Measuring User Experience Inclusivity in Human-AI Interaction via Five User Problem-Solving Styles

ANDREW ANDERSON, Oregon State University
JIMENA NOA GUEVARA, Oregon State University
FATIMA MOUSSAOUI, Oregon State University
TIANYI LI, Purdue University
MIHAELA VORVOREANU, Microsoft
MARGARET BURNETT, Oregon State University

Motivations: Recent research has emerged on generally how to improve AI products' Human-AI Interaction (HAI) User Experience (UX) , but relatively little is known about HAI-UX inclusivity. For example, what kinds of users are supported, and who are left out? What product changes would make it more inclusive?
Objectives: To help fill this gap, we present an approach to measuring what kinds of diverse users an AI product leaves out and how to act upon that knowledge. To bring actionability to the results, the approach focuses on users' problem-solving diversity. Thus, our specific objectives were: (1) to show how the measure can reveal which participants with diverse problem-solving styles were left behind in a set of AI products; and (2) to relate participants' problem-solving diversity to their demographic diversity, specifically gender and age.
Methods: We performed 18 experiments, discarding two that failed manipulation checks. Each experiment was a 2x2 factorial experiment with online participants, comparing two AI products: one deliberately violating one of 18 HAI guideline and the other applying the same guideline. For our first objective, we used our measure to analyze how much each AI product gained/lost HAI-UX inclusivity compared to its counterpart, where inclusivity meant supportiveness to participants with particular problem-solving styles. For our second objective, we analyzed how participants' problem-solving styles aligned with their gender identities and ages.
Results & Implications: Participants' diverse problem-solving styles revealed six types of inclusivity results: (1) the AI products that followed an HAI guideline were almost always more inclusive across diversity of problem-solving styles than the products that did not follow that guideline—but "who" got most of the inclusivity varied widely by guideline and by problem-solving style; (2) when an AI product had risk implications, four variables' values varied in tandem: participants' feelings of control, their (lack of) suspicion, their trust in the product, and their certainty while using the product; (3) the more control an AI product offered users, the more inclusive it was; (4) whether an AI product was learning from "my" data or other people's affected how inclusive that product was; (5) participants' problem-solving styles skewed differently by gender and age group; and (6) almost all of the results suggested actions that HAI practitioners could take to improve their products' inclusivity further. Together, these results suggest that a key to improving the demographic inclusivity of an AI product (e.g., across a wide range of genders, ages, etc.) can often be obtained by improving the product's support of diverse problem-solving styles.

Authors' addresses: Andrew Anderson, anderan2@oregonstate.edu, Oregon State University; Jimena Noa Guevara, noaguevg@oregonstate.edu, Oregon State University; Fatima Moussaoui, moussaof@oregonstate.edu, Oregon State University; Tianyi Li, li4251@purdue.edu, Purdue University; Mihaela Vorvoreanu, Mihaela.Vorvoreanu@microsoft.com, Microsoft; Margaret Burnett, burnett@oregonstate.edu, Oregon State University.

2160-6455/9999/99-ART99 $15.00
https://doi.org/99.9999/9999999.9999999

CCS Concepts: • **Human-centered computing** → **User studies**; • **Computing methodologies** → *Intelligent agents.*

Additional Key Words and Phrases: Intelligent User Interfaces, Human-Computer Interaction

ACM Reference Format:
Andrew Anderson, Jimena Noa Guevara, Fatima Moussaoui, Tianyi Li, Mihaela Vorvoreanu, and Margaret Burnett. 9999. Measuring User Experience Inclusivity in Human-AI Interaction via Five User Problem-Solving Styles . *ACM Trans. Interact. Intell. Syst.* 99, 99, Article 99 (9999), 39 pages. https://doi.org/99.9999/9999999.9999999

1 INTRODUCTION

Suppose the owner of an AI Product named "G3" ran a before/after user study, to find out whether potential customers had better user experiences (UX) with G3's new version than an older, not very good version of G3, and the study results came out like Figure 1.

Fig. 1. An outcome of an experiment comparing two versions of "G3" AI products [83]. X-axis shows the amount of improvement for 13 UX variables (unlabeled here) on the y-axis. *: statistically significant difference.

The product owner should be somewhat pleased—the product UX clearly improved. Of the 13 measured UX outcome variables (y-axis), 11 were positive. In fact, the *'s indicate that these 11 differences were significant between the old and new versions of G3. Still, the effect sizes were mainly small (yellow bars), with only two moderate effect sizes (blue bars).

What the product owner would now like to know is: who was included in those positive effects, and *who was left out*? What further changes are needed to enable G3 to better support more of the potential customers?

These kinds of questions are human-AI interactions (HAI) questions about the user experience (UX) quality that AI products offer their customers. In this paper, we abbreviate the concept of HAI user experiences as HAI-UX.

One method that the HAI community currently uses to improve AI products' user experiences is to develop and apply guidelines for human-AI interaction. At least three major companies—Apple, Google, and Microsoft—have each proposed guidelines, providing high-level advice for how to improve human-AI interaction, such as "Consider offering multiple options when requesting explicit feedback" [64], "Let users give feedback" [47], and "Support efficient correction" [6]. In fact, G3's product owner followed a guideline from the Microsoft set, and doing so improved the product (Figure 1).

In this paper, we consider how to measure beyond just *whether* products like G3 improve their HAI-UX. We investigate how to know *who*, of all the diverse humans who could be using products like G3, is *included* in our HAI-UX improvements and who has been left out. Applying the concept of inclusivity to human interactions with AI products, we will say that AI product A is more *inclusive* to some particular group of people than product B is, if product A provides those particular people with measurably better user experience outcomes than product B does.

The primary groups of interest in this paper are those who are diverse in the ways they go about *problem-solving*. We use the term *problem-solving* to mean any time that people are engaged in solving problems, such as whether and how to accept/reject an AI's recommendations. We consider participants' problem-solving diversity via the set of five problem-solving style spectra from the inclusive design method known as GenderMag [19]. We use the term *problem-solving styles* to refer to the approaches that individuals take to go about trying to solve a problem. GenderMag's

five problem-solving style spectra are people's diverse: attitudes toward risk, levels of computer self-efficacy, motivations, information processing style, and styles of learning technology.

For example, "risk-averse" is one endpoint of the risk attitude spectrum. Applying risk-aversion to technology, risk-averse users may be hesitant to invoke a new feature for fear that it may have undesirable side-effects (e.g., privacy), may waste their time, may not be worth learning about, etc. At the other end of the spectrum, "risk-tolerant" users may be more willing to take such risks, even if the feature has not been proven to work yet and requires additional time to understand [53, 98, 125].

In this investigation, we consider how user experiences of people with diverse problem-solving styles were impacted by design differences in AI-powered systems like G3 above. Specifically, we gathered 1,016 participants' five GenderMag problem-solving styles, and investigated inclusivity differences in 16 pairs of AI products. Each AI product had controlled differences: one AI product applied an HAI guideline from the Amershi et al. guidelines set [6], and its counterpart violated that guideline. All AI products were productivity software (e.g., Microsoft PowerPoint, etc.) that had added AI features. An earlier investigation on the same data, reported in Li et al. [83], investigated the "whether" questions of these data, i.e., whether HAI-UX differences between each pair AI products occurred. That investigation found that participants's HAI-UX outcomes were generally better when the guidelines were applied; Figure 1 is in fact one example of their findings. Our investigation instead focuses on "who" questions, i.e, who were included (and who were not) in the HAI-UX outcome changes, from the perspective of participants' diverse problem-solving styles.

To show how analyzing HAI-UX data by these five problem-solving styles can reveal actionable insights into how to improve an AI product's inclusivity, this paper presents a detailed analysis of one of these problem-solving style types, namely attitudes toward risk. However, space constraints prevent providing detailed analyses for all five of these problem-solving style types, so this paper summarizes the remaining four problem-style types' results with an eye toward generality; we also provide detailed analyses of all five problem-solving style types in the Appendices. We selected attitudes toward risk as the problem-solving type to present in detail, because of the preponderance of recent research literature and popular perception focusing on risks with AI, such as risks of inaccuracies, of privacy loss, of excessive or insufficient trust, of job loss, and more (e.g., [31, 35, 45, 57, 61, 62, 72, 118]). We investigate:

> **RQ1-Risk:** *When the HAI guidelines are violated vs. applied to AI products, how inclusive are the resulting AI products to users with diverse attitudes toward risk?*
> **RQ2-AllStyles:** *How inclusive are such products to users with diverse values of GenderMag's other four problem-solving styles?*

We also investigate the relationship between participants' problem-solving style diversity and their demographic diversity. Our reason for relating problem-solving diversity with demographic diversity is that knowing the demographic disparities in who a product serves may not lead to *actionable* ways to address those disparitarites; for example, if one gender is left out of high-quality user experiences with an AI product, how to fix it? In contrast, problem-solving style disparities often do suggest actionable ways forward; for example, if risk-averse participants are left out of high-quality user experiences, perhaps the product should be clearer about risks of using it (e.g., its privacy impacts):

> **RQ3-DemographicDiversity:** *How does AI product users' problem-solving diversity align with their demographic diversity?*

Thus, the new contributions of our research are:

- *Measuring HAI-UX inclusivity*: Presents an approach to measure inclusivity of an AI product's user experiences.

- *Risk-inclusivity in HAI-UX*: Uses the approach to reveal which of the participants with diverse attitudes toward risk are well-supported by a set of 16 AI products and which are not.
- *Beyond risk-inclusivity in HAI-UX*: Generalizes the above results to the other four GenderMag problem-solving style spectra.
- *Actionable inclusivity in HAI-UX*: Reveals whether results to the above suggest actionable steps an HAI practitioner can take to make an AI product more inclusive.
- *Problem-solving diversity and demographic diversity*: Reveals relationships between participants' problem-solving styles and their intersectional gender-and-age demographic diversity, to enable HAI practitioners to bring actionable results from problem-solving diversity investigations to bear on demographic disparities.
- *Implications for practitioners*: Suggests concrete ways HAI practitioners can use the approach on their own products, as well as starting points to develop new criteria, guidelines, and/or onboarding processes on designing their own AI products to be more inclusive of and equitable to diverse customers.

