Guidelines for Human-AI Interaction

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

Microsoft
Redmond, WA, USA
{samershi, mivorvor, adamfo, benushi, pennycol, jinsuh, shamsi, pauben, kori, teevan, ruthkg, horvitz}
@microsoft.com

[†]Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA, USA
weld@cs.washington.edu

ABSTRACT

Advances in artificial intelligence (AI) frame opportunities and challenges for user interface design. Principles for human-AI interaction have been discussed in the human-computer interaction community for over two decades, but more study and innovation are needed in light of advances in AI and the growing uses of AI technologies in human-facing applications. We propose 18 generally applicable design guidelines for human-AI interaction. These guidelines are validated through multiple rounds of evaluation including a user study with 49 design practitioners who tested the guidelines against 20 popular AI-infused products. The results verify the relevance of the guidelines over a spectrum of interaction scenarios and reveal gaps in our knowledge, highlighting opportunities for further research. Based on the evaluations, we believe the set of design guidelines can serve as a resource to practitioners working on the design of applications and features that harness AI technologies, and to researchers interested in the further development of guidelines for human-AI interaction design.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Artificial intelligence.*

[*]Work done as a visiting researcher at Microsoft Research.

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

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300233

KEYWORDS

Human-AI interaction; AI-infused systems; design guidelines

ACM Reference Format:
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 *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300233

1 INTRODUCTION

Advances in artificial intelligence (AI) are enabling developers to integrate a variety of AI capabilities into user-facing systems. For example, increases in the accuracy of pattern recognition have created opportunities and pressure to integrate speech recognition, translation, object recognition, and face recognition into applications. However, as automated inferences are typically performed under uncertainty, often producing false positives and false negatives, AI-infused systems may demonstrate unpredictable behaviors that can be disruptive, confusing, offensive, and even dangerous. While some AI technologies are deployed in explicit, interactive uses, other advances are employed behind the scenes in proactive services acting on behalf of users such as automatically filtering content based on inferred relevance or importance. While such attempts at personalization may be delightful when aligned with users' preferences, automated filtering and routing can be the source of costly information hiding and actions at odds with user goals and expectations.

AI-infused systems[1] can violate established usability guidelines of traditional user interface design (e.g., [31, 32]). For example, the principle of consistency advocates for minimizing unexpected changes with a consistent interface appearance and predictable behaviors. However, many AI components are inherently inconsistent due to poorly understood,

[1]In this paper we use *AI-infused systems* to refer to systems that have features harnessing AI capabilities that are directly exposed to the end user.

probabilistic behaviors based on nuances of tasks and settings, and because they change via learning over time. AI-infused systems may react differently depending on lighting or noise conditions that are not recognized as distinct to end users. Systems may respond differently to the same text input over time (e.g., autocompletion systems suggesting different words after language model updates) or behave differently from one user to the next (e.g., search engines returning different results due to personalization). Inconsistent and unpredictable behaviors can confuse users, erode their confidence, and lead to abandonment of AI technology [7, 22]. Errors are common in AI-infused systems, rendering it difficult to reliably achieve the principle of error prevention. This has contributed to the large and growing body of work on AI explanations and interpretability to support human verification of proposed actions aimed at reducing the likelihood of unwarranted or potentially dangerous actions and costly outcomes (e.g., [14, 21, 23, 36, 38, 44]).

For over 20 years, the human-computer interaction (HCI) community has proposed principles, guidelines, and strategies for designing user interfaces and interaction for applications employing AI inferences (e.g., [16, 17, 33]). However, the variability of AI designs (e.g., varying capabilities and interaction styles of commercial conversational agents impacting user engagement and usability [26]) and high-profile reports of failures, ranging from humorous and embarrassing (e.g., autocompletion errors [8]) to more serious harm when users cannot effectively understand or control an AI system (e.g., collaboration with semi-autonomous cars [41]), show that designers and developers continue to struggle with creating intuitive and effective AI-infused systems. Ongoing advances in AI technologies will generate a stream of challenges and opportunities for the HCI community. While such developments will require ongoing studies and vigilance, we also see value in developing reusable guidelines that can be shared, refined, and debated by the HCI community. The development and use of such shared guidelines can help with the design and evaluation of AI-infused systems that people can understand, trust, and can engage with effectively.

In this work, we synthesize over 20 years of learning in AI design into a small set of generally applicable design guidelines for human-AI interaction. Specifically, our contributions are:

- A codification of over 150 AI-related design recommendations collected from academic and industry sources into a set of 18 generally applicable design guidelines for human-AI interaction (see Table 1).
- A systematic validation of the 18 guidelines through multiple rounds of iteration and testing.

We hope these guidelines, along with our examination of their applications in AI-infused systems, will serve as a resource for designers working with AI and will facilitate future research into the refinement and development of principles for human-AI interaction.

2 RELATED WORK

For over 20 years, the academic community has proposed numerous guidelines and recommendations for how to design for effective human interaction with AI-infused systems. For example, Norman [33] and Höök [16] both recommended building in safeguards like verification steps or controlling levels of autonomy to help prevent unwanted adaptations or actions from intelligent systems. Others recommended managing expectations so as not to mislead or frustrate users during interaction with unpredictable adaptive agents [16, 20, 33]. Horvitz's formative paper on mixed-initiative systems [17] proposed principles for balancing autonomous actions with direct manipulation constructs, such as supporting user-driven invocation of intelligent services, scoping actions based on inferred goals and confidences, and inferring ideal action in light of costs, benefits, and uncertainties. The latter guideline was operationalized via the introduction of a decision-theoretic methodology to guide decisions about acting on AI inferences versus waiting for user input, based on consideration of expected costs and benefits of performing AI automation under uncertainty.

In some cases, specific AI design recommendations have received considerable attention within the academic community. For example, a large body of work exists and continues to grow around how to increase transparency or explain the behaviors of AI systems (e.g., [14, 21, 23, 36, 38, 44], to name a few). Similarly, when and how to automatically adapt or personalize interfaces has been studied extensively in a variety of scenarios (e.g., [9, 11–13]).

Others in the community have studied how to design for specific human-AI interaction scenarios. For example, researchers have been studying how to effectively interact with intelligent agents for many years (e.g., [18, 33]). This scenario has also had a recent resurgence of interest given advances in natural language processing and embedded devices driving the proliferation of conversational agents [26, 29, 35]. Similarly, researchers have for decades studied human interaction with intelligent context-aware computing systems including how to design for understandability and control of the underlying sensing systems [3, 23] and how to support ambiguity resolution [10]. Recent advances in sensing technologies and the widespread availability of commercial fitness and activity trackers have continued to drive interaction research in these domains [37, 45].

