependent variables is statistically significant (adjusted p-values smaller than 0.05, see asterisks in the chart).

when there is an indication that the system is learning and might improve over time. "From the two [Product A] seemed to have a better user experience. Although it did make some mistakes they outweigh the ease of use and learning ability of [Product A]."

This guideline focuses on the system "learning" from user behavior. Note that some participants assumed that this learning would directly improve system behaviors. Because participants appear to conflate two guidelines, learning user behavior (G13) without actually improving the system appropriately (G14) could be perceived as problematic. In practice, it is also important to convey to users how their interactions are used for system improvement or other purposes (G16). Because in practice these three guidelines can be deeply interrelated, it is important to consider their interaction when designing and evaluating human-AI interaction. Prior research has studied adaptive user interfaces that learn from user behaviors in various contexts such as autonomous driving [110] and e-learning [62], to name a few. Our study results confirm the positive effects of adapting user interfaces based on user behaviors, and also points to the importance of combining multiple guidelines when designing such adaptive user interfaces.

G14: Update and Adapt Cautiously. The vignettes described a document editing application that adapted a part of its menu to the user's current actions, as opposed to [Product V], which changed its entire menu.

[Product A] was preferred by 55 (81%) participants. Among those who preferred [Product A], 17 participants liked [Product A]'s consistency or found [Product V] disruptive: "I feel like [Product V] would constantly change the entire menu bar and these changes would be disruptive or distracting to me as I tried to work. [Product A] would be less intrusive." These perceptions might align with the small quantitative effects on feeling less inadequate ($\eta^2_G = 0.03$), less uncertain ($\eta^2_G = 0.08$), and reliability ($\eta^2_G = 0.05$). Consistent with the substantiated effects on control ($\eta^2_G = 0.09$) and productivity ($\eta^2_G = 0.06$), 15 respondents mentioned these aspects in their open-ended comments: "I like having control of the main functions that I find useful and not having the entire bar change would be more beneficial for me"; "Since both programs occasionally make mistakes, I would be more productive when using [Product A] because I would remember where the tool buttons are."

Fig. 17. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 15. Applying Guideline 15 has substantiated small effects on most dependent variables except the *decreased expectation of harm*. The impact on all dependent variables is statistically significant (adjusted p-values smaller than 0.05, see asterisks in the chart).

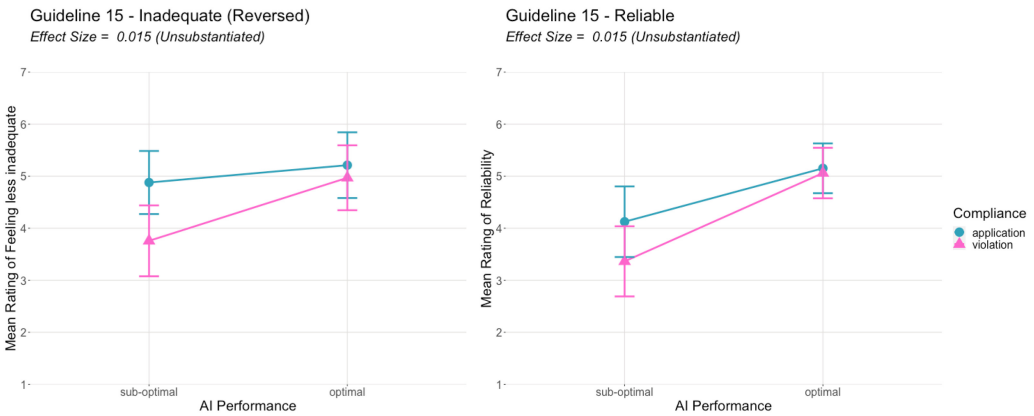

Fig. 18. Interaction effects were detected between compliance of Guideline 15 and two dependent variables: feeling of inadequacy and reliability. While the adjusted p-values are smaller than 0.05, the effect sizes for the interaction were unsubstantiated.

The results point to the importance of having a consistent UX and reducing the burden of learning an updated system, consistent with existing research on backward compatibility [13]. Previous work has recognized that maintaining consistency in adaptive user interfaces is challenging [57]. Future research can use this research protocol to experiment with multiple design ideas before investing in implementation.

G15: Encourage Granular Feedback. The vignettes were about a spreadsheet application. [Product A] had an option to provide feedback on suggested charts, while [Product V] did not.

A total of 60 (91%) participants preferred [Product A]. The strong preference for [Product A] is also reflected in the substantiated effects in almost all UX metrics, except the decreased expectation of harm. In the open-ended comments, 25 participants simply stated that the reason for their

Fig. 19. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 17. Applying Guideline 17 has substantiated small effects on most dependent variables, except *feeling less inadequate*, *feeling more secure*, *decreased suspicion*. The impact on the most dependent variable is statistically significant, except *feeling less inadequate* (adjusted p-values smaller than 0.05, see asterisks in the chart).

preference was the availability of the feedback feature, but others provided more nuanced reasons. Another 21 participants thought that because it asked for feedback, [Product A] would learn and adapt to user needs: "I feel that if I can let the program know which features are useful to me and which aren't, it may get better at predicting which features to suggest to me." This could explain the statistically significant interaction effects on perceptions on reliability (p = 0.04) and feeling less inadequate (p = 0.03): "I feel like in the long run, it will become more reliable, unlike [Product V] which has no way of knowing what it is doing wrong, making it unreliable."

The results suggest that asking for granular feedback from the users (G15) sets the expectation that the system will learn from the feedback (G13) and improve over time (G14). In a longer-term interaction scenario, user perception might change over time depending on system performance. Because these guidelines appear to be deeply interrelated, it is important to consider their interaction when designing and evaluating human-AI interaction.

G17: Provide Global Controls. The vignettes described an email application. [Product A] provided a setting where the user could teach the system that certain contacts as important, so their emails always went to the"Important" inbox and not be miscategorized by the system in the "Other" inbox. [Product V], on the other hand, did not have this global control.

[Product A] was preferred by 58 (89%) participants. Of those, 35 participants chose [Product A] because of the availability of the feature. Another 19 participants associated [Product A] with aspects of UX that are consistent with the small effects in quantitative results: feelings of **control** ($\eta^2_G = 0.10$): "Because it somehow gives you some control of the e-mails you want to be marked as important."; **trust** ($\eta^2_G = 0.05$): "With [Product A] I have more trust that the emails are where they are supposed to be ..."; **reliability** ($\eta^2_G = 0.03$): "[Product A] has more user functionality so you can customize the app to your liking and seems more reliable."; and **usefulness** ($\eta^2_G = 0.06$): "The setting to classify people would be useful if it works.". In the conditions with sub-optimal AI performance, four participants raised concerns about possible mistakes "Because it [Product V] is not promising something like marking people as important going forward and then failing to

Fig. 20. Effect sizes measured in generalized eta squared (noted as η^2_G, represented by the length of the bars and the first values next to the bars), adjusted p-values (noted as *adjusted pval*, the second values next to the bars), and the significance levels (represented by the number of asterisks next to the bars, no stars indicate no significance) of Guideline 18. Applying Guideline 18 has substantiated small effects on multiple dependent variables with statistic significance (adjusted p-values smaller than 0.05, see asterisks in the chart).

deliver something like that. I would expect [the product] to fail and because of that I would be watching it more closely and manually monitoring it."

The results revealed that users prefer to have control over system behaviors, but might have additional concerns when system performance is sub-optimal. When system performance is sub-optimal, users might not trust that the global controls will influence system behavior in the way they desire. This is not limited to a specific app or feature.

G18: Notify users About Changes. The vignettes used the same sorting feature in an email application as in G17. [Product A] would send a notification when the e-mail categorization feature underwent an update and changed the way it worked, while [Product V] did not.