2 BACKGROUND & RELATED WORK

2.1 Background

2.1.1 The Gender Inclusiveness Magnifier (GenderMag). The GenderMag problem-solving styles are foundational underpinnings to this investigation (Table 1). The GenderMag *method* is an inclusive design and evaluation method based on these five problem-solving style types [20]; software professionals use the method to detect "(gender-)inclusivity bugs."[1] GenderMag's problem-solving styles are particularly well-suited to our investigation into diverse users' experiences with AI features, because GenderMag was developed to improve technology's *inclusiveness* for *problem-solving technology*—such as spreadsheet development, software debugging, and any other domains where problems can arise with which users must grapple [20]. Because the GenderMag method is intended for practical use by developers without social science backgrounds, a set of criteria [20, 92] were applied to the original long list of applicable problem-solving style types [13] to reduce this list to the five styles in Table 1. These five style types have been repeatedly identified as having strong ties to both problem-solving and gender[2]; we will summarize some of this research shortly.

Each of these problem-solving style types have continuous ranges whose endpoints (Table 1's column 2 & 4) are the only distinguished values. Values at one end are assigned to a persona named "Abi," those at the other end are assigned to a persona named "Tim," and a mix of values are assigned to a persona named "Pat." For example, Abi and Pat are more risk-averse about technology risks than Tim (Table 1's row 1), so Abi and Pat might be less likely than Tim to use the same password on multiple sites.

To summarize these five styles (see [20] for details):

- Attitudes toward risk: Studies across multiple domains have reported a wide diversity in people's attitudes toward risk (e.g., [37, 126]) and how they solve problems [27, 63, 127]. Gender differences have also been reported in risk and decision in numerous problem-solving domains, with women almost always (statistically) less willing to take risks than other people [27, 127]. Note that risks relevant to decision-making include not only obvious risks like privacy and security, but also risk of wasting time and/or of failing [74].

[1] GenderMag finds the issues not by using people's gender identity, but rather by the five problem-solving style types. These problem-solving styles' values statistically cluster around genders.

[2] Some of these works gathered participants' biological sex rather than their gender identities; others by gender. Further, most of the literature has been binary, reporting only females/males or women/men, so the upcoming description of the styles is also necessarily binary. In this discussion, we use gender terminology (e.g., "woman" instead of "female") simply to avoid switching back and forth for different studies.

	 **Abigail/Abishek ("Abi")**	 **Patricia/Patrick ("Pat")**	 **Timara/Timothy ("Tim")**
Attitude toward Risk Range: Risk-averse – Risk-tolerant	Risk-averse	Risk-averse	Risk-tolerant
Computer Self-Efficacy Range: lower – higher	Lower (relative to peers)	Medium	Higher (relative to peers)
Motivations Range: task-oriented – tech-oriented	Task oriented: wants what technology can accomplish	Task oriented: wants what technology can accomplish	Tech oriented: technology is a source of fun
Information Processing Style Range: comprehensive – selective	Comprehensive	Comprehensive	Selective
Learning Style Range: by Process – by Tinkering	Process-oriented learner	Learns by tinkering: tinkers reflectively	Learns by tinkering (sometimes to excess)

Table 1. The five GenderMag problem solving style types (rows), each type's range of possible values, and the set of values for each. The "Abi" values (left) are the values at one end of each type, and the "Tim" values (right) are at the other end. Any individual can have any combination of values within these types, but in aggregate, the results have statistically clustered by people's self-identified gender (e.g., [20, 117, 125]).

- Computer self-efficacy: One specific form of confidence is self-efficacy—people's belief in their ability to succeed in a specific task [12]. Self-efficacy matters to problem solving because it influences people's use of cognitive strategies, effort exerted, persistence with a problem, and coping strategies [12]. Regarding gender, empirical data have shown that women tend statistically to have lower computer self-efficacy than other people [117]. Overall, ties between people's self-efficacy and how they approach a variety of problem-solving tasks have been well-documented in many domains (e.g., [98, 124, 132, 135]).
- Motivations: In the context of technology, motivations are the reasons an individual decides to interact with technology, such as using technology mainly to accomplish a task, versus having an interest and enjoyment in using and exploring technology (e.g., [116]). Gender differences have been reported in these "task-oriented" and "tech-oriented" motivations [117]. An individual's motivations can affect not only which technology features they decide to use, but also how they go about using those features [16, 51, 70, 97, 116].
- Information processing style: Solving problems often requires gathering information. However, individuals vary on how much information they gather and when. Some gather information comprehensively—i.e., in sizeable batches—to form an approach first and then carry it out, whereas others gather it selectively, acting upon the first promising information, then possibly gathering a little more information before taking the next action, and so on. Regarding gender, women are statistically more likely to process information comprehensively and men are statistically more likely to process it selectively [117]. In both research and practice, much attention has been given to how technology can enable different individuals to obtain the right amount of information at the right time (e.g., [26, 100, 119, 128]).
- Learning Style for Technology (By process vs. by tinkering): Learning style considers how people go about solving problems by how they structure their approach. For example, some

people prefer to learn new technology by an organized process, like a recipe. Others prefer to tinker, exploring options and experimenting in a "what if I did this" way. Regarding gender, women have been shown to be statistically less likely than other people to use the latter approach when encountering features new to them [117]. Because of such differences, technology organizations have begun to stress the importance of supporting the entire range of Learning Style values; one example is in Microsoft's design guidelines [74].

In GenderMag evaluations across the world, these five problem-solving style types have repeatedly shown impacts on which technology features diverse people decide to use and/or how they use them [4, 23, 30, 49, 75, 88, 95, 98, 111, 125]. However, most such evaluations have been outside the context of AI. This paper is within AI contexts and isolates the human-AI interactions from all other interactions.

2.1.2 Guidelines for Human-AI Interaction. A second key component of this paper is Amershi et al.'s 18 guidelines for human-AI interaction [6]. This set of 18 guidelines for human-AI interaction, depicted in Figure 2, provides high-level advice for HAI designers. Each guideline has three components (1) a number, (2) a name which provides high-level advice (e.g., "Make clear what the system can do"), and (3) a brief description of what the guideline means (e.g., "Help the user understand what the AI system is capable of doing"). Amershi et al. also ran an initial study to investigate how designers of AI-powered systems perceived these guidelines, and the designers found that the guidelines were clear and that they could find examples of these guidelines [6].

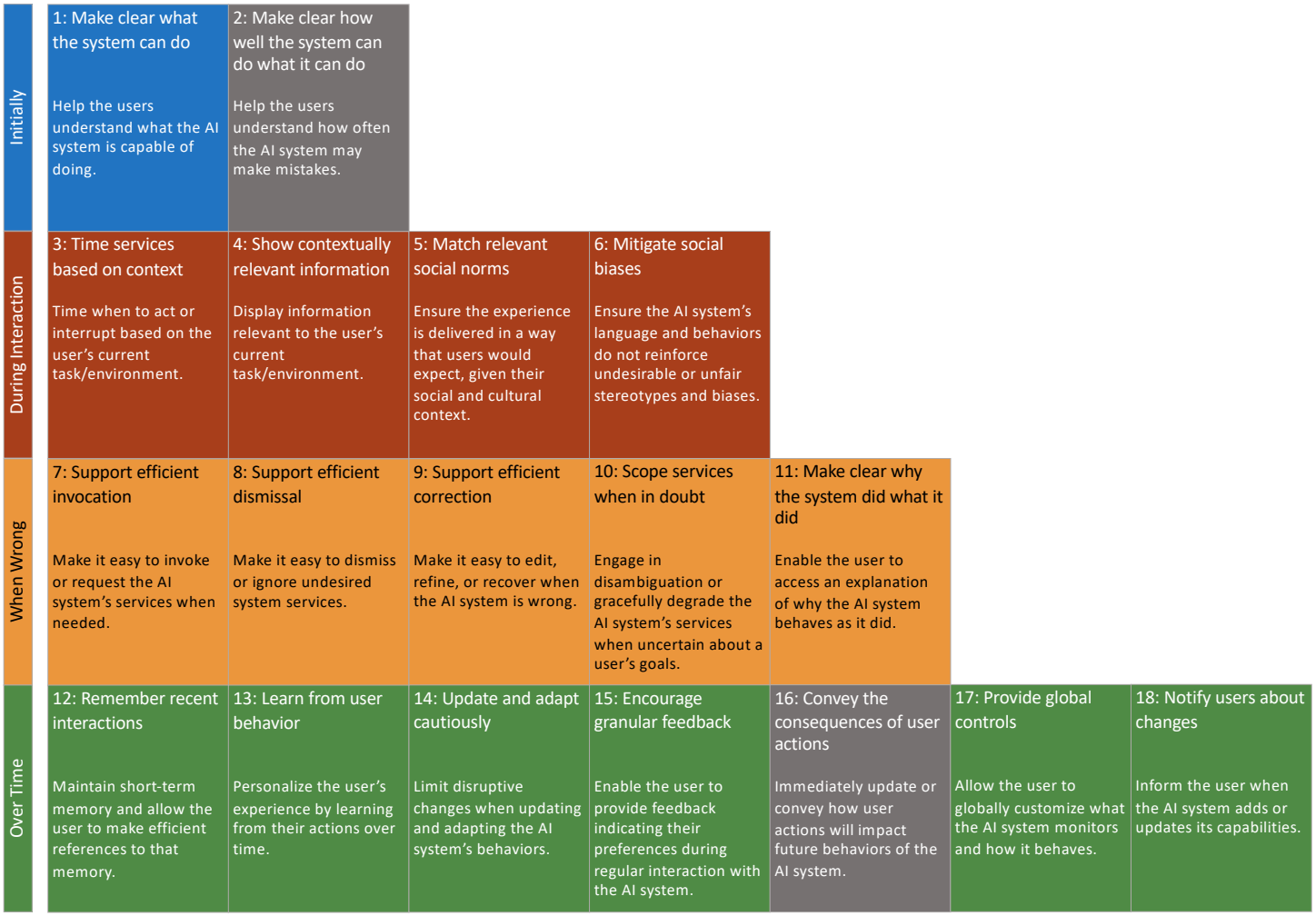


Fig. 2. Amershi et al.'s 18 guidelines for human-AI interaction [6]. For the 4 phases (left column), each guideline has a number, title, and brief description. Our analyses exclude the two guidelines' experiments (Guidelines 2 & 16, greyed out) which did not pass the manipulation check, as Li et al. also did [83].

2.2 Related Works

Inclusivity and related concepts like fairness in human-AI interaction can be thought of in two broad categories: 1) "under-the-hood" algorithmic inclusivity (i.e., how to detect and fix when *algorithms* behave inappropiately or unfairly for some groups of people) and 2) "over-the-hood" inclusivity of diverse users' *differing experiences* when they interact directly with AI products. There has been a host of literature for the former category (e.g., [15, 18, 50, 57, 71, 80, 133]), but this paper instead focuses on the latter category.