Despite all of this work, the ongoing stream of articles and editorials in the public domain about how to design in the face of AI (e.g., [2, 24, 25, 39, 42]) suggests designers need more guidance. This may be partly due to design suggestions

		AI Design Guidelines	Example Applications of Guidelines
Initially	G1	**Make clear what the system can do.** Help the user understand what the AI system is capable of doing.	[Activity Trackers, Product #1] "Displays all the metrics that it tracks and explains how. Metrics include movement metrics such as steps, distance traveled, length of time exercised, and all-day calorie burn, for a day."
	G2	**Make clear how well the system can do what it can do.** Help the user understand how often the AI system may make mistakes.	[Music Recommenders, Product #1] "A little bit of hedging language: 'we think you'll like'."
During interaction	G3	**Time services based on context.** Time when to act or interrupt based on the user's current task and environment.	[Navigation, Product #1] "In my experience using the app, it seems to provide timely route guidance. Because the map updates regularly with your actual location, the guidance is timely."
	G4	**Show contextually relevant information.** Display information relevant to the user's current task and environment.	[Web Search, Product #2] "Searching a movie title returns show times in near my location for today's date"
	G5	**Match relevant social norms.** Ensure the experience is delivered in a way that users would expect, given their social and cultural context.	[Voice Assistants, Product #1] "[The assistant] uses a semi-formal voice to talk to you - spells out "okay" and asks further questions."
	G6	**Mitigate social biases.** Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.	[Autocomplete, Product #2] "The autocomplete feature clearly suggests both genders [him, her] without any bias while suggesting the text to complete."
When wrong	G7	**Support efficient invocation.** Make it easy to invoke or request the AI system's services when needed.	[Voice Assistants, Product #1] "I can say [wake command] to initiate."
	G8	**Support efficient dismissal.** Make it easy to dismiss or ignore undesired AI system services.	[E-commerce, Product #2] "Feature is unobtrusive, below the fold, and easy to scroll past...Easy to ignore."
	G9	**Support efficient correction.** Make it easy to edit, refine, or recover when the AI system is wrong.	[Voice Assistants, Product #2] "Once my request for a reminder was processed I saw the ability to edit my reminder in the UI that was displayed. Small text underneath stated 'Tap to Edit' with a chevron indicating something would happen if I selected this text."
	G10	**Scope services when in doubt.** Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.	[Autocomplete, Product #1] "It usually provides 3-4 suggestions instead of directly auto completing it for you"
	G11	**Make clear why the system did what it did.** Enable the user to access an explanation of why the AI system behaved as it did.	[Navigation, Product #2] "The route chosen by the app was made based on the Fastest Route, which is shown in the subtext."
Over time	G12	**Remember recent interactions.** Maintain short term memory and allow the user to make efficient references to that memory.	[Web Search, Product #1] "[The search engine] remembers the context of certain queries, with certain phrasing, so that it can continue the thread of the search (e.g., 'who is he married to' after a search that surfaces Benjamin Bratt)"
	G13	**Learn from user behavior.** Personalize the user's experience by learning from their actions over time.	[Music Recommenders, Product #2] "I think this is applied because every action to add a song to the list triggers new recommendations."
	G14	**Update and adapt cautiously.** Limit disruptive changes when updating and adapting the AI system's behaviors.	[Music Recommenders, Product #2] "Once we select a song they update the immediate song list below but keeps the above one constant."
	G15	**Encourage granular feedback.** Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.	[Email, Product #1] "The user can directly mark something as important, when the AI hadn't marked it as that previously."
	G16	**Convey the consequences of user actions.** Immediately update or convey how user actions will impact future behaviors of the AI system.	[Social Networks, Product #2] "[The product] communicates that hiding an Ad will adjust the relevance of future ads."
	G17	**Provide global controls.** Allow the user to globally customize what the AI system monitors and how it behaves.	[Photo Organizers, Product #1] "[The product] allows users to turn on your location history so the AI can group photos by where you have been."
	G18	**Notify users about changes.** Inform the user when the AI system adds or updates its capabilities.	[Navigation, Product #2] "[The product] does provide small in-app teaching callouts for important new features. New features that require my explicit attention are pop-ups."

Table 1: Our 18 human-AI interaction design guidelines, roughly categorized by when they likely are to be applied during interaction with users, along with illustrative applications (rated as "clearly applied" by participants) across products tested by participants in our user study.

being scattered throughout different academic circles and venues, making them difficult to find (e.g., there is relevant work in a wide variety of venues including AAAI, UbiComp, RecSys, SIGIR, HRI, KDD). Moreover, potential design suggestions for AI are often not presented explicitly as such. In many cases, researchers identify usability issues with AI systems and suggest possible solutions in the discussion or future work sections of their academic papers. For example, Lugar and Sellen [26] identify variability in user expectations of conversational agents as causing usability issues and propose setting realistic expectations as a possible solution in their discussion. Similarly, Lee et al [22] studied automatic changes to search result lists during user interaction and suggested caution in updating those lists to balance stability with presenting new content to users. While these proposed solutions could be generalized into principles for designers, not presenting them as such makes them difficult to discover.

It can also be difficult to understand if and how design guidance stemming from one community or interaction scenario extends to others. For example, Bunt et al. [4] showed that, while explanations of AI behaviors have shown promise in complex and high-risk scenarios such as sensor-based ubiquitous computing systems or decision-support systems for medical or financial domains, they may be less important for relatively low-cost scenarios such as search and music or movie recommenders.

In this work we 1) synthesize a unified set of design guidelines from a variety of communities and sources and 2) systematically examine those guidelines in a variety of AI-infused systems to validate their applicability and relevance. The closest to our work is Horvitz's set of principles for mixed-initiative systems [17], noting that 8 of our 18 guidelines map to principles outlined in that work. We celebrate its 20-year anniversary by reflecting on learnings from the community since its publication. Moreover, recent work has warned that the lack of rigorous validations of proposed design heuristics in specific domains makes it difficult to gauge the utility of those heuristics [15]. We developed the guidelines shown in Table 1 using a four-phase process. In Phase 1 we consolidated more than 150 design recommendations from multiple sources into a set of 20 guidelines. In Phase 2 we conducted an internal modified heuristic evaluation of the guidelines, revising the set down to 18. Phase 3 consisted of a user study in which 49 participants used heuristic evaluation to assess the guidelines' relevance and clarity. Based on their feedback, we rephrased some of the guidelines to improve clarity and, in Phase 4, conducted an expert evaluation of the revisions to validate the final set.

3 PHASE 1: CONSOLIDATING GUIDELINES

We gathered AI design recommendations from three sources:

- A review of AI products and guidelines originating from industry. We collected guidelines asserted internally in our company and externally, and grouped them into themes; we audited a sample of AI products within and outside our company against the themes; and cross-referenced themes with internal customer feedback (reviews and bugs reported about our company's AI products).
- Recent public articles and editorials about AI design (e.g., [2, 24, 25, 39]).
- Relevant scholarly papers about AI design (see Related Work section).

While we drew AI design guidelines from the academic literature, the list we captured may not be exhaustive because, as discussed in Related Work, potential design guidelines are often not presented explicitly as such, making them difficult to search for via terms or combinations of terms such as "AI", "machine learning", "design", "principle" or "guideline". Further, as the field is evolving rapidly, we found the most up to date guidance about AI design in industry sources via articles published in the public domain.