[Product A] was preferred by 61 (90%) participants. When explaining their preference, 51 participants stated they wanted to know and stay informed: "I would want to know changes coming ahead of time"; "I would prefer to use [Product A] because it would notify me when it made changes, which would make me feel like I was more in the loop with what was going on and that I wouldn't be surprised when I signed onto my email account." These sentiments align with the quantitative effects of feeling more secure ($\eta^2_G = 0.04$), less uncertain ($\eta^2_G = 0.03$), and perhaps trusting the product more ($\eta^2_G = 0.02$). Also, participants mentioned [Product A] would make them feel more in control, which also showed a substantiated quantitative effect ($\eta^2_G = 0.03$): "Because it keeps me aware of what actions is taking, make me feel more secure and in control of [Product A]. With [Product V] I feel like I'm left in the dark."

The results suggest that the action of notifying users of system changes influences UX positively and minimizes surprises caused by a system update. This result is consistent with industry best practices (e.g., [76]), but in reality might be contingent on the assumption that users will pay attention to such notifications.

5 DISCUSSION

Overall, the results support the effectiveness of the guidelines [5] and highlight the importance of assessing ideas for each guideline's implementation in context. Our 16 successful factorial surveys suggest that the guidelines have a positive effect on UX, and explain associated user perceptions.

We also find that applications of the guidelines are not always successful or translate to positive user preference. There are various ways of applying each guideline and the same design might be received differently or even contrarily by different users. Multiple iterations with a sufficient target audience are especially important for identifying a more suitable way to apply a guideline in a given context. For example, the studies of Guidelines 2 and 16 failed partially due to the subtlety of our selected ways of applying the guidelines. We also weren't able to identify the issues with a small number of participants in the pilot study. Even with the successful studies, the participants' comments, as well as the effects on UX metrics, provide detailed insights about each guideline's application. Particularly in instances where some participants preferred [Product V], the results show possible pitfalls of guideline application that need to be considered and mitigated. While this article focuses on early-stage assessments, it is also important to validate specific human-AI interaction designs with users in real systems. The research protocol used in our studies could be adapted and repurposed to this end.

The results of the 16 successful factorial surveys support the application of the HAI guidelines. For some guidelines, such as Guidelines 6 through 9, the results strongly and enthusiastically support their application. Similarly, the results support the application without reservations of Guidelines 1, 10, 11, 14, 15, 17, and 18. Guidelines 11 and 15 showed interaction effects with AI performance that indicate their application might mitigate the effects of sub-optimal performance.

For some of the guidelines, however, the results suggest the need to consider tradeoffs, primarily concerning privacy. That was the case for Guidelines 3, 4, 12, and 13, whose implementation relies more on passive data collection (without explicitly asking the user for the data). One possibility to navigate this tradeoff is to couple these guidelines with Guideline 17, to empower users to globally control what data the system collects. In fact, the results point to other possible interactions among the guidelines, which would occur in a naturalistic context.

The results also provided unique insights into applying the guidelines in human-AI interaction design. We highlight a few here. Some guidelines can influence a user's expectations about the AI system. For example, a product applying G1: Make clear what the system can do can sometimes be interpreted as being able to do more. Similarly, a more blunt tone (violating G5: Match relevant social norms) was occasionally interpreted as a sign of confidence and better product performance. Therefore, it is important to also consider G2: Make clear how well the system can do what it can do when applying G1 and G5 to avoid creating unrealistic expectations about the AI system. In addition, the study results also revealed an interesting tension between control and convenience. In the study about G3: Time services based on context, some participants felt more in control when AI systems make a decision to pause notifications for them, while others felt more in control when the product did not make such decisions. A related insight is from the study about G17: Provide global controls, where participants became doubtful about the system under sub-optimal performance conditions.

The findings point to design implications for using the HAI guidelines to create human-AI interaction, which we discuss below.

5.1 Design Implications for Human-AI Interaction

A Guideline's Impact Can Be Influenced by Other Guidelines and AI Performance. For example, the comments for Guidelines 3 and 5 suggest how multiple guidelines might interact with each other to improve UX. From the participants who preferred [Product V] for G3: Time services based on context, we learn that automatically suspending notifications when the system senses the user is busy does not work for everyone, either because some people want to see all their notifications promptly or because they want more control. These comments suggest that pairing the application of Guideline 3 with the application of G17: Provide global controls could provide better UX. If

Guideline 17 were also applied, users would be able to globally control how they prefer to receive notifications. The study for G5: Match relevant social norms points out the need to balance a polite tone with communicating system confidence. This suggests connections with G2: Make clear how well the system can do what it can do. While some participants preferred the blunt tone and found it more confident, it is important to use language that calibrates user expectations of AI performance, as also suggested by [61]. Indeed, in real AI products and features, HAI guidelines would be used in parallel to orchestrate UX. Future research is needed to rigorously assess how multiple guidelines interact, building on their individual impact as reported in this article.

In terms of mitigating sub-optimal AI performance, the interaction effects suggest that participants interpreted applying G13: Learn from user behavior and G15: Encourage granular feedback as a promise that the system will improve over time. Applying G11: Make clear why the system did what it did and also helped mitigate sub-optimal AI performance with regards to trust and perception of harm. However, the results for this guideline, consistent with previous literature, also point to potential pitfalls.

Pitfalls of Guideline Application. Applying a guideline appropriately not only requires a deep understanding of the guidelines, the products, and the users, but it also takes multiple iterations. While no AI systems can produce a perfect UX, there are some reoccurring pitfalls that we observe and recommend designers watch out for.

For G11, our results are consistent with previous findings that the mere presence of an explanation can increase trust [56, 96]. Note that the G11 vignette did not provide the contents of an explanation–it just stated an explanation was available. While this is an encouraging finding for applying G11, it also implies a potential pitfall: Users might trust the system just because an explanation is available, regardless of how informative the explanation is. This can lead to over-trusting the system. When applying G11, it is important to be aware of the benefits and drawbacks of different kinds of explanations [61, 73] and be cautious about not over-inflating user trust in the system.

From the studies for G4: Show contextually relevant information and G13: Learn from user behavior, we learn about another pitfall of applying HAI guidelines: Some people might perceive contextualization and personalization as limiting, in that it does not allow them to see all available options. These concerns echo larger issues related to personalization, such as the Internet filter bubble [87].

Moreover, in multiple studies, some participants raised privacy concerns. In the case of G1: Make clear what the system can do, the privacy concern was due to the nature of the feature in the vignette, a presentation coach that would listen to the user delivering presentations. But for Guidelines 3, 4, and 12, which are all related to personalization, some participants stated they did not want their activities to be tracked to the degree that [Product A] did. The privacy paradox [15] might explain some of these findings. However, even if user acceptance of technologies that leverage personal data outweighs stated privacy concerns, it is important to align products and features with users' privacy preferences [55].

In our studies, however, each guideline was applied in only one way for that feature. This research protocol can also help collect user perceptions about different ways to implement each guideline. For example, there are different types of explanations [74] that can be used to apply G11: Make clear what the system can do. While we demonstrated that using this research protocol produces useful insights, it is also important to consider its limitations.

5.2 Limitations

While previous research supports the validity of factorial surveys [84, 95], reading a scenario cannot be an exact substitute for interacting with a product in the context of daily life and over a

period of time. Future work is needed to further investigate how the results from this research protocol compare to what we would see through actual use in a system in repeated sessions over time. Using this research protocol; however, can provide insights into user perceptions when it is too early in the planning process or too costly to develop and test an AI-powered prototype. Our intention is that it will help with planning human-AI interaction early in the design and development process, but not replace other assessments of the users' experience.