2.2.1 Investigations into individuals' problem-solving styles in Human-AI contexts. Because this paper addresses five problem-solving style ranges—risk, computer self-efficacy, motivations, information processing style, and learning style with technology (e.g., by process or by tinkering), we focus on those five styles here.

RQ1-Risk focuses on attitudes toward risk. There is a preponderance of research identifying risks associated with AI products (e.g., risks of inaccuracies, privacy loss, misaligned trust, etc.) In light of AI's risks, Cohen et al. [28] has posited that AI's explanations should consider both risk-averse and risk-tolerant users. Schmidt & Biessmann [106] also considered people's attitudes toward risk, classifying participants as either risk-averse or risk-tolerant using an incentivized gambling task. Their three conditions manipulated the level of transparency in the system, and they found that as system transparency increased, participants who were more risk-averse exhibited a more pronounced algorithmic bias (trusting the AI too much) than risk-tolerant participants, which they suggested was "a sign of blind trust in ML predictions that can be attributed to increased transparency" [106].

RQ2-AllStyles considers GenderMag's remaining four problem-solving style types, one of which is computer self-efficacy. Kulesza et al. [78] explicitly measured their participants' computer self-efficacy. Their work measured the change in their participants' computer self-efficacy as an outcome of explaining "why"-oriented eXplainable AI (XAI) approaches. They showed that scaffolding participants' experiences with "behind the scenes" training led to higher self-efficacy improvements and increased mental model soundness when compared to the participants without the scaffolding. Additionally, Jiang et al. [67] found that when their participants used an AI system, those with elevated self-confidence were less likely to accept the system's proposed solution, preventing them from being persuaded by the system in the presence of system discrepancies.

Another problem-solving style type considered in this paper is motivations. In this paper, motivations refers to reasons why people are interacting with the technology. Other researchers have also considered motivations in this sense, in the context of AI systems. For example, Shao & Kwon [110] identified four such motivations: people using AI products for the purposes of entertainment, companionship, functional utility, and dynamic control. They found that when participants interacted with AI products for functional utility and dynamic control, there was a positive relationship between their participants' satisfaction and these two motivations. Skjuve et al. [115] contributed six user motivations among respondents' responses for why they interacted with ChatGPT, including productivity, novelty, creative work, learning/development, and as a means of social interaction/support. Li et al. investigated motivations and user satisfaction in AI settings [82], and found a significant interaction between explanation type and motivations with respect to an AI product's persuasiveness (i.e., neighbor-rating explanation and hedonic motivation) for three types of explanations in a movie recommender system.

Other researchers have investigated a different meaning of motivation for users of AI products, namely how motivated a user was (i.e., how great their desire to participate, succeed, or win). For example, Eisbach et al. [39] found that when participants were more motivated to succeed at a task, they were more likely to intentionally process AI recommendations and explanations.

Visser [123] found that when participants perceived an AI as masculine, they were more motivated to play games with more intensity to win than when the AI was perceived as feminine. Some researchers have also investigated how motivated (i.e., likely) users are to interact with an AI product in the future. For example, Baek & Kim [11] found that when ChatGPT was perceived as creepy, participants were less motivated to interact with it in the future.

Another problem-solving style type considered in this paper is information processing style. As explained in Section 2.1, information processing style refers to the diverse ways that people gather information to solve problems. A closely-related concept is need for cognition [29, 65, 73, 104], which refers to the extent to which individuals are inclined towards effortful cognitive activities [22], ranging from those with a lower need for cognition to those with higher. In AI contexts, when considering need for cognition among participants, Dodge et al. [35] found that diverse needs for cognition came with different needs for explanation types and amount of explanation. Millecamp et al. [94] found that their participants with a higher need for cognition put in more effort to find the "best" AI recommendation, and Riefle et al. [103] found that their participants with higher need for cognition felt they understood an AI's explanations more than their counterparts.

One particularly pertinent inclusivity result while considering need for cognition in AI contexts was that of Buçinca et al. [17], researching how to reduce over-reliance on AI explanations. They found that adding cognitive forcing functions benefited only those with higher need for cognition, creating intervention-generated inequalities[3] because those with higher need for cognition have historically been a more advantaged group.

The fifth problem-solving style type considered in this paper is learning style, and some researchers have considered process-oriented versus tinkering-oriented learning styles in their analyses of AI products and development. For example, Nam et al. [96] used the GenderMag problem-solving style survey to investigate how 32 developers' information processing styles and learning styles affected their ways of using Large Language Model (LLM) developer tools. They leveraged three linear regression models, one for each of the investigated LLM features, and they used both participants' information processing style and learning style as explanatory variables for feature usage. For learning style, they found that process-oriented learners were significantly more likely to probe the LLMs with follow-up queries, whereas tinkering-oriented learners tended to jump directly into tinkering with the code after getting minimal direction from the LLM.

This paper differs from the above works by considering all five of these user problem-solving styles (as opposed to subsets of them), and comparing their effects on different versions of the same products. This paper also differs by showing how investigating these styles can help reveal actionable steps toward improving AI-powered technologies' inclusivity across diverse users.

2.2.2 Investigations of HAI inclusivity to diverse humans, from a demographic perspective. Particularly relevant to **RQ3-DemographicDiversity** are works in AI contexts that investigate gender differences while analyzing human data. van Berkel et al. [120] studied perceived fairness in AI recidivism and loan predictions and found that their participants who identified as men were significantly more likely to say that both systems were fairer than those who identified as women. de Graaf et al. [33] found that gender influenced participants' willingness to accept robotic technologies. Derrick & Ligon [34] also found gender differences on how likable the AI was, depending on how it behaved. Similarly, Joseph et al. [69] utilized a regression model to report on how awareness of AI impacted perception and utilization of AI tools. They found that increases in male students' awareness of AI resulted in an increase in their utilization of AI tools. However, increases in female students' awareness resulted in a *decrease* in their utilization of AI. Of particular interest

[3]According to Veinot et al., intervention-generated inequalities occur when a technological intervention disproportionately benefits a group of people who are already privileged in a particular context [122].

to this paper, Hu & Min [61] found that, although both men and women were concerned about the "watching eye" of AI, the participants who identified as women were more concerned about privacy violations than the men.

Another line of research particularly relevant to **RQ3-DemographicDiversity** are investigations of age differences in human-AI interaction. Gillath et al. found that older participants were significantly less likely to trust AI. Similarly, both Shahid et al. [108] and Martinez-Miranda [89] found that age impacted their participants' perceptions of AI-powered robots. (Their participants were much younger than ours; i.e., under 18 years old.) Other works have identified that regarding user experience, utilizing technologies like augmented reality and affective computing can help social robots become better companions for older adults [7] or that aging populations have been empowered by using artificial intelligence to personalize smart home interfaces [44]. Additionally, Zhou et al. [136] found that considering participants' age while adapting human-facing AI interfaces in the smart home domain led to improvements in usability for elderly participants. Other works related to user experience have similarly found how people's age might influence attitudes towards AI, impacting things like trust [45] and acceptance [68]. As with gender, some works have also discovered age differences in risky situations, such as Shandilya et al. [109], who interviewed 15 participants, all aged 60 or over, and their findings included user experience themes which may cluster by attitudes toward risk, such as the perceived annoyance when AI-enabled products deviated from expected behavior or data privacy threats.

Although these works considered participants' demographic diversity to find differences in user experience while interacting with AI via demographic dimensions such as gender, our work differs by instead providing actionable avenues for HAI practitioners through the alignment with participants' five GenderMag problem-solving style values with the participants' demographic differences. We establish these ties in Section 6.

2.2.3 Actionable recommendations for human-AI interaction. This investigation occurred within the context of Amershi et al.'s guidelines for human-AI interaction, but there are other ongoing efforts to support human-AI interaction. In January, 2022, Xu et al. [131] suggested that the set of design standards and guidelines supporting Human Computer AI-based systems was quite sparse, corroborating Yang et al.'s [134] observations that designing for quality HAI experiences remains a challenge for researchers and designers. Some of the challenges Yang et al. identified included assisting users in understanding AI capabilities, how to craft thoughtful interactions, and collaborating with AI engineers throughout the design process.

To address these challenges, other works have also proposed (and evaluated) sets of design principles for human-AI interaction. In 1999, Horvitz [59] identified 12 critical factors for mixed-initiative user interfaces, since humans would transition towards performing collaborative tasks with intelligent agents[4]. Some of the critical factors pointed towards the need to consider things like the uncertainty of a user's goals, as well as how to empower the user to infer ideal actions in light of costs, benefits, and uncertainties. Since then, researchers have proposed multiple principles towards aspects of human-AI interaction, such as Kulesza et al.'s [77] principles of explanatory debugging, with situational considerations like principles for explaining how an AI made its decisions in the event that is wrong. Other proposed principles focus on specific technologies, such as Ahmad et al.'s [5] focus on personality-adaptive conversational agents. Ahmad et al.'s work produced six principles, some of which suggest a need to design agents in such a way that they can support diverse users in a mental health setting.

Others have investigated methods of informing the design of human-AI interaction through guidelines. Wright et al. [130] survey guidelines from three major companies—Apple, Google, and

[4]Amershi et al. point out that 8 of their guidelines map to principles outlined in Horvitz's work.

Microsoft—and unify more than 200 guidelines into multiple categories. In their work, they classify the guidelines into categories such as initial considerations of AI, curating the models themselves, the deployment of the AI-powered system, and the human-AI interface. As Wright et al. point out, both Apple's [64] and Google's [47] guidelines are developed with the *developer* in mind, whereas Amershi et al.'s guidelines focus on how the design pertains to the *user*. The closest work to our own that does an empirical investigation of guidelines for human-AI interaction comes from the first user investigation of the Amershi et al. guidelines, reported in Li et al. [83]. The results, discussed in more detail in Section 3.3, found that in almost all of the experiments, participants preferred products which applied the guidelines, and applying the guidelines positively impacted participants' user experience.

All of these works investigated guidelines for human-AI interaction. Although our paper's context was applying a set of HAI guidelines, its foci are to present an *empirical approach to measuring* HAI's *inclusivity outcomes* in AI-powered systems, and to show that the approach can produce *actionable* results for HAI designers.