From these sources, we obtained 168 potential AI design guidelines. Three members of our team conducted an asynchronous affinity diagramming process, clustering the guidelines into related concepts. This resulted in 35 concepts which we then filtered by removing concepts we deemed to be either too vague to design for directly (e.g., "build trust"), too specific to a particular AI scenario (e.g., "establish that the bot is not human"), or not AI specific (e.g., "display output effectively"). Filtering reduced our set of concepts to 20, each of which we then summarized in a sentence or phrase, forming our first iteration of the guidelines. We organized the guidelines into four top-level categories based on when during the user's interaction they applied: "Initially" (Guideline 1 & Guideline 2), "During interaction" (Guideline 3 - Guideline 6), "When wrong" (Guideline 7 - Guideline 11), and "Over time" (Guideline 12 - Guideline 18). Next, we tested the guidelines via a modified heuristic evaluation.

4 PHASE 2: MODIFIED HEURISTIC EVALUATION

We conducted an evaluation to test and iterate on the initial set of 20 AI design guidelines. We modeled our study after a heuristic evaluation [31], a common discount usability testing method where evaluators examine an interface for violations of a given set of usability guidelines. As the primary goal was to evaluate our design guidelines rather than to evaluate an interface, we modified the heuristic evaluation by asking evaluators to attempt to identify both applications and violations of the proposed guidelines in an interface and to reflect on the guidelines themselves during the evaluation.

Eleven members of our team participated in this evaluation. Team members selected AI-infused products or features

of their choice and then looked for applications or violations of our initial set of design guidelines over a one-hour period. In total, we inspected 13 AI-infused products or features including: two different email products with a feature for filtering unimportant emails, a navigation system, an e-commerce website with product recommendations, two photo organization products, a design assistance feature in a productivity software, a research assistance feature in a productivity application, a social network news feed feature, a web search service, and an image search service. These products were different from the products used in Phase 3.

After the modified heuristic evaluation, we reviewed the findings and reflections about each guideline and discussed issues and revision strategies for conflicting interpretations and ambiguities. For example, our initial phrasing of Guideline 9 ("allow efficient correction") and Guideline 17 ("allow coarse controls") caused several evaluators to confuse instance-level corrections with global-level settings (several evaluators identified adjusting settings as applications of Guideline 9 rather than Guideline 17). We subsequently rephrased Guideline 17 to include the term "global".

We also identified opportunities for merging related or redundant guidelines. For example, the initial set included "informing the user when to take control" and "fallback to a human where appropriate". Our evaluations found few applications of these guidelines, and we determined both of them to be instances of Guideline 10 (initially phrased "scope services when uncertain") and therefore removed them as distinct guidelines. Similarly, applications of "enable users to change privacy permissions" and "allow private mode" were deemed as instances of Guideline 17 (initially phrased "allow coarse controls") and were merged with that guideline.

We also decided to remove some guidelines that resulted in few or no applications during our evaluations. For example, neither the guideline to "explore vs. exploit in moderation" nor to "be especially conservative in the beginning" resulted in any identifiable usage across the products or features we examined. While these guidelines are important at the AI modeling level, they appeared to be difficult to observe or design for in an interface.

After these sessions we reformulated the remaining guidelines to follow a consistent format and to clarify issues identified by evaluators. Specifically, we proposed that each guideline adhere to the following criteria:

- It should be written as a rule of action, containing about 3-10 words and starting with a verb.
- It should be accompanied by a one-sentence description that qualifies or clarifies any potential ambiguities.
- It should not contain conjunctions so that designers can clearly validate whether it is applied or violated in an interface.

Removing conjunctions meant splitting some guidelines. For example, an initial guideline to "allow efficient invocation, correction, and dismissal" became three (to "support efficient invocation," "support efficient dismissal," and "support efficient correction," Guidelines 7-9).

Phase 2 produced a set of 18 guidelines that closely match the guidelines in Table 1. In the following sections we describe a user study that tested these 18 guidelines and a subsequent expert validation of the guidelines that we slightly rephrased after the user study (resulting in the final proposed set shown in Table 1).

5 PHASE 3: USER STUDY

We conducted a user study with 49 HCI practitioners to 1) understand the guidelines' applicability across a variety of products; and 2) get feedback about the guidelines' clarity.

Procedure

We modeled the user study after a heuristic evaluation. We assigned each participant to an AI-driven feature of a product they were familiar with and asked them to find examples (applications and violations) of each guideline.

First, we helped participants become familiar with the guidelines by providing a document that included at least one application and one violation for each. The examples came from a range of AI-infused products and were presented with a 1-2 sentence description and a screenshot where appropriate.

Participants were then instructed to play around with their assigned feature and fill out a form asking a series of questions. For each guideline, the form asked participants to first determine if the guideline "does not apply" to their assigned feature (i.e., irrelevant or out of scope) and, if not, to explain why. If a participant judged that a guideline should apply to their assigned feature and they observed applications or violations, the form requested participants to provide their own examples, and, for each example, a rating of the extent of the application or violation on a 5-point semantic differential scale from "clearly violated" to "clearly applied", along with an explanation of the rating. Participants were incentivized with an additional monetary gratuity to include screenshots to illustrate the examples. After completing the evaluation, participants submitted their examples and ratings and filled in a final questionnaire which asked them to rate each guideline on a 5-point semantic differential scale from "very confusing" to "very clear" and provide any additional comments about the guidelines.

We estimated the study would take approximately one hour to complete based on our modified heuristic evaluation study from Phase 2. Participants were given one week to complete the study on their own time and were compensated with an Amazon Gift Card worth a minimum of $50 and up

to $70 based on the number of applications or violations for which they provided screenshots.

Products

One objective of our study was to determine if and how each of our design guidelines manifests in a variety of AI-infused products. We used a maximum-variance sampling strategy [28] to select popular AI-infused products that covered a wide range of scenarios.

First, we searched online for rankings of top apps, software, and websites in the U.S. for both mobile and desktop devices. This search resulted in 13 lists from sources such as app stores (Apple, Google Play, Windows), and Web traffic rankings [1, 6, 40]. From these lists, we selected the top 10 products in each and then filtered out any that were offensive, game related, or did not currently use AI to drive any of their main end-user facing services (determined by examination of the product or reading supplemental help documentation and news media articles when necessary).

Next, we grouped the remaining products by their primary use case, resulting in 10 categories (e.g., email, e-commerce, social networking). We then selected two products per category based on market share as determined by recent online statistics reports (e.g., [19]). Finally, we selected a prominent AI-driven feature to evaluate per product. In total, we selected 20 products, two of which were from Microsoft.

Many of the products we selected were available on multiple platforms and devices. We attempted to evaluate products on a variety of platforms. Table 2 shows our final list of product categories, features and platforms.

Participants

We recruited participants via HCI and design distribution lists at a large software company. During recruitment we screened for people with at least one year of experience working in or studying HCI (e.g., in roles such as user experience design and user experience research) and familiarity with discount usability testing methods (e.g., heuristic evaluation, cognitive walkthrough). We listed all possible product and platform combinations, and asked respondents to select the options they were familiar with and comfortable evaluating.

We endeavored to assign 2-3 participants to each product according to recommendations for heuristic evaluations. Nielsen [30] recommends 2-3 evaluators when evaluators have both usability experience and familiarity with the product being tested. We also assigned participants so that each product was evaluated by people with a range of experience in discount usability techniques and no product was evaluated by participants with only limited experience. When participants dropped out of the study, we replaced them by assigning new participants from a wait list of eligible respondents, trying to maintain 2-3 evaluators per product.