It is possible that the results of our studies would have been different if the vignettes described different products or even different features of the same products. We attempted to mitigate this limitation through careful vignette design, so that each vignette focuses specifically on a behavior clearly corresponding to one of the HAI guidelines. We characterized the results of each study in light of the specifics of the product in the vignette and in the context of existing research. Many of our findings echoed insights from previous research, as pointed out at the end of each study's results section. This reinforces our results' validity and transferability to other product types. In terms of generalizability, in line with best practices for analyzing the validity and rigor of qualitative research [91], the question is not whether the results are generalizable, but whether the insights we derived about each guideline can be transferred to different contexts—such as different products or features. Participants' responses about their perceptions frequently made reference to principles that are by now universally accepted, such as the need to feel in control. This evidence supports the transferability of the findings. The quantitative effect sizes provide further evidence, but we do not have a statistical basis for claiming that, with other products or features, the effect size results would be similar. Nonetheless, more in-depth studies on potential confounding factors are needed to understand their impact.

Our studies considered the effects of guideline compliance from the average perspective of a "universal user." Because we opted for breadth instead of depth, we did not investigate UX effects for particular, well-defined, or intersectional groups [28]. Future research could probe deeper into particular guidelines and identify the nuanced effects of their application on particular groups. For example, people with different valences on the different cognitive facets explained in [23] are likely to be served differently by different ways of implementing the same guideline.

5.3 Implications for Assessing Human-AI Interaction with Factorial Surveys

It is useful to weigh the pros and cons of using a vignette-based study to assess human-AI interaction. In comparison to user interviews, this vignette-based research protocol allows larger-scale data collection with crowdsourcing, which mitigates potential biases in a small sample of users. In fact, the sample size (N = 5) in the pilot study failed to detect the issues with the vignettes in G2 and G16, which also suggests the importance of involving sufficient number of participants in similar studies.

In comparison to case studies, the use of factorial surveys allows researchers to better control the experiment variables and study a greater number of different human-AI interaction designs at an early stage. Low-fidelity prototypes can be included in the vignettes, where the descriptions in the vignettes can provide additional contexts about the AI systems' behaviors.

Two key limitations of this research protocol are the lack of in-person interaction with the users to collect more nuanced data about participants' perceptions and the lack of realistic interaction with an implemented system. In summary, when the goal is to validate human-AI interaction designs at an early stage, or when the number of experiment variables exceeds the implementation and deployment capacity, this research protocol should be considered and can be used to collect evidence about user perceptions of multiple human-AI interaction designs.

Of course, when using factorial surveys for assessing human-AI interaction, it is important to pay attention to the limitations listed in Section 5.2. If the purpose is to validate a specific design

for a specific feature and product, the issue of generalizability to other product types does not apply, but it is important to use best practices in vignette creation so as not to influence participant responses, and to conduct manipulation checks to ensure participants can notice the differences among alternate designs. In addition to careful vignette design and conducting manipulation checks, we offer four additional lessons learned from our studies:

First, using fictional product names and providing some real-world examples can mitigate the influence of brand loyalty and biases from participants.

Second, determining an appropriate sample size with power analysis is important for the successful detection of potential issues.

Third, randomizing the order of factors that are not independent variables, such as survey questions and fictional product names, can help mitigate unnecessary confounding.

Fourth, our methods also demonstrate the importance of data triangulation [26]. By complementing quantitative data about product preference and UX effects with qualitative data, we were able to access nuanced insights about the guidelines' application, learn about potential pitfalls, and explain the quantitative results. While it is possible to add to or change the set of UX metrics we used, quantitative metrics alone risk missing potentially important information. Just because a guideline does not show effects on some of the quantitative metrics, this does not mean it does not impact UX. It is important to mitigate this limitation by collecting qualitative data. As the results for G12: Remember recent interactions show, applying this guideline resulted in only one small substantiated effect, but participants had a strong preference for [Product A] and their open-ended answers commented on various UX aspects that our metrics did not capture. This is one possible reason why the effect sizes for some of our studies were small or unsubstantiated.

5.4 Implications for Vignette Design

Our results suggest the importance of testing how a guideline is applied to a given product. There are multiple ways of implementing each guideline, and different user groups might perceive the same implementation differently. Therefore, it is indispensable to assess how potential users perceive each proposed guideline implementation. We find that factorial surveys provide an efficient way to collect data about guideline implementation and user preference. Because factorial surveys rely so heavily on vignettes, we expand on lessons learned about how to write vignettes to use with this research protocol.

To facilitate understanding, the vignettes we used elicited participants' previous knowledge by grounding the fictional products among well-known products in the same category: *You are using a presentation app similar to Microsoft PowerPoint, Google Slides, Apple Keynote to make slides for a presentation.*

To simulate interaction with the system over time, we used brief statements such as *After using [Product Name] a few times, you notice it has learned your preferences and now features blue designs prominently.* Participants referred to this longitudinal aspect of interaction in their comments, which suggests they noticed and considered it in their responses.

Conveying different levels of AI performance was a challenge, as we were not able to describe how the specific feature performed sub-optimally without using negative language that would influence respondents' perceptions. Therefore, we used generic statements about product performance. The results show that participants perceived the difference because they passed the manipulation check for the AI performance variable (same as the attention check) and made reference to product performance in their open-ended comments. However, the lack of interaction effects with AI performance suggests that this statement might not be sufficient. The lived experience and frustration of sub-optimal AI performance were not appropriately substituted for. Until future

research creates a better way of simulating AI performance in vignettes, this variable could be omitted when using this protocol.

The utility of randomizing the fictional product names was also confirmed by a few participants' comments that showed their product preference was motivated by liking the product's name. We aimed to avoid referring to products using letters or numbers that would imply hierarchy. Therefore, we picked two short, gender-neutral names from a list of small towns: Kelso and Ione. Multiple participants expressed a preference for the name "Kelso," which we had not anticipated, but had mitigated through randomizing the product names.

Even though we used a systematic approach to vignette writing, following guidelines from the literature [10], the pilot tests were indispensable. Despite our two rounds of pilot testing, there were problems with manipulating the independent variable for Guidelines 2 and 16 that did not surface in the pilot tests.

We hope these reflections on vignette design can help future researchers interested in using this protocol for evaluating AI products or features.

Future research could collect more evidence about this method's ecological validity. While previous studies have shown that factorial survey results can be used to predict behavior [84, 95], it would be useful to prove this in the context of interacting with a computing system.

Furthermore, the vignettes we developed and tested artificially separated the guidelines, when, during regular interaction with a product, they might interact. Future research could create strategies for assessing the guidelines as they interact with each other.

6 CONCLUSION

In this work, we provided insights about how each of the 18 HAI guidelines [5] impacts product preference, user perceptions, and UX metrics. The results provide nuanced design implications for the application of each guideline and suggest pitfalls to avoid. The results also highlight deep connections among some of the guidelines and suggest how they should be used and assessed in practice.

The results also point to the importance of assessing user perceptions of planned human-AI interactions early, before investing in engineering. The use of factorial surveys in a research protocol similar to the one we employed provides a feasible option for what has traditionally been a challenge in the UX of AI [118]. UX researchers can draw upon our experience with using factorial surveys to compare multiple design ideas.

The participants' comments suggest that different ways of implementing the guidelines in different features and types of products, as well as different audience and interaction scenarios, could lead to different user perceptions of the same product design. We hope that the research protocol used in this work can be adapted and validated in future research and contribute to the design of effective and responsible human-AI interaction.

Future work is needed to consider how these guidelines interact with each other, as they are not meant to be used alone. Also, future research could compare various ways of applying each guideline and assess their effectiveness for specific, intersectional user groups.