3 METHODOLOGY

To investigate our research questions, we performed 18 independent experiments, one for each of Amershi et al.'s 18 HAI guidelines [6] (listed earlier in Section 2.1). We used these experiments to perform two investigations. Investigation One, reported in Li et al. [83], investigated the impacts of violating/applying these guidelines. Investigation Two, which is the one we report in this paper, investigated potential disparities in the user experiences of participants with diverse problem-solving style values. Our investigation used GenderMag's five problem-solving style spectra —the spectrum of participants' attitudes toward risk, of their computer self-efficacy, of their motivations, of their information processing styles, and of their learning styles (by process vs. by tinkering).

To answer these research questions, we generated the following statistical hypotheses before data collection. For any dependent variable and any of the five problem-solving styles, our statistical hypotheses between applications (app) and violations (vio) of any guideline were:

$$H_0 : \mu_{app} - \mu_{vio} = 0$$
$$H_A : \mu_{app} - \mu_{vio} \neq 0$$

3.1 Study Design

The experiments' context was productivity software, such as document editors, slide editors, search engines, email applications, and spreadsheet applications. Each experiment was a 2x2 factorial experiment, where each factor had two levels.

The first factor, the "guideline adherence" factor, was within-subjects, and the factor's levels were "guideline violation" and "guideline application". For any one guideline's experiment, the "guideline violation" condition violated that particular HAI guideline; for example, in Guideline 1's experiment (make clear what the system can do), the "guideline violation" did *not* make clear what the system can do. Similarly, in Guideline 11's experiment (make clear why the system did what it did), the "guideline violation" did *not* make clear why the system did what it did. In contrast, the "guideline application" level applied each guideline; for example, in Guideline 1's experiment (make clear what the system can do), the "guideline application" condition added clarifying information about what the system can do.

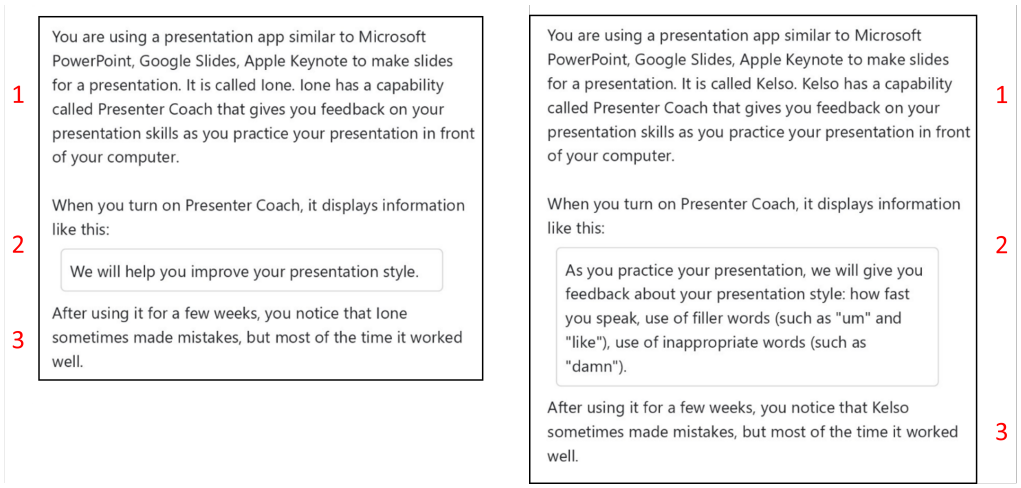
The second factor, the "AI performance" factor, was between-subjects. This factor's levels were "AI optimal" and "AI sub-optimal". In the "AI optimal" level, the AI sometimes made mistakes but worked well most of the time, whereas in the "AI sub-optimal" level, the AI sometimes made mistakes and sometimes worked well.

In each experiment, both the product that violated the guideline and the product that applied it were represented by vignettes, as in several other works in human-AI interaction [1, 32, 81, 85, 91]. The vignettes were developed in two phases: in the first phase, two researchers went through an iterative brainstorming process, where they independently thought about how the 18 guidelines might show up in productivity software, drafting between 5–8 interaction scenarios for each guideline. Then, the researchers collaborated to review, rewrite, and sometimes replace the scenarios. In the second phase, the researchers adhered to Auspurg et al.'s [10] best practices to make the vignettes simple, clear, and realistic. In cases where the interaction description was not understandable through text, images were used to promote understandability. Before deploying the study, each vignette went through two rounds of piloting. In the first round, each vignette received feedback from 7 HCI researchers not familiar with the project, and changes were made based on that feedback. In the second round, we piloted the updated vignettes on Amazon MTurk with five participants per vignette; no issues were identified from this second pilot. Each of the final vignettes was composed of three parts: (1) a product/feature introduction; (2) a description of what the AI feature did; and (3) a summary of how well the AI performed.

Figure 3 provides an example of the two vignettes from the experiment for Guideline 1 ("Make clear what the system can do"). In first part, the only difference between the two conditions' vignettes was in the name (Ione and Kelso), generic names given to each product to distinguish them from each other and to avoid the influence of prior familiarity with a real product. The second part manipulates the "guideline adherence" factor. In Figure 3a, part 2 states:

"*We will help you improve your presentation style*"

without giving specific examples or details, thus violating the guideline by *not* making clear what the system can do. In contrast, Figure 3b's part 2 applies the guideline to make clear exactly what the system can do, stating:



(a) The `Violation AI product`. (b) The `Application AI product`.

Fig. 3. Guideline 1's ("make clear what the system can do") two vignettes. Each vignette had three components: (1) A product and feature introduction, describing what the product was and what it did, (2) the behavior description of the manipulated AI feature that differentiated the guideline's violation from its application, and (3) the AI performance description. Note that participants were never exposed to the concept of guideline violations or applications; instead, they saw only generic names (Ione & Kelso).

Dependent Variable Name	Dependent Variable Wording	Reverse-Coded?
I would feel in control	"*I would feel in control while using the product.*"	
I would feel secure	"*I would feel secure while using the product.*"	
I would feel inadequate	"*I would feel inadequate while using the product.*"	✓
I would feel uncertain	"*I would feel uncertain while using the product.*"	✓
I would feel productive	"*I would feel productive while using the product.*"	
I perceived it as useful	"*I would find the product useful.*"	
I would be suspicious	"*I would be suspicious of the intent, action, or outputs of the product.*"	✓
It would be harmful	"*I would expect the product to have a harmful or injurious outcome.*"	✓
I find the product reliable	"*I would expect the product to be reliable.*"	
I would trust the product	"*I would trust the product.*"	

Table 2. The 10 dependent variables regarding users' perceived feelings [14], usefulness [102], and trust [66] questions. Participants answered these 7-point agreement scale questions for both the Violation product *and* the Application product, which they saw in a random order. We indicate the reverse-coded questions (✓) – Feel Inadequate, Feel Uncertain, Suspicious, Harmful – which Li et al. also did. As such, they became: Feel Adequate, Feel Certain, Not Suspicious, Not Harmful. Participants saw *only* the wording shown in the "Dependent Variable Wording" column.

"*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').*"

We will refer to the vignette that violated the guideline as the `Violation AI product`. Similarly, we will refer to the vignette that applied the guideline as the `Application AI product`.

Table 2 lists the questions that the participants responded to for both the `Violation AI product` and the `Application AI product`. These dependent variables gather information about different dimensions of participants' user experiences. The first five questions (control, secure, inadequate, uncertain, and productive) follow Benedek & Miner's [14] approach of measuring end users' feelings in user experience. Perceived usefulness was taken from Reichheld and Markey [102] and has been known to relate to acceptance and use of AI-infused systems. The last four questions (suspicious, harmful, reliable, and trust) came directly from Jian et al. [66], who focused on scales for trust in automated systems. The answer to each question was an agreement scale, ranging from "extremely unlikely" (encoded as a 1) to "extremely likely" (encoded as a 7).

3.2 Participants & Procedures

1,300 participants were recruited from Amazon Mechanical Turk (MTurk), a popular crowdsourcing platform. To ensure quality data, participants had to meet certain performance criteria on MTurk before they could participate in the study, such as having at least 100 approved human intelligence tasks (HITs) and having above a 95% acceptance rate on the platform. Additionally, participants had to be located in the USA and be at least 18 years old. After workers accepted the HIT, they were presented with an IRB consent form, and then answered three screening questions. The first two asked about their familiarity with productivity software and the last confirmed that they were above the minimum age requirements. Upon completion of the screening survey, participants were provided $0.20.

Once participants had completed the screening survey, they were randomly assigned to one (and only one) of the 18 experiments, one for each guideline. First, participants randomly saw either the `Violation AI product` or `Application AI product`, such as the example provided in Figure 3[5]. Participants then responded to the user experience questions shown in Table 2, asked in a random order. Once participants completed their responses for the first AI product, they saw the second product and answered the same user experience questions in another random order.

Once participants had seen both products and answered the user experience questions for each, they were asked to select which product they preferred and explain why they preferred it. As detailed in Li et al. [83], 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. Other team members conducted spot checks to verify the qualitative coding. Participants were then asked two manipulation check questions[6], one closed- and one open-ended. The closed-ended manipulation check asked participants whether or not they agreed with text that mirrored the guidelines themselves (e.g., "make clear what the system can do", "make clear why the system did what it did", etc.). For example, if participants in Guideline 1's experiment agreed that the `Application AI product` made clear what the system can do, and they disagreed with the statement for the `Violation AI product`, then they passed the manipulation check. The open-ended manipulation check asked participants to "...briefly describe the differences between Kelso and Ione" (the fictitious names randomly assigned to the `Violation AI product` and `Application AI product`). The open-ended answers were qualitatively coded to check whether or not each participant had successfully perceived the experimental manipulation.

Participants then filled out a questionnaire with their demographic data, including their age, self-identified gender, race, highest education level, and field of employment[7]. They also filled out the problem-solving style questionnaire (Section 3.4) and were paid a bonus of $5 for completing the experiment.

3.3 Investigation One Results Summary

As mentioned in Section 1, Investigation One, reported in Li et al. [83], compared user experience outcomes of AI products that had applied the guidelines against AI products that had not. That investigation's measures were generalized eta-squared (η^2) effect sizes for each of the dependent variables in each of the experiments.

The primary takeaway from Investigation One was that, for most of the guidelines, participants perceived the `Application AI products` as more useful and as providing better user experiences than the `Violation AI products` did. Figure 4 shows thumbnails of their results for each guideline's experiment. The more color-filled each thumbnail, the larger the positive effect sizes were for that guideline's experiment. For example, G3's thumbnail shows significant differences with small or medium effect sizes on most of the HAI-UX aspects measured. G6's experiment produced particularly strong results. Its thumbnail is almost filled with color, indicating that G6's experiment produced significant differences on all HAI-UX measures, with medium or large effect sizes for all but one.