Product Category	Feature	Participants
E-commerce (Web)	Recommendations	6
Navigation (Mobile)	Route planning	5
Music Recommenders (Mobile)	Recommendations	5
Activity Trackers (Device)	Walking detection and step count	5
Autocomplete (Mobile)	Autocomplete	5
Social Networks (Mobile)	Feed filtering	5
Email (Web)	Importance filtering	5
Voice Assistants (Device)	Creating a reminder with a due date	5
Photo Organizers (Mobile)	Album suggestions	4
Web Search (Web)	Search	4

Table 2: Product categories and features tested in the user study, and the number of participants assigned to each.

In the end, 49 people (29 female, 18 male, 2 preferred not to answer) participated in our study. Participants spanned ages 18-55: 5 were in the age range of 18-24, 24 were in the age range of 25-34, 13 were aged 35-44 and 7 were aged 45-54. Of these participants, 19 were researchers, 12 were designers, 11 were HCI or design interns from various universities worldwide, and the remaining 7 were a mix of engineers, product managers or vendors. The participants' experience working in or studying HCI/UX was as follows: 1-4 years (23 participants), 5-9 years (14 participants), 10-14 years (9 participants), 15-19 years (1 participant), 20+ years (2 participants). Thirty-nine participants self-reported as being highly or very highly experienced at discount usability testing methods while 10 reported as having medium to low levels of experience (we screened out participants with "very low" levels of experience). Participants were from 4 different countries spanning 3 continents. While we recruited participants using internal mailing lists, we took steps to mitigate sampling bias to ensure the results do not exclusively represent one organization's mindset, in addition to including the 11 external participants. Our questionnaires asked participants to rate the extent to which an example is illustrative of a guideline and the clarity of each guideline's wording on Likert scales. These questions are unlikely to be influenced by company values. Moreover, our main sampling criterion was experience with discount usability methods. It is unlikely that the entirety of participants' professional training and experience were internal.

Adjustments and Misinterpretations

To obtain accurate counts of examples of the proposed guidelines across products, we reviewed participant responses for the following cases:

- Duplicate applications or violations of a guideline for any given product (55 instances). For example, two different participants identified the same application of Guideline 1 for an activity tracker: "This guideline is applied in the activity summary view, where it shows a summary of my 'move', 'exercise' and 'stand' metrics." and "Displays all the metrics that it tracks and explains how. Metrics include movement metrics such as steps, distance traveled..." The 55 duplications were removed from the analysis.
- Instances where participants used "does not apply" to indicate that they could not find examples of a guideline rather than to indicate that the guideline is not relevant for the product they were testing, as we intended by this designation (73 instances). For example, "To be quite honest I believe that this would apply, however I can't think of a way to show it." and "Cannot find examples of application or violation.". These 73 instances were also removed from the analysis.
- Instances where participants used "does not apply" to indicate that a guideline was violated (20 instances). For example, "Even in the setting page, there's no option for changing or customizing anything for the autocomplete function." and "[Voice Assistant, Product #1] did not provide additional hints or tips to educate me on what the system is capable of achieving beyond the task I had already asked it to run." We reclassified these instances as violations.
- Instances where participants misinterpreted one guideline for another, discussed further below (56 instances).

We identified these cases using a two-pass process where participant responses were first reviewed by one member of our team to identify each case and then those cases were verified or invalidated by another member of our team. We removed 14 additional instances from our analysis when the two reviewers from our team disagreed on any of these cases.

Results

Our evaluation in this phase focused on two key questions, each addressed in one of the subsections below: 1) Are the guidelines relevant? That is, can we identify examples of each guideline across a variety of products and features? 2) Are the guidelines clear? That is, can participants understand and differentiate among them?

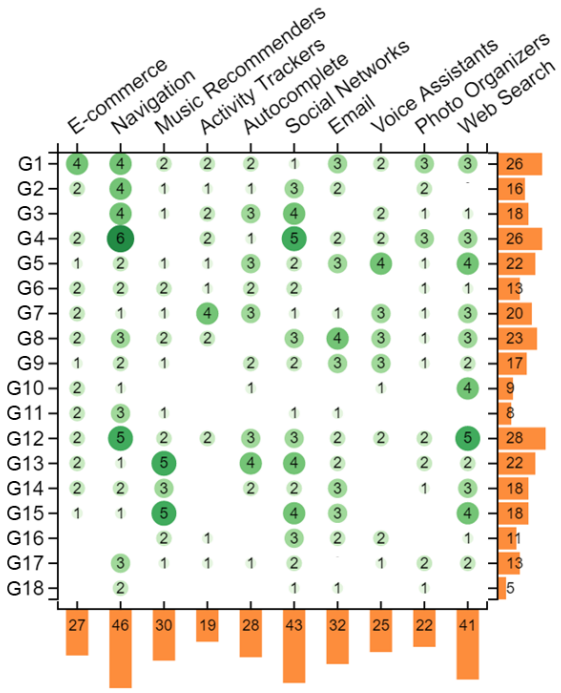
Relevance. Across the 20 products they evaluated, participants identified 785 examples of the 18 guidelines, after the adjustments described earlier: 313 applications, 277 violations, 89 neutrals (rated at the mid-point between "clearly applied" and "clearly violated"), and 106 instances of "does not apply". Figures 1a-1c show the guideline counts per product category for applications, violations, and "does not apply", respectively. Figure 1d shows an aggregate of all applicable ratings, including neutral responses. Finally, Table 1 shows example applications participants provided for each guideline (marked as "clearly applied" by the participant).

In this analysis, we use the following interpretation constructs to better understand results from Figure 1. First, we use the total number of applications and violations as an indicator of the overall evidence of a guideline being relevant (e.g., Guidelines 1, 12, 17). Second, relevant guidelines with a high positive difference between the number of applications and violations are guidelines which are not only relevant but also widely implemented for the set of products in the study (e.g., Guidelines 1, 4, 12). Third, relevant guidelines with a high negative difference between the number of applications and violations are guidelines which, despite their importance, are still not widely implemented (e.g., Guidelines 2, 11, 17). Fourth, we discuss guidelines with the highest numbers of "does not apply" (e.g., Guidelines 3, 5, 6).