APPENDIX

A PARTICIPANT BACKGROUND AND INTERACTION EFFECTS

Table 2. Age of Participants

	AI Performance	18–24 years old	25–34 years old	35–44 years old	45–54 years old	55–64 years old	65–74 years old	Didn't Respond
G1	Optimal	2	19	5	4	2	0	1
	Sub-optimal	5	17	8	1	0	0	0
G3	Optimal	1	14	9	6	3	0	0
	Sub-optimal	4	15	6	8	2	0	0
G4	Optimal	4	13	9	6	0	0	0
	Sub-optimal	7	11	8	4	4	0	0
G5	Optimal	3	11	13	2	2	2	0
	Sub-optimal	3	16	11	2	3	1	0
G6	Optimal	5	14	7	7	2	0	0
	Sub-optimal	4	10	8	7	2	2	0
G7	Optimal	2	20	6	3	0	0	0
	Sub-optimal	5	13	3	5	3	0	0
G8	Optimal	2	14	7	6	2	0	0
	Sub-optimal	7	14	5	4	0	0	0
G9	Optimal	2	15	7	3	0	0	0
	Sub-optimal	5	14	10	3	1	0	0
G10	Optimal	2	13	10	8	0	0	0
	Sub-optimal	4	10	9	5	4	1	0
G11	Optimal	2	17	7	3	1	1	0
	Sub-optimal	2	11	9	10	2	0	0
G12	Optimal	7	5	14	1	6	1	0
	Sub-optimal	6	13	9	4	2	1	0
G13	Optimal	7	16	8	1	1	0	0
	Sub-optimal	0	17	4	5	1	0	0
G14	Optimal	4	15	11	2	0	2	0
	Sub-optimal	3	17	9	4	1	0	0
G15	Optimal	3	11	11	7	1	0	0
	Sub-optimal	4	13	7	6	3	0	0
G17	Optimal	4	10	13	3	1	1	0
	Sub-optimal	5	15	7	2	4	0	0
G18	Optimal	6	14	8	3	2	1	0
	Sub-optimal	8	10	8	4	3	1	0
	Total	128	437	266	139	58	14	1

Table 3. Genders of the Participants

	AI Performance	Woman	Man	Non-binary	Gender-non-conforming	Prefer-not-to-answer	Didn't Respond
G1	Optimal	15	17	0	0	0	1
	Sub-optimal	7	24	0	0	0	0
G3	Optimal	17	15	1	0	0	0
	Sub-optimal	21	12	1	0	1	0
G4	Optimal	21	10	1	0	0	0
	Sub-optimal	21	11	1	0	0	0
G5	Optimal	18	15	0	0	0	0
	Sub-optimal	20	15	0	0	1	0
G6	Optimal	17	18	0	0	0	0
	Sub-optimal	18	15	0	0	0	0
G7	Optimal	11	20	0	0	0	0
	Sub-optimal	18	11	0	0	0	0
G8	Optimal	6	25	0	0	0	0
	Sub-optimal	15	14	1	0	0	0
G9	Optimal	11	16	0	0	0	0
	Sub-optimal	20	13	0	0	0	0
G10	Optimal	20	12	0	0	0	0
	Sub-optimal	22	11	0	0	0	0
G11	Optimal	14	17	0	0	0	0
	Sub-optimal	14	20	0	0	0	0
G12	Optimal	20	14	0	0	0	0
	Sub-optimal	21	13	0	0	0	0
G13	Optimal	14	18	0	0	0	0
	Sub-optimal	15	12	0	0	0	0
G14	Optimal	17	17	0	0	0	0
	Sub-optimal	19	15	0	0	0	0
G15	Optimal	23	10	0	0	0	0
	Sub-optimal	15	18	0	0	0	0
G17	Optimal	16	15	0	0	0	0
	Sub-optimal	17	16	0	0	0	0
G18	Optimal	18	15	0	0	0	0
	Sub-optimal	18	14	2	0	0	0
Total		539	488	8	5	2	1

Table 4. Attitude Towards AI Question: *How Much do you Support or Oppose the Development of AI?*

	AI Performance	Strongly oppose (1)	Somewhat oppose (2)	Neither support nor oppose (3)	Somewhat support (4)	Strongly support (5)	Didn't Respond	Average Rating
G1	Optimal	0	2	6	9	15	1	4.16
	Sub-optimal	0	1	4	12	14	0	4.26
G3	Optimal	0	3	4	10	16	0	4.18
	Sub-optimal	0	1	3	18	13	0	4.23
G4	Optimal	0	1	4	6	21	0	4.47
	Sub-optimal	0	4	4	16	10	0	3.94
G5	Optimal	0	0	3	18	12	0	4.27
	Sub-optimal	1	3	2	19	11	0	4
G6	Optimal	2	2	3	11	17	0	4.11
	Sub-optimal	1	2	3	14	13	0	4.09
G7	Optimal	1	2	1	14	13	0	4.16
	Sub-optimal	2	3	1	16	7	0	3.79
G8	Optimal	0	5	2	11	13	0	4.03
	Sub-optimal	0	0	6	11	13	0	4.23
G9	Optimal	1	1	4	12	9	0	4
	Sub-optimal	3	0	4	11	15	0	4.06
G10	Optimal	0	2	7	16	8	0	3.91
	Sub-optimal	1	0	0	15	17	0	4.42
G11	Optimal	1	4	1	14	11	0	3.97
	Sub-optimal	0	2	2	13	17	0	4.32
G12	Optimal	2	1	3	12	16	0	4.15
	Sub-optimal	0	2	4	10	19	0	4.31
G13	Optimal	0	1	2	16	14	0	4.3
	Sub-optimal	0	2	1	13	11	0	4.22
G14	Optimal	0	3	1	14	15	1	4.24
	Sub-optimal	1	2	4	15	12	0	4.03
G15	Optimal	0	2	4	18	9	0	4.03
	Sub-optimal	1	3	2	11	16	0	4.15
G17	Optimal	1	3	4	15	9	0	3.88
	Sub-optimal	0	3	7	13	10	0	3.91
G18	Optimal	0	5	2	14	12	1	4
	Sub-optimal	1	2	3	18	9	1	3.97
Total		19	67	101	435	417	4	4.12

Table 5. Attitude Towards AI: *Do you Think that Society Will Become Better or Worse from Increased Automation and AI?*

	AI Performance	Much worse (1)	Worse (2)	Won't change (3)	Better (4)	Much better (5)	Didn't Respond	Average Rating
G1	Optimal	0	4	5	18	5	1	3.75
	Sub-optimal	0	4	1	23	3	0	3.81
G3	Optimal	1	3	5	19	5	0	3.73
	Sub-optimal	0	6	3	22	4	0	3.69
G4	Optimal	0	3	2	21	6	0	3.94
	Sub-optimal	0	8	5	19	2	0	3.44
G5	Optimal	1	3	3	21	5	0	3.79
	Sub-optimal	1	2	9	20	4	0	3.67
G6	Optimal	0	6	6	19	4	0	3.60
	Sub-optimal	0	7	4	18	4	0	3.58
G7	Optimal	1	4	2	23	1	0	3.61
	Sub-optimal	3	4	2	19	1	0	3.38
G8	Optimal	1	5	3	15	7	0	3.71
	Sub-optimal	0	4	3	16	7	0	3.87
G9	Optimal	1	5	3	13	5	0	3.59
	Sub-optimal	2	2	6	19	4	0	3.64
G10	Optimal	2	6	8	14	3	0	3.30
	Sub-optimal	2	5	4	20	2	0	3.45
G11	Optimal	3	4	3	19	2	0	3.42
	Sub-optimal	1	4	4	22	3	0	3.65
G12	Optimal	2	2	6	19	5	0	3.68
	Sub-optimal	2	6	4	17	6	0	3.54
G13	Optimal	1	2	3	23	4	0	3.82
	Sub-optimal	1	2	6	14	4	0	3.67
G14	Optimal	0	4	3	21	6	0	3.85
	Sub-optimal	1	4	6	18	5	0	3.65
G15	Optimal	0	4	8	20	1	0	3.55
	Sub-optimal	0	2	3	21	7	0	4.00
G17	Optimal	3	6	5	16	2	0	3.25
	Sub-optimal	0	7	4	18	4	0	3.58
G18	Optimal	0	5	3	23	3	0	3.71
	Sub-optimal	0	4	8	18	4	0	3.65
Total		29	137	140	608	128	1	3.65