In addition, Investigation One's analysis informed two aspects of Investigation Two's analysis. First, Investigation One's analysis revealed that 2 of the 18 experiments failed the manipulation checks (Section 3.2)—the experiments for Guideline 2 and Guideline 16—and as such were dropped

[5]Participants were not told that one product violated/applied a guideline and one did the opposite.

[6]In experimental design, a manipulation check is a test used to determine the effectiveness of a manipulation in an experimental design. Passing manipulation checks indicates that the manipulation in an experimental design was effective, whereas failing manipulation checks indicates that it was not.

[7]Counts reported in Appendix C

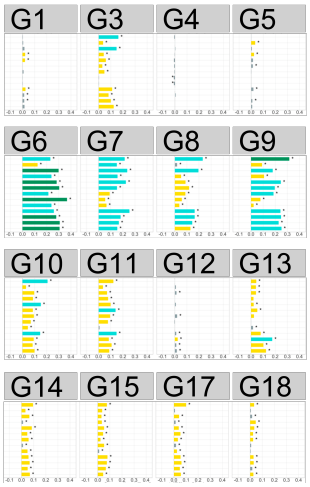


Fig. 4. Thumbnails of Investigation One's results for each guideline's experiment. More color indicates larger effect sizes. *: difference was statistically significant. See Li et al. [83] for full details.

from Investigation One. Thus, our investigation also drops those two experiments, which leaves a total of 1,043 participants across the remaining 16 experiments. Second, Investigation One's analysis of these remaining 16 experiments revealed that the AI optimality factor (Section 3.1) was significant in only one of these experiments. This resulted in Investigation One dropping this experimental factor, and we do the same for Investigation Two.

3.4 Investigation Two (Current Investigation) Data Analysis

This paper's investigation analyzes the same independent experiments' data from a new perspective: the inclusivity that the violation vs. application AI products afforded diverse participants. Specifically, we consider diversity in terms of participants' diverse problem-solving styles (**RQ1-Risk** and **RQ2-AllStyles**) and their diverse gender/age demographics (**RQ3-DemographicDiversity**).

To collect demographics, we used a questionnaire asking participants their gender identity and age group. To collect participants' diverse problem-solving styles, we used the GenderMag facets survey [55], a validated survey that measures participants' values of the five GenderMag problem-solving style types enumerated earlier in Section 2.1, termed "facets" in GenderMag publications. Each problem-solving style type has multiple Likert-style questions that run from *Disagree Completely* (encoded as a 1) to *Agree Completely* (encoded as a 9), a few examples of which are shown in Table 3. For example, using this instrument, if one participant answers the first question (top row) closer to *Agree Completely* than a second participant, the first participant is considered to be more risk-averse than the second participant. 27 of the 1043 participants failed at least one attention check in the problem-solving style survey, leaving 1,016 participants for this investigation. Appendix A lists the full questionnaire, including the attention checks.

The GenderMag survey has previously been validated in multiple ways. Hamid et al. [55] summarize the six-step validation process; among the steps were literature searches, multiple statistical analyses, demographic validation, and problem-solving style validation. Particularly relevant to this paper was Guizani et al.'s [53] participant validation of the problem-solving styles the survey captures. In that study, participants took the survey, then spoke aloud throughout problem-solving tasks. Participants' in-the-moment verbalizations when problem-solving validated their own questionnaire responses 78% of the time, a reasonably good measure of consistency [48].

For this Problem Solving Style:	Sample Question:
Attitude toward risk	"*I avoid using new apps or technology before they are well-tested*"
Computer self-efficacy	"*I am able to use unfamiliar technology when I have seen someone else using it before trying it.*"
Motivations	"*It's fun to try new technology that is not yet available to everyone, such as being a participant in beta programs to test unfinished technology.*"
Information processing style	"*I always do extensive research and comparison shopping before making important purchases.*"
Learning style (by process vs. by tinkering)	"*I enjoy finding the lesser-known features and capabilities of the devices and software I use.*"

Table 3. Examples of questions from the validated problem-solving style survey (full survey in Appendix A).

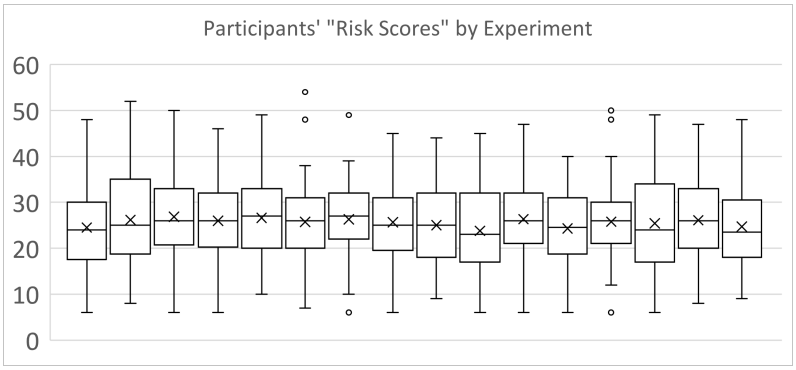


Fig. 5. Participants' risk scores (y-axis) for each experiment (x-axis). "x"s mark the means, horizontal lines mark the medians. Participants above the median are more risk-averse than their peers below the median.

To score a participant's problem-solving style values, we summed up that participant's responses to the risk questions; then the self-efficacy questions, and so on. Each sum is the participant's "score" for that problem-solving style. Comparing these scores reveals a participant's placement in that problem-solving style type compared to others in the same peer group, such as among computer science professors, or among residents of eldercare facilities, etc.; in our case, the peer group is the adult productivity software users who participated in the study.

These scores formed 16 distributions, one for each experiment (e.g., see Figure 5 for the risk score distributions). Using each experiment's median[8], which is robust against outliers, we then defined participants as being either more risk-averse than their peers (i.e., above the median) or more risk-tolerant, and similarly for the other four problem solving styles[9].

To analyze the dependent variables for each of the 16 experiments, we used t-tests after ensuring that the assumptions held, as follows. To investigate inclusivity (Section 4), we compared within-subjects using paired t-tests, treating each `Violation AI product` as a "before" and `Application AI product` as an "after". As Table 4 shows, each of the 16 experiments had over

[8]This approach has also been used in other measures of problem-solving styles, such as *need for cognition* [17].
[9]These classification rules are detailed in Table A7 in Appendix A.

Experiment	G1	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G17	G18
Risk-Averse Participants	26	31	29	32	28	26	26	26	31	31	35	27	31	29	27	32
Risk-Tolerant Participants	31	37	36	37	36	32	35	32	34	32	34	30	37	35	36	35
Total	57	68	65	69	64	58	61	58	65	63	69	57	68	64	63	67

Table 4. The number of risk-averse vs. risk-tolerant participants (rows) per experiment (columns). The group sizes were similar, with the smaller group at least 43% of the total in every experiment.

30 participants, suggesting normality of every experiment by the Central Limit Theorem.[10] In addition, in cases where the sample size fell beneath 30, we used Shapiro-Wilk tests to validate that the underlying reference distribution was not significantly different than normal (i.e., $p \geq .05$).

Satisfying these assumptions indicated that the above t-tests were appropriate analysis techniques for these data. Each of the 16 experiments were designed with pre-planned hypotheses for each dependent variable, so we do not report statistical corrections in the body of this paper. As other researchers [8, 99] point out, statistical corrections (e.g., Bonferroni, Holm Bonferroni, Benjamini-Hochberg, etc.) are necessary only if "...a large number of tests are carried out without pre-planned hypotheses" [8, 9]. Still, we recognize that not all readers may agree with this choice, so we also show all the Holm-Bonferroni corrections [58] in Appendix D.

4 RESULTS: WHAT PARTICIPANTS' RISK STYLES REVEALED

RQ1-Risk considers the 16 pairs of AI products described in Section 3—one violating a guideline and its counterpart applying that guideline—and how the two differed in their inclusivity of risk-diverse participants' HAI experiences. (We will generalize beyond risk in Section 5.)

In this paper, we measure whether/how applying a guideline to an AI product *changed* the product's inclusivity toward some particular group of participants. For any user experience dependent variable in an AI product, we will say the HAI-UX is *more (less) inclusive* to a group of participants if the Application AI product's result for that variable are *significantly higher (lower)* than the Violation AI product's for *those* participants.[11]

To answer this question, we performed an in-depth analysis of all HAI-UX measurements' inclusivity by considering participants' attitudes toward risk. HAI-UX inclusivity could change in only four possible ways: (1) inclusivity changes for both the risk-averse and risk-tolerant participants, (2) inclusivity changes for neither of them, (3) inclusivity changes for the risk-averse only, and (4) inclusivity changes for the risk-tolerant only. As Table 5 shows, instances of all of these categories occurred.

Table 5 also shows that the risk results fell mainly in categories (1) and (2) above. Perhaps most important, the table shows that whenever applying a guideline produced a change in inclusivity, it was almost always a *positive* change for at least some risk-attitude group of participants—without loss of inclusivity for the other group.

Result #1: *Following the guidelines usually led to inclusivity gains.* Applying these guidelines led to 115 (75+13+27) inclusivity gains for either or both risk groups, and only 1 inclusivity loss.

[10] "The Central Limit Theorem asserts that averages based on large samples have approximately normal sampling distributions, regardless of the shape of the population distribution" [101]. By convention, the rule of thumb for a large enough sample is often considered to be >=30 [101].

[11] Recall that, because this inclusivity measure is within-subject, we used paired t-tests to measure HAI-UX in the before (`Violation AI product`) vs. after (`Application AI product`) versions of the AI products.

Dependent Variable	G01	G03	G04	G05	G06	G07	G08	G09	G10	G11	G12	G13	G14	G15	G17	G18	Total Gains
Feel In Control		A T			A T	A T	A T	A T	A T	A T			A T	A T	A T	A	11
Feel Secure	T	T		A	A T	A T	A T	A T	A T	A T		T		T		A T	12
Feel Adequate		T		T	A T	A T	T	A T	A	T		T	A	T			11
Feel Certain		T			A T	A T	A T	A T	A T	A T			A T	A T	A T	T	11
Feel Productive	A	A T			A T	A T	A T	A T	A T	A T	T	T	A T	A	A T	T	14
Perceived Useful		A T			A T	A T	A T	A T	A	A T	T	A T	A T	A T	A T	T	13
Not Suspicious			T		A T	A T	A T	A T	A T	A T			T	A T			8
Not Harmful		T			A T	A T	T	T	A T				A T	A	A		9
Product Reliable	A	A T			A T	A T	A T	A T	A T	A T	A	T	A	T	T	T	14
Trust Product		T			A T	A T	A T	A T	A T	A T		A T	A	A T	A T	T	12

Table 5. Risk results summary, by guideline experiment (columns) and dependent variable (rows). Reading columnwise reveals the guidelines with the most inclusivity gains, and reading rowwise reveals the dependent variables with the most inclusivity gains (rightmost column). There was also one inclusivity loss.
Total occurrences: 75 inclusivity gains for both risk-Averse and risk-Tolerant (A T); 44 without inclusivity gains/losses (blank); 13 inclusivity gains for risk-Averse only (A); 27 inclusivity gains for risk-Tolerant only (T); 1 inclusivity loss for risk-Tolerant only (T). Total possible: 160.