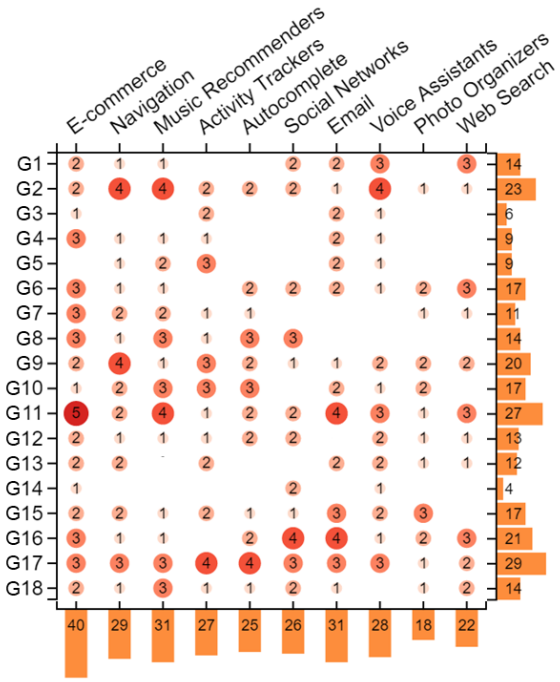
Participants found at least one application or violation of each of our guidelines in each product category we tested, suggesting broad evidence of the guidelines' relevance. While participants were able to identify examples of each guideline in most of the product categories we tested, voice assistants had the largest number of "does not apply" instances reported, while photo organizers, activity trackers and voice assistants had the fewest numbers of total applications or violations. Interestingly, each of these product categories involves a mode of operation or input data type beyond simple graphical user interfaces and text (specifically, interaction over images or sensor data, or voice-based interaction).

No instances of Guideline 10 "Scope services when in doubt" were reported for the two social networks we tested and no instances of Guideline 14 "Update and adapt cautiously" were reported for the two activity trackers. Some participants reported that these guidelines were hard to observe in a single session or without knowledge about the underlying AI algorithms. For example, one participant noted that Guideline 10 was "More difficult to assess unless you have a lengthy period of time with the product - and potential guidance for understanding the behind-the-scenes mechanisms," possibly referring to understanding when the AI system was "in doubt". Similarly, for Guideline 14, one participant said, "It's a bit difficult to assess this in a single session." These guidelines were, however, observed in all other products that participants tested in our study, so such difficulty could be attributed to the guidelines not being applied or being difficult to observe in these particular products.

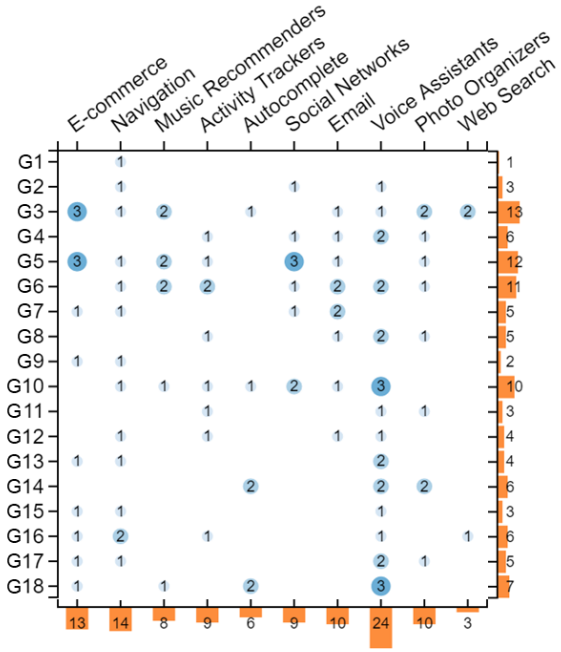
Relevant guidelines that have significantly (at least 40%) more applications than violations are evidence of being widely implemented across products. This is also an indicator that there exist current mechanisms in the intersection of AI



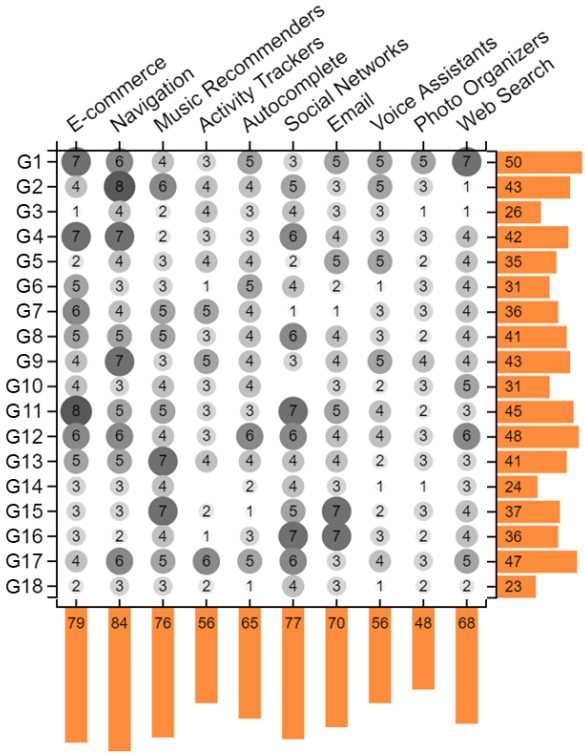
(a) **Counts of "clear application" or "application" responses.**



(b) **Counts of "clear violation" or "violation" responses.**



(c) **Counts of "does not apply" responses.**



(d) **Counts of all responses, excluding "does not apply" and including neutral.**

Figure 1: Counts of applications (top left), violations (top right), and "does not apply" (bottom left) responses in our user study. Rows show counts by guideline, while columns show counts by product category tested.

and design that facilitate the implementation of such guidelines. For example, frequent item sets and location detection were two common mechanisms used to support Guideline 4 in showing contextually relevant information (e.g., [E-commerce, Product #2] "The feature assumes I'm about to buy a gaming console and shows accessories and games that would go with it..." or [Web Search, Product #2] "Searching a movie title returns show times near my location for today's date"). For Guideline 12, multiple products leverage the history of user interactions to suggest a reduced cache of items that might be more useful to the user (e.g., [Navigation, Product #1] "Opening the app shows a list of recent destinations, as well as allows you to access 'favorite' locations.").

Some guidelines emerged as relevant, but not widely implemented, as indicated by the large number of violations. For example, Guideline 11 "Make clear why the system did what it did" had one of the highest number of violations, despite the large volume of active research in the area of intelligibility and explanations. This guideline also had one of the fewest reported instances of "does not apply", suggesting that participants could imagine opportunities for explanations, but were often unable to obtain them. In some cases, participants reported violations when they were unable to locate any explanation at all (e.g., [E-commerce, Product #1] "I have no idea why this is being shown to me. Is it trying to sell me stuff I do not need?" and [Music Recommender, Product #1] "Even when drilling down into a song there is no explanation for why this particular song was recommended."). In other cases, participants reported violations when explanations were provided but were seemingly inadequate for their purposes (e.g., [Email, Product #1] "This does list out things which affect it, but they don't explain it in a clear manner. Do each of these affect it equally?" and [Navigation, Product #1] "It always says the suggested route is the "best route" but it doesn't give you the criteria for why that route is the best."). These results suggest that participants could envision explanations being useful in most of the products we tested, but more work is necessary to understand the level of explanations people may desire and how designers can produce them. In some cases, explanations might be undesirable, for financial or business reasons (e.g. adversarial (gaming) behavior by Web page authors would be exacerbated if search engines explained their ranking.).