Table 6. Interaction Effects for Each Dependent Variable in Each Factorial Survey, Measured in F and P-value, and Generalized Eta-squared

DV	Stat	G1	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G17	G18
Control	F	2.19	0.01	0.97	0.12	1.58	0.71	1.32	0.87	0.23	0.61	0.29	3.82	0.67	0.40	0.02	0.67
	p	0.14	0.93	0.33	0.73	0.21	0.40	0.26	0.36	0.63	0.44	0.59	0.06	0.41	0.53	0.90	0.42
	ges	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
Inadequate	F	0.30	0.00	2.05	0.08	0.05	0.62	0.24	0.36	0.03	0.07	0.93	0.09	0.01	4.99	0.10	0.31
	p	0.59	0.97	0.16	0.77	0.83	0.44	0.63	0.55	0.87	0.79	0.34	0.76	0.92	**0.03**	0.75	0.58
	ges	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
Productive	F	1.61	2.61	3.87	0.23	0.16	0.11	0.01	2.65	0.01	0.21	0.12	0.18	0.62	0.29	1.40	2.89
	p	0.21	0.11	0.05	0.63	0.69	0.74	0.91	0.11	0.94	0.65	0.73	0.68	0.43	0.59	0.24	0.09
	ges	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Secure	F	0.63	1.66	0.08	0.04	0.06	0.29	0.00	0.00	1.80	2.76	1.24	4.36	1.01	3.41	0.72	0.11
	p	0.43	0.20	0.78	0.84	0.81	0.59	1.00	1.00	0.18	0.10	0.27	**0.04**	0.32	0.07	0.40	0.74
	ges	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.03	0.00	0.01	0.00	0.00
Uncertain	F	0.04	0.03	0.01	0.01	0.57	1.12	0.88	0.33	1.10	3.50	0.00	0.62	0.78	1.21	0.61	0.41
	p	0.85	0.86	0.92	0.91	0.45	0.30	0.35	0.57	0.30	0.07	0.96	0.44	0.38	0.28	0.44	0.53
	ges	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00
Trust	F	1.03	0.09	0.40	1.22	0.46	0.00	0.73	1.31	1.24	4.61	0.01	0.01	0.36	0.05	0.02	3.85
	p	0.32	0.77	0.53	0.27	0.50	0.99	0.40	0.26	0.27	**0.04**	0.92	0.94	0.55	0.82	0.90	0.05
	ges	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01
Reliability	F	0.22	0.55	0.14	0.34	0.00	1.08	0.12	2.58	0.95	2.35	3.08	0.00	1.94	4.25	0.81	0.03
	p	0.64	0.46	0.71	0.56	0.99	0.30	0.73	0.11	0.33	0.13	0.08	0.99	0.17	**0.04**	0.37	0.86
	ges	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.00
Suspicion	F	0.04	0.03	0.75	1.42	0.02	0.08	0.49	0.08	0.31	0.72	1.34	0.91	0.54	0.77	0.02	0.07
	p	0.85	0.87	0.39	0.24	0.88	0.78	0.49	0.78	0.58	0.40	0.25	0.34	0.47	0.38	0.89	0.80
	ges	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
Harmful	F	1.43	0.51	0.01	0.01	2.16	0.03	0.04	0.20	2.06	4.72	0.19	0.78	0.02	0.12	0.48	0.27
	p	0.24	0.48	0.94	0.91	0.15	0.87	0.84	0.66	0.16	**0.03**	0.67	0.38	0.87	0.74	0.49	0.61
	ges	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Performance	F	1.51	0.03	0.31	1.90	0.01	0.22	0.33	2.40	1.70	0.48	0.07	0.12	0.41	0.62	3.41	2.07
	p	0.22	0.87	0.58	0.17	0.92	0.64	0.57	0.13	0.20	0.49	0.79	0.73	0.53	0.43	0.07	0.15
	ges	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01
Useful	F	1.36	0.01	0.72	1.25	1.93	1.14	2.20	3.10	0.43	0.02	0.01	2.43	0.27	1.04	0.36	1.35
	p	0.25	0.93	0.40	0.27	0.17	0.29	0.14	0.08	0.51	0.90	0.93	0.12	0.60	0.31	0.55	0.25
	ges	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00
NPS	F	0.29	0.50	0.39	0.00	0.03	0.91	0.71	1.54	2.34	3.91	0.42	0.72	0.20	0.00	0.01	1.21
	p	0.59	0.48	0.53	0.99	0.86	0.34	0.40	0.22	0.13	0.05	0.52	0.40	0.65	0.95	0.94	0.28
	ges	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
BI	F	1.46	0.36	0.63	0.17	0.00	0.09	0.73	0.21	0.03	1.72	0.01	5.06	2.58	1.35	0.31	0.02
	p	0.23	0.55	0.43	0.68	0.95	0.77	0.40	0.65	0.85	0.19	0.94	**0.03**	0.11	0.25	0.58	0.90
	ges	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.01	0.01	0.00	0.00

Cells with value 0.00 are values <0.005. P-values < 0.05 are bolded, and the corresponding generalized eta-squared values are highlighted with colors indicating the effect sizes (unsubstantiated , small , medium , and large).

ACKNOWLEDGMENTS

Anonymized for review.

REFERENCES

[1] A. Adadi and M. Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). In *IEEE Access*, vol. 6. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052

[2] Prilly Putri Adinda and Amalia Suzianti. 2018. Redesign of user interface for e-government application using usability testing method. In *Proceedings of the 4th International Conference on Communication and Information Processing*. 145–149. DOI: https://doi.org/10.1145/3290420.3290433

[3] Icek Ajzen. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes* 50, 2 (1991), 179–211. DOI: https://doi.org/10.1016/0749-5978(91)90020-T

[4] Amazon. 2020. What Is Conversational AI? - Alexa Skills Kit Official Site. Retrieved 21 Nov, 2021 from https://developer.amazon.com/en-US/alexa/alexa-skills-kit/conversational-ai.

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–13. DOI: https://doi.org/10.1145/3290605.3300233

[6] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–23. DOI: https://doi.org/10.1145/3214262