4.1 When everybody gained: More inclusivity for both the risk-averse and the risk-tolerant

At first glance, it may appear that the "when everybody gained" category of results is a natural consequence of the overall success rates shown by Investigation One. For example, many of the experiments that produced strong positive results in Investigation One also did so in Investigation Two, as with G6–G9. Still, even the G6–G9 experiments reveal relationships between outcomes and perceptions of risk that shed new light on the *why*s of these results.

For example, consider Guideline 8. In this experiment, the AI-powered feature was a design helper to automatically provide design suggestions for alternative layouts in a presentation application. In the `Violation AI product`'s vignette, the feature's behavior was: *"You are working on a slide and Design Helper pops up, showing you some design suggestions. You do not need any design help at this time, but there is no way to hide the design suggestions."* In contrast, the `Application AI product`'s vignette started out the same, but its last sentence was: *"You do not need any design help at this time, so you click on a button visible on screen to hide the design suggestions."*

The second row of Table 6 shows one of Guideline 8's outcomes,[12] with the risk-averse participants' suspicions of the `Violation AI product` (left hatched boxplot) significantly worse than their suspicions of the `Application AI product` (left clear boxplot)[13] ($t(25) = 3.2354$, $p = .003$, $d = .648$)[14]. Likewise, the risk-tolerant participants also were significantly less suspicious of the `Application AI product` than of the `Violation AI product` ($t(34) = 3.0020$, $p = .005$, $d = .507$), as shown in the right boxplots.

Yet, despite their agreement on these outcomes, participants' free-text remarks showed that their reasoning differed with their attitudes toward risk. The risk facet is nuanced—it includes aversion/tolerance risks with privacy/security, of producing low-quality work, of wasting too much time, of having trouble with the product, etc. In Guideline 8's experiment, about a fourth (14/61) of the participants' comments focused on the second of these, the risk of low-quality work.

[12] Appendix D provides boxplots of results for all dependent variables in all experiments.
[13] Recall from Table 2 that the "suspicious" dependent variable was one of the variables that reverse-coded for presentation clarity, so that more positive outcomes were always higher on the scales.
[14] This result was derived using Student's t-test, although these data are not continuous. We validated all results in this paper using Wilcoxon signed rank test, and the non-parametric results agreed with our own 97% of the time.

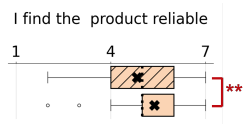
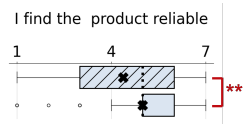
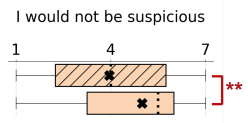
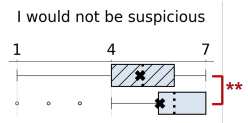
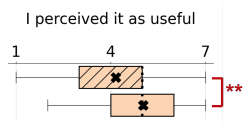
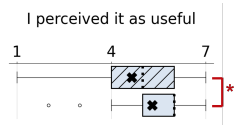
Guideline	Risk-Averse Inclusivity	Risk-Tolerant Inclusivity
3: Time services based on context		
8: Support efficient dismissal		
14: Update and adapt cautiously		

Table 6. Inclusivity gains: A few examples of the 75 times inclusivity improved for both risk-averse and risk-tolerant participants. E.g., in Guideline 8's experiment (second row), both the risk-averse and the risk-tolerant participants were significantly more suspicious of the `Violation AI product` (top, hatched boxplots) than of the `Application AI product` (bottom, clear boxplots). x = average, ¦ = median. * = p<.05, ** = p<.01, *** = p<.001, NS=not significant.

This focus on risk of low-quality work was especially true of risk-averse participants. 31% (8/26) of this experiment's risk-averse participants wrote about preferring the increased control they had over their work quality with the `Application AI product`.

> G08-1921-risk-averse: "*...very convenient and still make me feel very much in control of my choices.*"
>
> G08-3619-risk-averse: "*I don't trust [`Application AI product`]..., but the fact I can turn the feature off lets me be in more control.*"

Even the more risk-tolerant were worried about this type of risk, and 17% (6/35) of these participants expressed the same sentiments. However, for these more risk-tolerant participants, annoyance and frustration also figured prominently in their reasoning (26%: 9/35), compared to only 1 risk-averse participant expressing this sentiment.

> G08-2831-risk-tolerant: "*Because I can get rid of the content that might ... influence me to do something stupid. If I am going to do something stupid it will be my idea.*"
>
> G08-3681-risk-tolerant: "*...[`Application AI product`] would allow me more freedom, and be less annoying with its suggestions, even when they are wrong.*"
>
> G08-2627-risk-tolerant: "*Without an option to turn off an unnecessary feature, I would be extremely frustrated...as it would be a severe distraction... never would I use [`Violation AI product`]...*"

Comments like these, when coupled with the risk-averse and risk-tolerant participants' feeling both significantly more in control and less suspicious of the `Application AI product`, suggest relationships between an expectation of risk and four particular HAI-UX inclusivity outcomes. As Table 5 shows, across all 7 experiments where the `Application AI product` gained inclusivity for both the risk-averse and risk-tolerant participants' (not)-suspicious outcome (row 7), it also gained inclusivity for their certainty (row 4), control (row 1), and trust outcomes (row 10).

Result #2: *Suspicion, control, trust, and certainty changed in tandem, for both risk groups.* In all experiments, every inclusivity gain in (1) (not-)suspicious for both the risk-averse and risk-tolerant was coupled with an inclusivity gain in all three of (2) in-control and (3) trust, and (4) certainty.

This result provides insight into why the five experiments that gained the most inclusivity across the risk spectrum participants—G6, G7, G8, G9, and G15—did as well as they did. What these five *guidelines* have in common is that they all give users more control over the product. What their five *experiments* have in common (from Table 5) is that, in all of them, the `Application AI product` gained inclusivity in all four of the above variables:

Result #3: *Giving users control mattered for both risk groups.* The five experiments with the most inclusivity gains across risk-diverse participants were those whose guidelines increased users' control over the AI products.

4.2 When nobody gained: No inclusivity improvements for either risk group

Not all the results were as positive for diversity. Some changes did not change inclusivity outcomes for either of the two groups, measured in this paper as no significant difference in HAI-UX inclusivity for either the risk-averse or the risk-tolerant participants between the `Violation AI product` and `Application AI product`. This was the second-most prevalent category, occurring 44 times across 10 experiments.

Consider Guideline 4's ("show contextually relevant information") results; examples are in Table 7. Guideline 4's experiment produced 9 instances of the "nobody gained" category. In that experiment, the application was a document editor, and the AI-powered feature was an acronym explainer. The `Violation AI product` violated the guideline: *"When you highlight an acronym to see what it stands for, [Violation] shows you a standard list of possible definitions taken from a popular acronym dictionary."* In contrast, the `Application AI product`: *"When you highlight an acronym to see what it stands for, [Application] shows you definitions that are used in your workplace and pertain to the topic of the current document."*

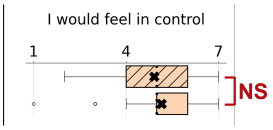
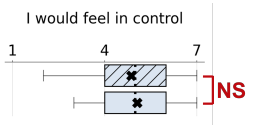
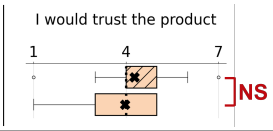
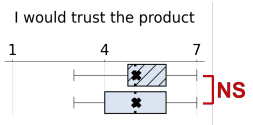
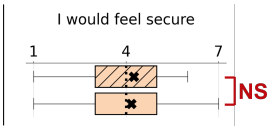
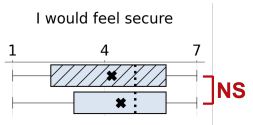
Guideline	Risk-Averse Inclusivity	Risk-Tolerant Inclusivity
1: Make clear what the system can do		
4: Show contextually relevant information		
12: Remember recent interactions		

Table 7. Inclusivity unchanged: Examples from the 44 instances in which HAI-UX inclusivity did not significantly change for either the risk-averse or the risk-tolerant participants.

In some ways, the participants' reasoning for their unchanging responses to the `Violation AI product` vs. the `Application AI product` echoed those of the previous subsection, namely wanting to avoid the risk of low-quality work. As in the previous section, this reasoning was especially common among the risk-averse participants (34% = 10/29), although 22% (8/36) of the risk-tolerant also used it. However, whereas in the previous section participants gave this risk

as an *asset* of the `Application AI product`, in this section they gave it as a *liability* of the `Application AI product`.

> G4-4098-risk-averse: "*[`Violation AI product`] may make mistakes ... but its use of a generic dictionary makes it easier to recognize mistakes... With [`Application AI product`], I would be more likely to miss mistakes.*"
>
> G4-3799-risk-averse: "*...if [`Application AI product`] 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.*"

Guideline 4 also raised privacy concerns among some participants:

> G4-3905-risk-averse: "*... I would be nervous that [`Application AI product`] is pulling data from things like my other software and my browsing history.*"
>
> G4-3947-risk-tolerant: "*I don't like the idea of [`Application AI product`] taking definitions from my workplace. It makes me worry I'm being listened to...*"

In the "everybody gained" category (previous section), the five most inclusive guidelines across the risk spectrum revealed a relationship among risk-inclusivity and trust, control, certainty, and (not)-suspicious. The five *least* inclusive guidelines as per Table 5—G1, G4, G5, G12, and G13—show that the relationship persisted in the "nobody gained" category. None of G1, G4, G5, and G12 produced any inclusivity gains for any of these interrelated variables; and G13 showed only two such gains. These results not only confirm **Result #2**, but also provide a complement to **Result #3**:

Result #4: *Not having user control mattered, for both risk groups.* None of the five guidelines showing the fewest risk-inclusivity gains offered increased user control over the AI products.