Guidelines 3, 5, and 6 had the highest number of "does not apply" ratings. Several participants indicated that Guideline 3 was not applicable because the products they were testing presented services only when explicitly requested by the user. For example, for one of the E-commerce products, one participant stated, "I feel this guideline does not apply for the recommendations page. It [is] a very 'pull' kind of interaction."; i.e., the user views recommendations while browsing and there are no 'push' notifications. Similarly, for one of the Web Search engines we tested, one participant stated, "[the search engine] does not generally interrupt a user at any point. The mobile app has notifications, which might be relevant here, but the desktop website does not. Generally speaking, AI services pop up based on when the user searches and what he or she searches for, not based on an ongoing session." This guideline is therefore likely more relevant for products that take proactive actions without explicit user requests, such as sending notifications.

Guideline 5 "Match relevant social norms" and Guideline 6 "Mitigate social biases" had some of the most reported instances of "does not apply". Examination of these instances revealed that in some cases participants firmly believed these guidelines were not relevant for the products they were testing while other participants reported either applications or violations of these guidelines in those same product categories. For example, one participant reported about one of the navigation products tested that "information is not subject to biases, unless users are biased against fastest route". However, a different participant was able to identify a violation of this guideline for the same product category "Regards the 'Walking' transport there's no way to set an average walking speed. [The product] assumes users to be healthy." Similarly, one participant reported about one of the voice assistants we tested that "Nothing in this interaction had any social biases that it could reinforce.", while another stated about the same product that "While it's nice that a male voice is given as an option, the default [voice assistant] voice is female, which reinforces stereotypical gender roles that presume a secretary or receptionist is female." Some participants, however, had no trouble identifying bias: "I typed in 'black' in the search bar and it came back with images of me as well as my niece [...] it saw a black face and used that as its frame of reference for all pictures, then returned all pictures of me and my family without images of other black spaces in an environment".

Guidelines 5 and 6 were noted as the least clear by our participants in their (see Figure 2), with several participants remarking about the difficulty of imagining social norms beyond their own or recognizing potential sources of bias (e.g., "Hard for a designer to implement, because it requires them to think outside of their own social context", "Doesn't apply to me but to potential other people.", and "This is hard to measure. Who defines what is undesirable and unfair?"). These assessments suggest that a diverse set of evaluators may be necessary to effectively recognize or apply these guidelines in practice. Alternatively, designers may need specific training or tools to recognize social norms and biases. GenderMag[5], a method for identifying gender biases in user interfaces, is one such tool, but further work is needed in this area.

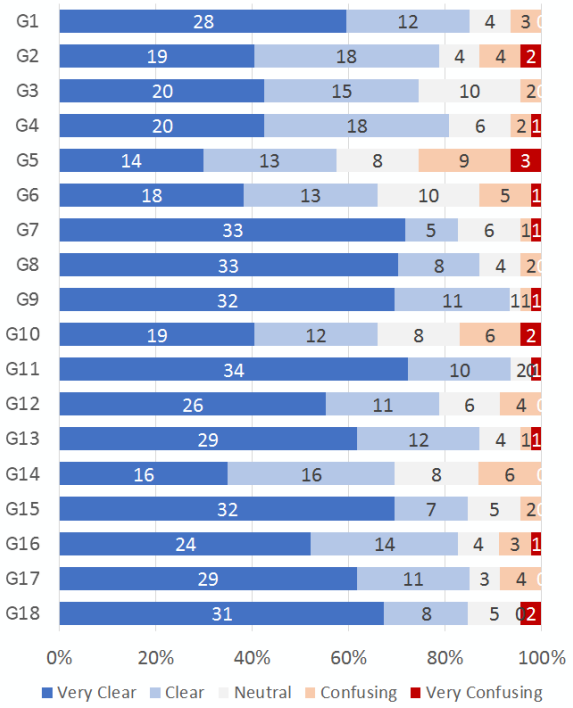


Figure 2: Subjective evaluations by study participants about the clarity of the 18 AI design guidelines.

Clarity and Clarifications. Figure 2 presents clarity ratings for all guidelines. To identify guidelines in need of further clarification, we reviewed these ratings and the 56 misinterpretations explained in the section Adjustments and Misinterpretations. We noted guidelines as needing further clarification when errors were determined to be systematic, which we defined as having four or more instances confused with another guideline or having multiple participants making similar comments about clarity. From this analysis, we identified and addressed the following issues:

Guidelines 1, 2, and 11 (originally phrased as "Make capabilities clear", "Set expectations of quality" and "Make explanations of behavior available") had 13 misinterpreted instances (5 between Guidelines 1 and 2; 8 between Guidelines 1 and 11) and several comments about these being hard to differentiate (e.g., one participant commented on Guideline 2 that "I don't know what is different between this guideline and the guideline #1"). To clarify, we revised these guidelines using parallel language while emphasizing the intended differences (Guideline 1 is about *what* the system can do, Guideline 2 is about how *well* the system can do it, and Guideline 11 is about explaining *why* something happened, after the fact).

Guideline 4 (originally phrased as "Show contextually relevant information. Display information about the user's inferred goals and attention during interaction.") was confused with Guideline 13 (originally phrased as "Learn from user behavior. Personalize the experience based on the user's past actions") six times. Examination revealed that most of these errors were due to "preferences" and "personalization" being considered as "relevant context". To clarify, we rephrased these guidelines as in Table 1, emphasizing the difference between a user's "current context" (e.g., "current task and environment") and personalization which we intended to mean learning about preferences "over time".

Guidelines 3 and 4 (originally phrased as "Time services based on context" and "Show contextually relevant information") were confused with each other in four instances. Several participants commented that this was because *what* is displayed and *when* it is displayed are often related (e.g., "provide the right information at the right time" and "The time when I'm specifically looking for DP to HDMI cable should be the most ideal time to recommend possible variations in DP to HDMI"). However, we decided to keep these guidelines separate to avoid conjunctions and updated Guideline 3 to use the same language of "current task and environment" as Guideline 4.

Guideline 12 (originally phrased as "Maintain working memory") was confused with Guideline 13 ("Learn from user behavior") seven times, seemingly because the term "memory" was being interpreted as something that happens over time. To clarify, we revised these guidelines as in Table 1 to emphasize the difference between maintaining short term memory of recent interactions and learning behaviors over time.

Guidelines 15 and 17 (originally phrased as "Encourage feedback" and "Provide global controls", respectively) were confused six times, seemingly because the difference between local (or instance-level) feedback and global feedback (e.g., settings that impact behaviors on all instances) was still unclear despite introducing the term "global" after our first heuristic evaluation in Phase 2. We therefore revised these Guidelines as in Table 1 to further emphasize that Guideline 15 is about granular feedback that happens during a specific interaction, while Guideline 17 is about global customization of behaviors.

These revisions resulted in the final set of guidelines presented in Table 1, which we further evaluated with experts as described in the following section.

6 PHASE 4: EXPERT EVALUATION OF REVISIONS

To verify whether the revisions we proposed in the previous section improved our guidelines, we conducted an expert review. Expert reviews have been shown to be effective at identifying problems related to wording and clarity [27, 34, 43]. For this purpose, we defined experts as people who have work experience in UX/HCI and who are familiar with discount usability methods such as heuristic evaluation.