[7] E. Aronson, T. D. Wilson, and M. B. Brewer. 1998. Experimentation in social psychology. In *The handbook of Social Psychology*, D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.). McGraw-Hill, 99–142.
[8] L. N. Yaddanapudi. 2016. The American Statistical Association statement on P-values explained. *J Anaesthesiol Clin Pharmacol.* 32, 4 (2016), 421–423. DOI: 10.4103/0970-9185.194772
[9] Christiane Atzmüller and Peter M. Steiner. 2010. Experimental vignette studies in survey research. *Methodology* 6, 3 (2010), 128–138. DOI: https://doi.org/10.1027/1614-2241/a000014
[10] Katrin Auspurg, Thomas Hinz, and Stefan Liebig. 2009. Complexity, learning effects and plausibility of vignettes in the factorial survey design. *Methoden - Daten - Analysen* 3, 1 (2009), 59–96. DOI: https://doi.org/10.1007/BF01974149
[11] Katrin Auspurg, Thomas Hinz, Stefan Liebig, and Carsten Sauer. 2015. The factorial survey as a method for measuring sensitive issues. In *Proceedings of the Improving Survey Methods: Lessons from Recent Research.* 137–149.
[12] Roger Bakeman. 2005. Recommended effect size statistics for repeated measures designs. *Behavior Research Methods* 37, 3 (2005), 379–384. DOI: https://doi.org/10.3758/BF03192707
[13] Gagan Bansal, Besmira Nushi, Ece Kamar, Dan Weld, Walter Lasecki, and Eric Horvitz. 2019. A case for backward compatibility for human-ai teams. arXiv:1906.01148. Retrieved June 17, 2022 from https://arxiv.org/abs/1906.01148.
[14] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of AI explanations on complementary team performanceIn *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/3411764.3445717
[15] Susan B. Barnes. 2006. A privacy paradox: Social networking in the united states. *First Monday* 11, 9 (2006), 5. DOI: https://doi.org/10.5210/fm.v11i9.1394
[16] Joey Benedek and Trish Miner. 2002. Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association* 2003, 8–12 (2002), 57.
[17] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300.
[18] Enrico Bertini, Silvia Gabrielli, Stephen Kimani, Tiziana Catarci, and Giuseppe Santucci. 2006. Appropriating and assessing heuristics for mobile computing. In *Proceedings of the Working Conference on Advanced Visual Interfaces.* Association for Computing Machinery, New York, NY, 119–126. DOI: https://doi.org/10.1145/1133265.1133291
[19] Jaspreet Bhatia and Travis D. Breaux. 2018. Empirical measurement of perceived privacy risk. *ACM Transactions on Computer-Human Interaction* 25, 6, 1–47. DOI: https://doi.org/10.1145/3267808
[20] Jaspreet Bhatia, Travis D. Breaux, Joel R. Reidenberg, and Thomas B. Norton. 2016. A theory of vagueness and privacy risk perception. In *Proceedings of the 2016 IEEE 24th International Requirements Engineering Conference.* Institute of Electrical and Electronics Engineers Inc., 26–35. DOI: https://doi.org/10.1109/RE.2016.20
[21] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. DOI: https://doi.org/10.1191/1478088706qp063oa
[22] Michael Brown, Tim Coughlan, Jesse Blum, Glyn Lawson, Robert Houghton, Richard Mortier, Murray Goulden, and Unna Arunachalam. 2015. Tailored scenarios: A low-cost online method to elicit perceptions of home technologies using participant-specific contextual information. *Interacting with Computers* 27, 1 (2015), 60–71. DOI: https://doi.org/10.1093/iwc/iwu028
[23] Margaret Burnett, Simone Stumpf, Jamie Macbeth, Stephann Makri, Laura Beckwith, Irwin Kwan, Anicia Peters, and William Jernigan. 2016. GenderMag: A method for evaluating software's gender inclusiveness. *Interacting with Computers* 28, 6 (2016), 760–787. DOI: https://doi.org/10.1093/iwc/iwv046
[24] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the 2015 International Conference on Healthcare Informatics.* 160–169. DOI: https://doi.org/10.1109/ICHI.2015.26
[25] Pascale Carayon, Peter Hoonakker, Ann Schoofs Hundt, Megan Salwei, Douglas Wiegmann, Roger L. Brown, Peter Kleinschmidt, Clair Novak, Michael Pulia, Yudi Wang, Emily Wirkus, and Brian Patterson. 2020. Application of human factors to improve usability of clinical decision support for diagnostic decision-making: A scenario-based simulation study. *BMJ Quality and Safety* 29, 4 (2020), 329–340. DOI: https://doi.org/10.1136/bmjqs-2019-009857
[26] Nancy Carter, Denise Bryant-Lukosius, Alba Dicenso, Jennifer Blythe, and Alan J. Neville. 2014. The use of triangulation in qualitative research. *Oncology Nursing Forum* 41, 5 (2014), 545–547. DOI: https://doi.org/10.1188/14.ONF.545-547
[27] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Routledge. DOI: https://doi.org/10.4324/9780203771587
[28] Sasha Costanza-Chock. 2020. *Design Justice: Community-led Practices to Build the Worlds We Need.* MIT Press.
[29] Tim Coughlan, Richard Mortier, Michael Brown, Robert J. Houghton, Glyn Lawson, and Murray Goulden. 2013. Tailored scenarios: A low-cost online method to elicit perceptions on designs using real relationships. In *Proceedings of the Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, New York, 343–348. DOI: https://doi.org/10.1145/2468356.2468417

[30] Fred D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13, 3 (1989), 319–340.
[31] Nicola Davis. 2021. From oximeters to AI, where bias in medical devices may lurk. *The Guardian* (2021). Retrieved from https://www.theguardian.com/society/2021/nov/21/from-oximeters-to-ai-where-bias-in-medical-devices-may-lurk.
[32] Heather Desurvire and Charlotte Wiberg. 2009. Game usability heuristics (PLAY) for evaluating and designing better games: The next iteration. In *Proceedings of the Online Communities and Social Computing*, A. Ant Ozok and Panayiotis Zaphiris (Eds.), Springer, Berlin, 557–566.
[33] Julie S. Downs, Mandy Holbrook, and Lorrie Faith Cranor. 2007. Behavioral response to phishing risk. In *Proceedings of the ACM International Conference Proceeding Series*. ACM Press, New York, New York, 37–44. DOI: https://doi.org/10.1145/1299015.1299019
[34] Facebook. [n.d.]. General Best Practices - Messenger Platform. Retrieved June 17, 2022 from https://developers.facebook.com/docs/messenger-platform/introduction/general-best-practices.
[35] Celina Friemel, Stefan Morana, Jella Pfeiffer, and Alexander Maedche. 2018. On the role of users' cognitive-affective states for user assistance invocation. In *Proceedings of the Information Systems and Neuroscience*. Springer, 37–46.
[36] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660. DOI: https://doi.org/10.5465/annals.2018.0057
[37] Google. [n.d.]. People + AI Guidebook | PAIR. Retrieved June 17, 2022 from https://pair.withgoogle.com/guidebook.
[38] Nina Grgic-Hlaca, Christoph Engel, and Krishna P. Gummadi. 2019. Human decision making with machine advice: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). DOI: https://doi.org/10.1145/3359280
[39] David Gunning. 2019. DARPA's explainable artificial intelligence (XAI) program. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI'19)*. Association for Computing Machinery, New York, NY, USA, ii. https://doi.org/10.1145/3301275.3308446
[40] Anders Gustafsson, Michael D. Johnson, and Inger Roos. 2005. The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of Marketing* 69, 4 (2005), 210–218. DOI: https://doi.org/10.1509/jmkg.2005.69.4.210
[41] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2018. Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital Journalism* 6, 3 (2018), 330–343.
[42] Matthew K. Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for natural language failures with the AI playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY. DOI: https://doi.org/10.1145/3411764.3445735
[43] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 159–166. DOI: https://doi.org/10.1145/302979.303030
[44] Eric Horvitz, Johnson Apacible, and Muru Subramani. 2005. Balancing awareness and interruption: Investigation of notification deferral policies. In *Proceedings of the International Conference on User Modeling*. Springer, 433–437.
[45] Edward Cutrell, Mary Czerwinski, and Eric Horvitz. 2001. Notification, disruption, and memory: Effects of messaging interruptions on memory and performance. In *Proceedings of the Human-Computer Interaction: INTERACT*, Vol. 1. 263.
[46] Roberto Hoyle, Luke Stark, Qatrunnada Ismail, David Crandall, Apu Kapadia, and Denise Anthony. 2020. Privacy Norms and Preferences for Photos Posted Online. *ACM Trans. Comput.-Hum. Interact.* 27, 4, Article 30 (August 2020). 27 pages. https://doi.org/10.1145/3380960
[47] R. Hughes and M. Huby. 2004. The construction and interpretation of vignettes in social research. *Social Work and Social Sciences Review* 11, 1 (2004), 36–51. https://doi.org/10.1921/17466105.11.1.36
[48] Mathias Humbert, Benjamin Trubert, and Kévin Huguenin. 2019. A survey on interdependent privacy. *ACM Computing Surveys* 52, 6 (2019), 1–40. DOI: https://doi.org/10.1145/3360498
[49] IBM. [n.d.]. IBM Design for AI. Retrieved June 17, 2022 from https://www.ibm.com/design/ai/.
[50] Sarah Janböcke, Diana Löffler, and Marc Hassenzahl. 2020. Using experimental vignettes to study early-stage automation adoption. In arXiv:2004.07032. Retrieved June 17, 2022 from https://arxiv.org/abs/2004.07032.
[51] Yong Gu Ji, Jun Ho Park, Cheol Lee, and Myung Hwan Yun. 2006. A usability checklist for the usability evaluation of mobile phone user interface. *International Journal of Human-Computer Interaction* 20, 3 (2006), 207–231. DOI: https://doi.org/10.1207/s15327590ijhc2003_3
[52] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. DOI: https://doi.org/10.1207/S15327566IJCE0401_04
[53] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399. DOI: https://doi.org/10.1038/s42256-019-0088-2