4.3 Selective inclusivity: who gained, who did not, and why?

The third and fourth categories of HAI-UX inclusivity changes were inclusivity gains for the risk-averse participants only or the risk-tolerant participants only. Neither category was very large, with 13 and 27 total instances, respectively. Table 8 shows a few examples.

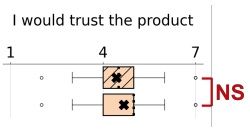
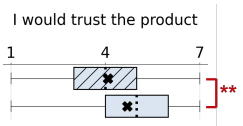
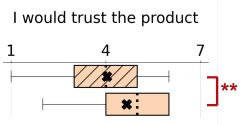
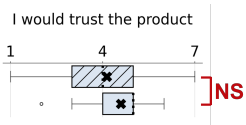
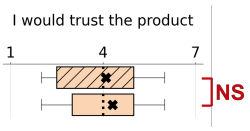
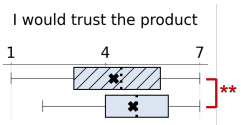
Guideline	Risk-Averse Inclusivity	Risk-Tolerant Inclusivity
3: Time services based on context		
14: Update and adapt cautiously		
18: Notify users about change		

Table 8. Example inclusivity gains by the risk-averse only (e.g., the Guideline 14 experiment) or by the risk-tolerant only (e.g., the Guideline 3 and 18 experiments).

Despite the relatively small totals, the fourth category, that of bringing gains to the risk-tolerant only, reveals a unique pattern shared by three experiments—Guideline 3's, Guideline 13's, and Guideline 18's. As a column-wise reading of Table 5 shows, in these three experiments, inclusivity gains for the risk-tolerant participants abounded, but the risk-averse participants rarely gained.

Guideline 3's experiment offers a case in point. In that experiment, the `Application AI product` provides services only when it decides the user's current task/environment would benefit. The `Application AI product`'s vignette applied this guideline by stopping email notifications "*...when you are busy.*"

As Table 5 shows, for Guideline 3's risk-tolerant participants, the `Application AI product` showed inclusivity gains on every dependent variable except one for the risk-tolerant participants. However, only four of these gains extended to the risk-averse participants.

Why such differences? For Guideline 3's experiments, the risk-averse participants' concerns about risks to their work or their privacy appeared to outweigh the benefits of fewer notifications, whereas for the risk-tolerant, the balance seemed to tip the other way. For example, 26% (8/31) of the risk-averse participants explicitly brought up concerns about these risks, but only 11% (4/37) of the risk-tolerant did.

> G3-3504-risk-averse: "*...Also, I wouldn't be sure that [`Application AI product`] would be able to accurately qualify my activities.*"
>
> G3-3054-risk-averse: "*[`Application AI product`]... would have to be able to monitor your online activity...that would be a little invasion of privacy...*"

A risk-oriented commonality among these three experiments lies in what these AI products were actually doing. All three of these `Application AI products` learn from the *user's own data*, as opposed to learning mainly from huge datasets mostly consisting of *other people's* data. Specifically, both Guideline 3's and Guideline 13's AI products involved learning from the user's own context and behaviors, and Guideline 18's AI product involved learning from that user's emails. In the latter case, the AI product also moved that user's emails around, adding the risk that the user might not find some of their emails later.

Two other `Application AI products`, Guideline 4's and Guideline 12's, also had this attribute. Guideline 4 involved learning from the user's contexts, and Guideline 12 learned from the user's recent interactions. Most of these two products' outcomes were in the "nobody gained" category: the risk-averse were very uncomfortable with these products, and even the risk-tolerant saw too many risks (e.g., recall the discussion of Guideline 4's experiment in the previous subsection).

These five guidelines' `Application AI products` were the only ones with this attribute. And for these five experiments, the risk-averse participants hardly ever experienced any inclusivity gains (from Table 5).

Result #5: *Learning from "my" data mattered.* Whenever the `Application AI products` learned from participants' own data, inclusivity gains for risk-averse participants were rare.

4.4 The risk results and actionability

The AI products in these experiments were designed to isolate effects of applying vs. violating each guideline. However, if they were real products for sale, the products' owners would probably hope to make each product as well-received by as many of its customers as possible.

The risk results provide actionable ideas for such product owners, for seven of these AI products. Figure 6 points out which products those were.

The green boxes in the figure mark those products: they are G1's, G3's, G4's, G5's, G12's, G13's, and G18's. For example, Section 4.1 revealed how lack of user control affected some of the low-performing AI products (e.g., G1's, G4's, G5's, G12's, and G13's). An actionable implication for these products is that those products would improve by offering users more control. As another example, Section 4.3 revealed the sensitivity the risk-averse participants had to products that potentially did "too much" with *their* data (e.g., G3's, G4's, G12's, G13's, and G18's products). One actionable idea for those products would be to provide information on what else the user's personal data

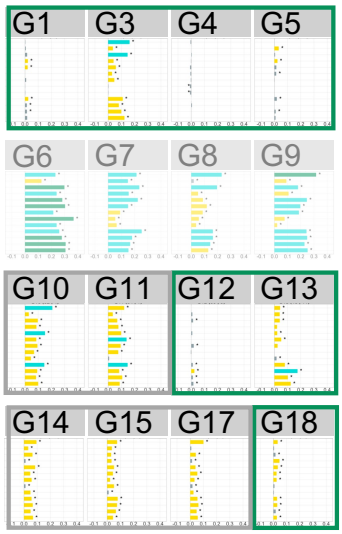


Fig. 6. Revisiting Investigation One's results, annotated with Risk implications. Green boxes: Risk results point to ways to improve those AI products. Gray boxes: AI products we might like to improve, but Risk results were "too inclusive" to help. Unboxed AI products were already very well-received.

are used for and how long these data are stored. More generally, results like these suggest that a way to improve AI products favored by only the risk-averse participants or only the risk-tolerant participants, is to attend to risk-oriented attributes of the product that were not tolerated well by participants at that end of the risk spectrum.

Result #6: *Some risk results were actionable.* For the seven `Application AI products` associated with G1, G3, G4, G5, G12, G13, and G18, the Risk results provided actionable ideas for further improving the HAI-UX those products offered.

5 RESULTS: BEYOND RISK—THE OTHER FOUR PROBLEM-SOLVING STYLES

Section 4 considered only one type of problem-solving diversity, namely participants' diverse attitudes toward risk. We now turn to **RQ2-AllStyles**, which asks "How inclusive are such products to users with diverse values of GenderMag's other four problem-solving styles?" Although space does not permit an in-depth analysis of each remaining problem-solving style—motivations, learning style, computer self-efficacy, and information processing style—we summarize in this section whether and how the Risk results of Section 4 generalize to analogous results. If they do, we also consider whether those new results add anything to our understanding of the user experiences the AI products offer to diverse problem-solvers. Full analyses for all of these styles are in the Appendices.

Recall from Table 1 that GenderMag uses two personas, "Abi" and "Tim", to identify the distinguished endpoints of each of GenderMag's five problem-solving style types. As per the table's definitions, we classify participants as more "Abi"-like if they had any of the following problem-solving style values: more risk-averse, lower computer self-efficacy, task-oriented motivations for using technology, had a comprehensive information processing style, or were a process-oriented learner. Participants nearer the opposite endpoint of these problem-solving spectra are classified to be more "Tim"-like; i.e., more risk-tolerant, had higher computer self-efficacy, had tech-oriented

motivations, had a more selective information processing style, or learned more by tinkering. As in other persona research [3], we use these persona names for two reasons: 1) to provide a vocabulary for an associated collection of traits and collection of traits (recall Section 2), and 2) Using the "Abi" and "Tim" vocabulary helps emphasize which of the "distinguished endpoints" of each problem-solving style type tend to co-occur, helping to keep clear which problem-solving value belongs to the underserved population.

As the upcoming Tables 9, 10, 11, 12, and earlier Table 5 show, the first result of these analyses was very good news for most of the `Application AI products`. As was also true of risk diversity results, whenever applying an `Application AI product` produced a change in inclusivity, it was almost always a *positive* change for at least some <problem-solving value> group of participants—without loss of inclusivity for the other group. For example, as Table 9 shows, whenever applying a guideline produced a gain for either the task-motivated group or the tech-motivated group, it almost never produced an inclusivity loss for the other group, with only one exception.

Result #7: *Following the guidelines usually led to inclusivity gains—for every one of these five problem-solving styles.* Applying these guidelines led to 115, 116, 116, and 116 inclusivity gains, respectively, for motivations-diverse, learning-style-diverse, self-efficacy-diverse, and information-processing-diverse participants; with only 1 or 2 inclusivity losses for any of these types of problem-solving diversity.

Motivations

Dependent Variable	G01	G03	G04	G05	G06	G07	G08	G09	G10	G11	G12	G13	G14	G15	G17	G18	Total Gains
Feel In Control		A T			A T	A T	A T	A T	A T	A T		A	A T	A T	A T	A T	12
Feel Secure	A	T		A	A T	A T	A T	A T	A T	A T			A	T		A	12
Feel Adequate		A			A T	A T	A	T	A	T		A	A T	A			10
Feel Certain		A T		A	A T	A T	A T	A T	A T	A T			A T	A	A T	T	12
Feel Productive		A T			A T	A T	A T	A T	A T	A T	T	A	A T	T	A T	A	13
Perceived Useful		A T			A T	A T	A T	A T	A T	A T	A T	A T	A T	A T	A T	A T	13
Not Suspicious		T			A T	A	A T	A T	A T	A T			T	A T			9
Not Harmful			T		A T	A T	T	T	A				A T	A	A		8
Product Reliable	A	A T			A T	A T	A T	A T	A T	A T	A T	T	A T		T	A	13
Trust Product		T		A	A T	A T	A T	A T	A T	A T		A	A	A T	A T	A	13

Table 9. Motivations results summary, by guideline experiment (columns) and dependent variable (rows), for those with "Abi"-like task-oriented and "Tim"-like tech-oriented motivations. Total occurrences: 77 inclusivity gains for both motivations (A T); 44 without inclusivity gains/losses (blank); 24 inclusivity gains for task-oriented only (A); 14 inclusivity gains for tech-oriented only (T); 1 inclusivity loss for tech-oriented only (T). Total possible: 160.

One result from **RQ2-AllStyles** was *who* were advantaged across these 16 product pairs. Note that the GenderMag assignment of endpoints to "Abi" vs. "Tim" followed widespread statistical skews of genders toward these particular styles [20]; previous research has shown that "Abi" styles have statistical tendencies to cluster, and so do the "Tim" styles. Thus, one might expect color patterns in risk's results (Table 5) to be similar to the columnar color patterns in, say, the Motivations results (Table 9).