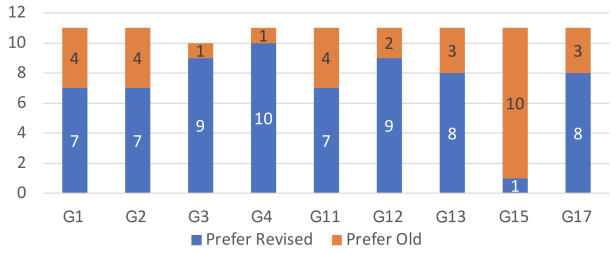


Figure 3: Number of experts out of 11 who preferred the revised or the old version. One participant suggested their own alternative for Guideline 3.

We reasoned that experts with experience in applying various guidelines to design solutions would be able to assess whether our guidelines would be easy to understand and therefore to work with.

We recruited 11 experts (6 female, 5 male) from the same large company through snowball sampling. Of these experts, 6 were UX designers, 3 were UX researchers, and two were in research and product planning roles. Their length of experience working in UX or HCI was more than 20 years (1), 16-20 years (4), 11-15 years (3), and 2-5 years (3). Participants self-reported their familiarity with discount usability methods as very high (5), high (4), and medium (2).

First, we asked each expert to review the 9 revised guidelines independently. They chose, for each guideline, the version they thought was easier to understand (the old version or the version we revised after in Phase 3). Then the experts reviewed the pairs of guidelines that emerged in Phase 3 as confusing or overlapping. For each pair, we asked experts to rate whether the two guidelines mean the same thing and the difficulty of distinguishing between them. We compensated participants with a $30 gift card for an estimated time commitment of 45 minutes.

Figure 3 shows that experts preferred the revised versions for all but Guideline 15. Revisions appear to have helped distinguish between the pairs of guidelines Phase 3 participants had trouble with, but five experts still found Guidelines 1 and 2 somewhat difficult to distinguish (Table 4). Since the revision of Guideline 15 made it easy to distinguish it from 17, we decided to keep it.

Table 3 illustrates the evolution of the first two guidelines through the four phases.

Phase 1: Consolidating guidelines
Set appropriate expectations.
Set accurate expectations to give people a clear idea of what the experience is and isn't capable of doing.
Phase 2: Internal evaluation
Set appropriate expectations.
Phase 3: User study
G1: Make capabilities clear. Help the user understand what the AI system is capable of doing.
G2: Set expectations of quality. Help the user understand what level of performance the AI system is capable of delivering.
Phase 4: Expert evaluation of revisions
G1: Make clear what the system can do. Help the user understand what the AI system is capable of doing.
G2: Make clear how well the system can do what it can do. Help the user understand how often the AI system may make mistakes.

Table 3: Evolution of Guidelines 1 and 2.

7 DISCUSSION & FUTURE WORK

We synthesized guidance proposed over the past 20 years about the design of human-AI interaction into a set of 18 AI usability guidelines. These guidelines were iteratively refined in four phases by a team of 11 researchers, and were applied or reviewed by an additional 60 designers and usability practitioners. Over the various stages of development, the guidelines were applied to AI-infused products across 10 product categories. These efforts provide evidence for the relevance of the guidelines across a wide range of common AI-infused systems. In terms of utility, we anticipate the guidelines will be useful to evaluate existing products and emerging design ideas. Our evaluation methods show that the guidelines lend themselves well to usability inspection methods such as heuristic evaluation. Future work could examine the uses and value of these guidelines at various stages of design.

We recognize that there is a tradeoff between generality and specialization, and that these guidelines might not adequately address all types of AI-infused systems. For example, we reported that some guidelines do not directly apply to AI systems that lack graphical user interfaces (e.g., voice-based virtual assistants and activity trackers). Additional guidelines may be necessary to help designers and developers create intuitive and effective products with these properties or in these product categories. Likewise, specialized guidelines may be required in certain high-risk or highly regulated areas such as semi-autonomous vehicles, robot-assisted surgery, and financial systems. We hope the 18 guidelines presented here and their validation process stimulate and inform future research into the development of domain-specific guidance.

Our work also intentionally focused on AI design guidelines that we believed could be easily evaluated by inspection of a system's interface. For example, we excluded broad principles such as "build trust", and focused instead on specific and observable guidelines that are likely to contribute to building trust. Previous work, however, has proposed guidelines that impact the usability of AI-infused systems but must

be considered when constructing the AI model. For example, we excluded Horvitz's [17] principle of "inferring ideal action in light of costs, benefits, and uncertainties" and guidance about being "especially conservative in the beginning" because these require decisions to be made at the modeling layer of a system. We foresee the value of future work to investigate how designers and model developers can work together to effectively apply these guidelines in AI-infused systems. For example, given the expected performance of an AI model, designers may recommend specific designs that reduce costs while optimizing benefits to users (e.g., displaying multiple options to users until the performance of the AI model is improved enough to take proactive action on the user's behalf).

Our decisions to optimize for generality, and to focus on observable properties, serve as a reminder that interaction designers routinely encounter these types of trade-offs. We anticipate situations where there will be interactions and trade-offs in attempts to employ several of the guidelines. As an example, if a system uses a complex or deep model to achieve a high level of performance, it may be challenging to both convey the consequences of user actions (Guideline 16), while also actively learning from user behavior (Guideline 13). Further research is necessary to understand the implications of these potential interactions and trade-offs for the design of AI systems and to understand how designers employ these guidelines "in the wild."

Finally, we recognize that our guidelines only begin to touch on topics of fairness and broader ethical considerations. Ethical concerns extend beyond the matching of social norms (Guideline 5) and mitigating social biases (Guideline 6). As an example, an AI system may adhere to each of these guidelines and yet impact people's lives or livelihoods in a consequential manner. It is imperative that system designers carefully evaluate the many influences of AI technologies on people and society, and that this remains a topic of ongoing research and intense interest. Ethics-focused guidelines can be difficult to fully evaluate in a heuristic evaluation, and successful detection of problems may depend on who is performing the evaluation. Our results related to Guidelines 5 "Match relevant social norms" and 6 "Mitigate social biases" suggest that diversity among evaluators helps identify a range of issues that might be invisible to members of majority groups.

8 CONCLUSION

We proposed and evaluated 18 generally applicable design guidelines for human-AI interaction. We distilled the guidelines from over 150 AI-related design recommendations and validated them through three rounds of evaluation. We are hopeful that application of these guidelines will result in better, more human-centric AI-infused systems, and that our synthesis can facilitate further research. As the current technology landscape is shifting towards the increasing inclusion of AI in computing applications, we see significant value in working to further develop and refine design guidelines for human-AI interaction.

Guidelines	Meanings: Different	Distinguish: Easy	Distinguish: Hard	Distinguish: Neutral/ Medium
1 & 2	10	6	5	0
1 & 11	11	6	3	2
3 & 4	10	6	2	3
4 & 13	11	9	1	1
12 & 13	9	7	1	3
15 & 17	10	9	1	1

Table 4: Number of experts out of 11 who rated each pair of guidelines as different in meaning and distinguishable.

9 ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of our newest team member, Ever Zayn McDonald.

REFERENCES

[1] Alexa. 2018. Top sites in the United States. Retrieved July, 2018 from https://www.alexa.com/topsites/countries/US

[2] Kathy Baxter. 2017. How to Meet User Expectations for Artificial Intelligence. Medium. Retrieved September, 2018 from https://medium.com/salesforce-ux/how-to-meet-user-expectations-for-artificial-intelligence-a51d3c82af6

[3] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human–Computer Interaction* 16, 2-4 (2001), 193–212.