[54] Jeff Johnson. 2007. *GUI Bloopers 2.0: Common User Interface Design Don'ts and Dos*. Elsevier.
[55] Sabrina Karwatzki, Olga Dytynko, Manuel Trenz, and Daniel Veit. 2017. Beyond the personalization-privacy paradox: Privacy valuation, transparency features, and service personalization. *Journal of Management Information Systems* 34, 2 (2017), 369–400. DOI: https://doi.org/10.1080/07421222.2017.1334467
[56] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–14. DOI: https://doi.org/10.1145/3313831.3376219
[57] Don Kemper, Larry Davis, Cali Fidopiastis, and Denise Nicholson. 2007. Foundations for creating a distributed adaptive user interface. In *Proceedings of the International Conference on Foundations of Augmented Cognition*. Springer, 251–257.
[58] Sang Min Ko, Won Suk Chang, and Yong Gu Ji. 2013. Usability principles for augmented reality applications in a smartphone environment. *International Journal of Human-Computer Interaction* 29, 8 (2013), 501–515. DOI: https://doi.org/10.1080/10447318.2012.722466
[59] Yasumasa Kobayashi, Takahiro Tanaka, Kazuaki Aoki, and Kinya Fujita. 2015. Automatic delivery timing control of incoming email based on user interruptibility. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1779–1784. DOI: https://doi.org/10.1145/2702613.2732825
[60] Alfred Kobsa. 2007. Privacy-enhanced personalization. *Communications of the ACM* 50, 8 (2007), 24–33. DOI: https://doi.org/10.1145/1278201.1278202
[61] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–14. DOI: https://doi.org/10.1145/3290605.3300641
[62] Sucheta V. Kolekar, Sriram G. Sanjeevi, and D. S. Bormane. 2010. Learning style recognition using artificial neural network for adaptive user interface in e-learning. In *Proceedings of the 2010 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 1–5.
[63] Hannu Korhonen and Elina M. I. Koivisto. 2006. Playability heuristics for mobile games. In *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services*. Association for Computing Machinery, New York, NY, 9–16. DOI: https://doi.org/10.1145/1152215.1152218
[64] Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2016. Watch commander: A gesture-based invocation system for rectangular smartwatches using B2B-swipe. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 37–39.
[65] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2019. Automation transparency: Implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics* 62, 3 (2019), 345–360.
[66] Hosub Lee and Alfred Kobsa. 2019. Confident privacy decision-making in IoT environments. *ACM Transactions on Computer-Human Interaction* 27, 1 (2019). DOI: https://doi.org/10.1145/3364223
[67] Barbara Leporini and Fabio Paternò. 2008. Applying web usability criteria for vision-impaired users: Does it really improve task performance? *International Journal of Human-Computer Interaction* 24, 1 (2008), 17–47. DOI: https://doi.org/10.1080/10447310701771472
[68] Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, intimacy and self-disclosure in social media. In *Proceedings of the Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 3857–3869. DOI: https://doi.org/10.1145/2858036.2858414
[69] Olibário Machado Neto and Maria da Graça Pimentel. 2013. Heuristics for the assessment of interfaces of mobile devices. In *Proceedings of the 19th Brazilian Symposium on Multimedia and the Web*. Association for Computing Machinery, New York, NY, 93–96. DOI: https://doi.org/10.1145/2526188.2526237
[70] Microsoft. 2017. Principles of bot design - Bot Service | Microsoft Docs. Retrieved June 17, 2022 from https://docs.microsoft.com/en-us/azure/bot-service/bot-service-design-principles?view=azure-bot-service-3.0 Retrieved from https://docs.microsoft.com/en-us/azure/bot-service/bot-service-design-principles?view=azure-bot-service-4.0.
[71] Becca Monaghan. 2021. An artificial intelligence bot has turned racist because of humans. Retrieved June 17, 2022 from https://www.msn.com/en-gb/money/technology/an-artificial-intelligence-bot-has-turned-racist-because-of-humans/ar-AAQmDDl.
[72] John Morkes and Jakob Nielsen. 2021. Concise, SCANNABLE, and Objective: How to Write for the Web. Retrieved June 17, 2022 from https://www.nngroup.com/articles/concise-scannable-and-objective-how-to-write-for-the-web/.

[73] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, 607–617. DOI: https://doi.org/10.1145/3351095.3372850

[74] Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv:1902.01876. Retrieved June 17, 2022 from https://arxiv.org/abs/1902.01876.

[75] Roberto Munoz, Thiago Barcelos, and Virginia Chalegre. 2012. Defining and validating virtual worlds usability heuristics. In *Proceedings of the International Conference of the Chilean Computer Science Society*. 171–178. DOI: https://doi.org/10.1109/SCCC.2011.23

[76] Bence Mózer. 2017. Release Notes And Other Great Ways To Communicate Product Updates. Retrieved June 17, 2022 from https://uxstudioteam.com/ux-blog/communicate-product-updates/.

[77] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*. Association for Computing Machinery, New York, NY, 677–686. DOI: https://doi.org/10.1145/2566486.2568012

[78] Jakob Nielsen. 1994. 10 Heuristics for User Interface Design: Article by Jakob Nielsen. Retrieved June 17, 2022 from https://www.nngroup.com/articles/ten-usability-heuristics/.

[79] Jakob Nielsen. 1994. Enhancing the explanatory power of usability heuristics. In *Proceedings of the Conference on Human Factors in Computing Systemss*. ACM, New York, New York, 152–158. DOI: https://doi.org/10.1145/259963.260333

[80] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, New York, 249–256. DOI: https://doi.org/10.1145/97243.97281

[81] Donald A. Norman. 1983. Design principles for human-computer interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–10. DOI: https://doi.org/10.1145/800045.801571

[82] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, K. E. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernández, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab. 2020. Bias in data-driven artificial intelligence systems-an introductory survey. Arxiv, abs/2001.09762.

[83] Arfika Nurhudatiana and Jae Young Seo. 2020. An mHealth application redesign based on nielsen's usability heuristics. In *Proceedings of the 2020 The 6th International Conference on E-Business and Applications*. ACM, New York, NY, 85–89. DOI: https://doi.org/10.1145/3387263.3387267

[84] Susan M. O'Connor, John B. Davies, Dorothy D. Heffernan, and Robert van Eijk. 2003. An alternative method for predicting attrition from an alcohol treatment programme. *Alcohol and Alcoholism* 38, 6 (2003), 568–573. DOI: https://doi.org/10.1093/alcalc/agg112

[85] S. Olejnik and James Algina. 2003. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods* 8, 4 (2003), 434–47.

[86] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. Arxiv, abs/1907.12652.

[87] Eli Pariser. 2011. *The Filter Bubble: What the Internet is Hiding from You*. Penguin UK.

[88] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. 2013. Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. In *Proceedings of the IFIP Advances in Information and Communication Technology*. Springer, New York LLC, 366–378. DOI: https://doi.org/10.1007/978-3-642-39218-4_27

[89] Evelina Patsoule and Panayiotis Koutsabasis. 2014. Redesigning websites for older adults: A case study. *Behaviour and Information Technology* 33, 6 (2014), 561–573. DOI: https://doi.org/10.1080/0144929X.2013.810777

[90] M. Pattinson, C. Jerram, K. Parsons, A. McCormac, and M. Butavicius. 2011. Managing phishing emails: A scenario-based experiment. In *Proceedings of the 5th International Symposium on Human Aspects of Information Security and Assurance*. 75–85. Retrieved June 17, 2022 from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.663.4023.

[91] Michael Quinn Patton. 1999. Enhancing the quality and credibility of qualitative studies. *Health Serv Res*. 34, 5 Pt 2 (1999), 1189–1208.

[92] David Pinelle, Nelson Wong, and Tadeusz Stach. 2008. Heuristic evaluation for games: Usability principles for video game design. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM Press, New York, New York, 1453–1462. DOI: https://doi.org/10.1145/1357054.1357282

[93] David Pinelle, Nelson Wong, Tadeusz Stach, and Carl Gutwin. 2009. Usability heuristics for networked multiplayer games. In *Proceedings of the 2009 ACM SIGCHI International Conference on Supporting Group Work*. ACM Press, New York, New York, 169–178. DOI: https://doi.org/10.1145/1531674.1531700

[94] F. F. Reichheld and R. Markey. 2011. *The Ultimate Question 2.0: How Net Promoter Companies Thrive in a Customer-driven World*. Harvard Business Press. Retrieved June 17, 2022 from https://books.google.com/books?id=e8jhiYjQrU0C.

[95] Corbin Reno and Erika S. Poole. 2016. It matters if my friends stop smoking: Social support for behavior change in social media. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, New York, NY, 5548–5552. DOI: https://doi.org/10.1145/2858036.2858203

[96] 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*. Association for Computing Machinery, New York, NY, 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778

[97] Alissa L. Russ, Alan J. Zillich, Brittany L. Melton, Scott A. Russell, Siying Chen, Jeffrey R. Spina, Michael Weiner, Elizabette G. Johnson, Joanne K. Daggy, M. Sue McManus, Jason M. Hawsey, Anthony G. Puleo, Bradley N. Doebbeling, and Jason J. Saleem. 2014. Applying human factors principles to alert design increases efficiency and reduces prescribing errors in a scenario-based simulation. *Journal of the American Medical Informatics Association : JAMIA* 21, e2 (2014), e287–e296. DOI: https://doi.org/10.1136/amiajnl-2013-002045

[98] Luiz Henrique A. Salazar, Thaísa Lacerda, Juliane Vargas Nunes, and Christiane Gresse von Wangenheim. 2013. A systematic literature review on usability heuristics for mobile phones. *International Journal of Mobile Human Computer Interaction* 5, 2 (2013), 50–61.

[99] SAP. 2020. Designing Intelligent Systems | SAP Fiori Design Guidelines. Retrieved June 17, 2022 from https://experience.sap.com/fiori-design-web/designing-intelligent-systems/.

[100] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. 2015. Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 539–546.

[101] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM Press, New York, New York, 373–382. DOI: https://doi.org/10.1145/1753326.1753383

[102] T. B. Sheridan. 1989. Trustworthiness of command and control systems. In *Proceedings of the Analysis, Design and Evaluation of Man-Machine Systems*. J. RANTA (Ed.), Pergamon, Amsterdam, 427–431. DOI: https://doi.org/10.1016/B978-0-08-036226-7.50076-4

[103] Ben Shneiderman. 1987. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley, Reading, Mass.

[104] Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (6th ed.). Pearson.

[105] Ben Shneiderman, Catherine Plaisant, Maxine S. Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2016. *Designing the User Interface: Strategies for Effective Human-computer Interaction*. Pearson.

[106] Debbie Stone, Caroline Jarrett, Mark Woodroffe, and Shailey Minocha. 2005. *User Interface Design and Evaluation*. Elsevier.

[107] Kimberly Stowers, Nicholas Kasdaglis, Michael Rupp, Jessie Chen, Daniel Barber, and Michael Barnes. 2017. Insights into human-agent teaming: Intelligent agent transparency and uncertainty. In *Proceedings of the Advances in Human Factors in Robots and Unmanned Systems*. Springer, 149–160.

[108] Madiha Tabassum, Jess Kropczynski, Pamela Wisniewski, and Heather Richter Lipford. 2020. Smart home beyond the home: A case for community-based access control. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. ACM, New York, NY, 1–12. DOI: https://doi.org/10.1145/3313831.3376255

[109] M. Iftekhar Tanveer, Ru Zhao, Kezhen Chen, Zoe Tiet, and Mohammed Ehsan Hoque. 2016. Automanner: An automated interface for making public speakers aware of their mannerisms. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 385–396.

[110] Patrick Tchankue, Janet Wesson, and Dieter Vogts. 2011. The impact of an adaptive user interface on reducing driver distraction. In *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. Association for Computing Machinery, New York, NY, 87–94. DOI: https://doi.org/10.1145/2381416.2381430

[111] Ha Trinh, Koji Yatani, and Darren Edge. 2014. PitchPerfect: Integrated rehearsal environment for structured presentation preparation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1571–1580.

[112] Mr R. Valanarasu. 2019. Smart and secure IoT and AI integration framework for hospital environment. *Journal of ISMAC* 1, 03 (2019), 172–179.

[113] Mihaela Vorvoreanu, Lingyi Zhang, Yun Han Huang, Claudia Hilderbrand, Zoe Steine-Hanson, and Margaret Burnett. 2019. From gender biases to gender-inclusive design: An empirical investigation. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM Press, New York, New York. DOI: https://doi.org/10.1145/3290605.3300283

[114] Lisa Wallander. 2009. 25 years of factorial surveys in sociology: A review. *Social Science Research* 38, 3 (2009), 505–520. DOI: https://doi.org/10.1016/j.ssresearch.2009.03.004

[115] Kelly D. Wason, Michael J. Polonsky, and Michael R. Hyman. 2002. Designing vignette studies in marketing. *Australasian Marketing Journal* 10, 3 (2002), 41–58. DOI: https://doi.org/10.1016/s1441-3582(02)70157-2

[116] Jake Weidman, William Aurite, and Jens Grossklags. 2019. On sharing intentions, and personal and interdependent privacy considerations for genetic data: A vignette study. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16, 4 (2019), 1349–1361. DOI: https://doi.org/10.1109/TCBB.2018.2854785

[117] Kyle Wiggers. 2021. AI Weekly: Recognition of bias in AI continues to grow. Retrieved June 17, 2022 from https://venturebeat.com/2021/12/03/ai-weekly-recognition-of-bias-in-ai-continues-to-grow/.

[118] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–13. DOI: https://doi.org/10.1145/3313831.3376301

[119] Baobao Zhang and Allan Dafoe. 2019. Artificial intelligence: American attitudes and trends. *SSRN Electronic Journal* (2019). DOI: https://doi.org/10.2139/ssrn.3312874

[120] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, 295–305. DOI: https://doi.org/10.1145/3351095.3372852

Received 30 June 2021; revised 11 January 2022; accepted 13 January 2022