[4] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are Explanations Always Important?: A Study of Deployed, Low-cost Intelligent Interactive Systems. In *Proc. IUI '12*. ACM, New York, NY, USA, 169–178.

[5] 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.

[6] comScore. 2018. Latest rankings. Retrieved July, 2018 from https://www.comscore.com/Insights/Rankings

[7] Maartje de Graaf, Somaya Ben Allouch, and Jan van Dijk. 2017. Why Do They Refuse to Use My Robot?: Reasons for Non-Use Derived from a Long-Term Home Study. In *Proc. HRI '17*. ACM, New York, NY, USA, 224–233.

[8] Defy Media. 2015. Damn You Auto Correct! Retrieved September, 2018 from http://www.damnyouautocorrect.com/

[9] T. Deuschel and T. Scully. 2016. On the Importance of Spatial Perception for the Design of Adaptive User Interfaces. In *2016 IEEE 10th International Conference on Self-Adaptive and Self-Organizing Systems (SASO)*. 70–79.

[10] Anind Dey, Jennifer Mankoff, Gregory Abowd, and Scott Carter. 2002. Distributed Mediation of Ambiguous Context in Aware Environments. In *Proc. UIST '02*. ACM, New York, NY, USA, 121–130.

[11] Leah Findlater and Joanna McGrenere. 2004. A Comparison of Static, Adaptive, and Adaptable Menus. In *Proc. CHI '04*. ACM, New York, NY, USA, 89–96.

[12] Krzysztof Z. Gajos, Mary Czerwinski, Desney S. Tan, and Daniel S. Weld. 2006. Exploring the Design Space for Adaptive Graphical User

Interfaces. In *Proc. AVI '06*. ACM, New York, NY, USA, 201–208.

[13] Krzysztof Z Gajos, Katherine Everitt, Desney S Tan, Mary Czerwinski, and Daniel S Weld. 2008. Predictability and accuracy in adaptive user interfaces. In *CHI*. ACM, 1271–1274.

[14] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proc. CSCW '00*. ACM, New York, NY, USA, 241–250.

[15] Setia Hermawati and Glyn Lawson. 2016. Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus? *Applied Ergonomics* 56 (2016), 34 – 51.

[16] Kristina Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers* 12, 4 (2000), 409–426.

[17] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proc. CHI '99*. ACM, New York, NY, USA, 159–166.

[18] Eric Horvitz, Jack Breese, David Heckerman, David Hovel, and Koos Rommelse. 1998. The Lumière Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In *Proc. UAI '98*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 256–265.

[19] IDC Corporate USA. 2018. IDC: The premier global market intelligence firm. Retrieved September, 2018 from https://www.idc.com/

[20] Anthony Jameson. 2008. Adaptive interfaces and agents. In *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (2nd ed.), Andrew Sears and Julie A. Jacko (Eds.). CRC Press, Boca Raton, FL, 433–458.

[21] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proc. IUI '15*. ACM, New York, NY, USA, 126–137.

[22] Chia-Jung Lee, Jaime Teevan, and Sebastian de la Chica. 2014. Characterizing Multi-click Search Behavior and the Risks and Opportunities of Changing Results During Use. In *Proc. SIGIR '14*. ACM, New York, NY, USA, 515–524.

[23] Brian Y. Lim and Anind K. Dey. 2009. Assessing Demand for Intelligibility in Context-aware Applications. In *Proc. UbiComp '09*. ACM, New York, NY, USA, 195–204.

[24] Josh Lovejoy. 2018. The UX of AI. Google Design. Retrieved September, 2018 from https://design.google/library/ux-ai/

[25] Josh Lovejoy and Jess Holbrook. 2017. Human-Centered Machine Learning. 7 steps to stay focused on the user when designing with ML. Medium. Retrieved September, 2018 from https://medium.com/google-design/human-centered-machine-learning-a770d10562cd

[26] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proc. CHI '16*. ACM, New York, NY, USA, 5286–5297.

[27] Aaron Maitland and Stanley Presser. 2016. How Accurately Do Different Evaluation Methods Predict the Reliability of Survey Questions? *Journal of Survey Statistics and Methodology* 4, 3 (2016), 362–381.

[28] Joseph A Maxwell. 2012. *Qualitative research design: An interactive approach*. Vol. 41. Sage publications.

[29] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proc. CHI '18*. ACM, New York, NY, USA, Article 6, 7 pages.

[30] Jakob Nielsen. 1992. Finding Usability Problems Through Heuristic Evaluation. In *Proc. CHI '92*. ACM, New York, NY, USA, 373–380.

[31] Jakob Nielsen and Rolf Molich. 1990. Heuristic Evaluation of User Interfaces. In *Proc. CHI '90*. ACM, New York, NY, USA, 249–256.

[32] D.A. Norman. 1988. *The psychology of everyday things*. Basic Books, New York.

[33] Donald A. Norman. 1994. How Might People Interact with Agents. *Commun. ACM* 37, 7 (July 1994), 68–71.

[34] Kristen Olson. 2010. An examination of questionnaire evaluation by expert reviewers. *Field Methods* 22, 4 (2010), 295–318.

[35] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proc. CHI '18*. ACM, New York, NY, USA, Article 640, 12 pages.

[36] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations As Mechanisms for Supporting Algorithmic Transparency. In *Proc. CHI '18*. ACM, New York, NY, USA, Article 103, 13 pages.

[37] Ruth Ravichandran, Sang-Wha Sien, Shwetak N. Patel, Julie A. Kientz, and Laura R. Pina. 2017. Making Sense of Sleep Sensors: How Sleep Sensing Technologies Support and Undermine Sleep Health. In *Proc. CHI '17*. ACM, New York, NY, USA, 6864–6875.

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. KDD '16*. ACM, New York, NY, USA, 1135–1144.

[39] Katharine Schwab. 2017. 10 Principles For Design In The Age Of AI. Retrieved September, 2018 from https://www.fastcompany.com/3067632/10-principles-for-design-in-the-age-of-ai

[40] SimilarWeb. 2018. Top websites ranking. Retrieved July, 2018 from https://www.similarweb.com/top-websites/united-states

[41] Jack Stewart. 2018. Why Tesla's Autopilot Can't See a Stopped Firetruck. Retrieved September, 2018 from https://www.wired.com/story/tesla-autopilot-why-crash-radar/

[42] Erica Virtue. 2017. Designing with AI.

[43] Jolita Vveinhardt and Evelina Gulbovaitė. 2016. Expert evaluation of diagnostic instrument for personal and organizational value congruence. *Journal of business ethics* 136, 3 (2016), 481–501.

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

[45] Rayoung Yang, Eunice Shin, Mark W. Newman, and Mark S. Ackerman. 2015. When Fitness Trackers Don'T 'Fit': End-user Difficulties in the Assessment of Personal Tracking Device Accuracy. In *Proc. UbiComp '15*. ACM, New York, NY, USA, 623–634.