However, this sometimes did not happen. For example, Table 5 visually contained twice as many T cells for the risk-tolerant participants as there were A cells for the risk-averse participants (27 vs. 13 respectively). But upcoming Tables 9, 10, 11, and 12 show that who gained more inclusivity advantages depended on which problem-solving style was considered. For example, Table 9 reverses

Learning Style

Dependent Variable	G01	G03	G04	G05	G06	G07	G08	G09	G10	G11	G12	G13	G14	G15	G17	G18	Total Gains
Feel In Control		A T			A T	A T	A T	A T	A T	A T		A	A T	A T	A T	A	12
Feel Secure	T	T		A T	A T	A T	A T	A T	A T	A T			T	T	T	A	13
Feel Adequate		T		T	A T	A T	A	A T	A	T			A T	A T			10
Feel Certain		T	A		A T	A T	T	A T	A T	A T			A T	A T	A T	A	11
Feel Productive	A	A T	T		A T	A T	A T	A T	A T	A T	T	A	A T	A	A T	A	15
Perceived Useful		A T			A T	A T	A T	A T	A T	A T	T	A T	T	A T	A T	A	13
Not Suspicious					A T	A	A T	A T	A	A T			A	A		A	9
Not Harmful		T	T		A T	A	T	A	A			T	A T	A			9
Product Reliable	T	A T			A T	A T	A T	A T	A T	A T	A		A T		T	A	12
Trust Product		T			A T	A T	A T	A T	A T	A T		A	A T	A T	A T	A	12

Table 10. Learning style results summary, by guideline experiment (columns) and dependent variable (rows), for the "Abi"-like process-oriented and "Tim"-like tinkering-oriented learners. Total occurrences: 72 inclusivity gains for both learning styles (A T); 42 without inclusivity gains/losses (blank); 13 inclusivity gains for process-oriented only (A); 27 inclusivity gains for tinkering-oriented only (T); 1 inclusivity loss for process-oriented only (A). 1 inclusivity loss for tinkering-oriented only (T). Total possible: 160.

Computer Self-Efficacy

Dependent Variable	G01	G03	G04	G05	G06	G07	G08	G09	G10	G11	G12	G13	G14	G15	G17	G18	Total Gains
Feel In Control		A T			A T	A T	A T	A T	A T	A T			A T	A T	A T	A T	11
Feel Secure	A	T		A	A T	A T	A T	A T	A T	A T		T	T	A	T	A T	14
Feel Adequate	A	T		A	A T	A T		A T	A	T	T		A T	T			11
Feel Certain		T			A T	A T	A T	A T	A T	T			A T	A T	A T		10
Feel Productive		A T		T	A T	A T	A T	A T	A T	A T	T	T	A T	A	A T	A	14
Perceived Useful	T	A T			A T	A T	A T	A T	A T	A T	A	A T	A T	A T	A T	T	14
Not Suspicious			T		A T	A	A T	T	A T	A T			A	A			8
Not Harmful			T		A T	A T	T	A	A			T	A T	A	T		9
Product Reliable	A	A T			A T	A T	A T	A T	A T	A T	A	T	T		T	T	13
Trust Product		T			A T	A T	A T	A T	A T	A T		T	T	A T	A T	T	12

Table 11. Computer self-efficacy results summary, by guideline experiment (columns) and dependent variable (rows), for those with "Abi"-like lower and "Tim"-like higher computer self-efficacy. Total occurrences: 72 inclusivity gains for both (A T); 42 without inclusivity gains/losses (blank); 17 inclusivity gains for lower only (A); 27 inclusivity gains for higher only (T); 2 inclusivity losses for higher only (T). Total possible: 160.

Table 5's trend, with more A cells for the task-oriented participants than T cells for the tech-oriented.

This trend sometimes occurred even within individual experiments, demonstrated in Figure 7 for Guideline 13's experiment. Guideline 13's product was a presentation app, and the AI feature was a design helper that recommended designs for alternative layouts. When participants saw the `Violation AI product`, they were told that "*... Violation AI product has not learned your preferences and blue designs appear in the same place among the suggested designs as the first time you used it,*" whereas the `Application AI product` "*...has learned your preferences and now features blue designs prominently.*" Considering participants' attitudes toward risk (first column), the risk-tolerant participants derived the most benefit from the `Application AI product`. However, the second column shows these same participants' data, but instead considering their motivations. This adds to the risk results; the `Application AI product` benefited both the risk-tolerant and also those with task-oriented motivations.

	Information Processing Style																	
Dependent Variable	G01	G03	G04	G05	G06	G07	G08	G09	G10	G11	G12	G13	G14	G15	G17	G18	Total Gains
Feel In Control		A T			A T	A T	A T	A T	A T	A T		T	A T	A T	A T	T	12
Feel Secure	T	A		A	A T	A T	A T	T	A T	A T				A		T	11
Feel Adequate		A			A T	A T	A	A T	A	T	A		T	A	T		11
Feel Certain		A		A	A T	A T	A T	A T	A T	A		A	A T	A T	A T	T	13
Feel Productive		A T			A T	A T	A T	A T	A T	A T	T	A	A T	A	A T	T	13
Perceived Useful		A T	A		A T	A T	A T	A T	A T	A T	A T	A T	T	A T	A T	T	14
Not Suspicious			A		A T	A	A T	A T	A T	A T			T	A T	A	A	10
Not Harmful					A T	A T	A	T	A				A T	T			7
Product Reliable	T	A			A T	A T	A T	A T	A T	A T	A		A	A	A	T	13
Trust Product		A T			A T	A T	A T	A T	A T	A T		A	T	A T	A T	T	12

Table 12. Information processing style results summary, by guideline experiment (columns) and dependent variable (rows), for "Abi"-like comprehensive and "Tim"-like selective information processors. Total occurrences: 70 inclusivity gains for both information processing styles (A T); 43 without inclusivity gains/losses (blank); 26 inclusivity gains for comprehensive only (A); 20 inclusivity gains for selective only (T); 1 inclusivity loss for comprehensive only (A). Total possible: 160.

For HAI practitioners, results like these suggest that different design decisions can appeal to different problem-solving styles for different reasons. For example, 46% (13/28) of the participants with task-oriented motivations mentioned how efficient they would become with the `Application AI product` or how much time they would save while using it:

> G13-2178-task-oriented: "*I like to have software that anticipates my needs, because it makes working more efficient.*"
>
> G13-4099-task-oriented: "*It is more efficient to see designs similar to those I have used before...it will take me less time to find them.*"
>
> G13-2740-task-oriented: "*It [the `Application AI product`] learned my preferences quicker which in time will save me time and trouble.*"

Dependent Variable	Ri.	Mo.	In.	SE	Le.
Feel In Control		A	T		A
Feel Secure	T			T	
Feel Adequate	T	A			
Feel Certain			A		
Feel Productive	T	A	A	T	A
Perceived Useful	A T	A T	A T	A T	A T
Not Suspicious					
Not Harmful				T	T
Product Reliable	T	T		T	
Trust Product	A T	A	A	T	A

Fig. 7. Guideline 13's inclusivity gains (rows) across the five different problem-solving styles (columns). Notice how the values switches between the "Tim"-like (T) and "Abi"-like (A) participants, depending on the problem-solving style. Ri. = Risk, Mo. = Motivations, In. = Information Processing Style, SE = Self-Efficacy, and Le. = Learning Style.

Although the participants with tech-oriented "Tim"-like motivations also raised efficiency and time savings, they did so less frequently than their task-oriented peers (only 17%—5/29):

> G13-662-tech-oriented: "*because it saves time than starting from scratch every time I use it [the `Application AI product`].*"
>
> G13-662-tech-oriented: "*I prefer [`Application AI product`] because of its ability to learn my preferences...thus helping me to work more efficiently.*"

Comments like these also have ties to the research literature. In that body of work, people who are more task-oriented prefer to use technologies to accomplish their task, using methods they are already familiar and comfortable with [19, 21, 24, 60, 87, 114]. Task-oriented people do so in an attempt to focus on the tasks that they care about, which might explain why these task-oriented participants commented so frequently on how the `Application AI product` saved them time; if the

product saves them time, then the task-oriented participants could achieve their task more quickly, devoting more time to what they care about rather than having to spend additional time recreating designs.

As the examples and tables in this section have shown, the types of participants who did and did not benefit from the changes in the `Application AI product` varied by problem-solving style—attending to only the Risk style did not tell the whole story. This suggests that HAI practitioners wanting to create a more inclusive AI-powered product for those with diverse problem-solving styles should consider all five of GenderMag's problem-solving styles.

Result #8: *The union of these five styles' results revealed more about who was left out—and why—than any one style's results alone could do.* G03, G13 (Figure 7), and G18 are cases in point, illustrated through the change in the colors of the cells between "Abi" and "Tim" across these five problem-solving styles.

6 PARTICIPANTS' PROBLEM-SOLVING STYLES AND THEIR DEMOGRAPHICS

Some HAI research has suggested demographic differences in different AI systems' HAI usability (e.g., [40, 79, 137]). Here, we consider whether problem-solving style results like those in Section 4 and Section 5 can shed light on *why* such demographic differences exist.

RQ3-DemographicDiversity seeks to understand how participants' problem-solving diversity aligned with their demographic diversity. The answer to this question will show whether problem-solving disparities in HAI user experiences can help explain demographic disparities in HAI user experiences.

For example, consider the Guideline 18 outcome variable of "certainty" and the two genders for whom enough data are present for inferential statistics—women and men. A statistical peek at the G18 data by gender reveals that the men's inclusivity significantly increased with G18's `Application AI product` over the `Violation AI product` ($t(28) = 3.1777$, $p = .004$, $d = 590$), whereas the women's did not ($t(34) = 1.0359$, $p = .308$, $d = .175$). This gender disparity seems problematic, but knowing its presence does not suggest a solution[15].

6.1 Problem-solving style diversity, gender, and age

If the gender results for the G18 example above show alignment with, for example, the G18 risk results of Section 4, the risk-oriented solution ideas from that section might help remove the gender disparity. And indeed, these two results do align: risk analysis showed that G18's `Application AI product`, which added user control to the AI product, did not provide significant inclusivity gains for the risk-averse certainty outcome ($t(31) = 1.7261$, $p = .094$, $d = .256$) but did for the risk-tolerant ($t(34) = 2.1884$, $p = .036$, $d = .370$).

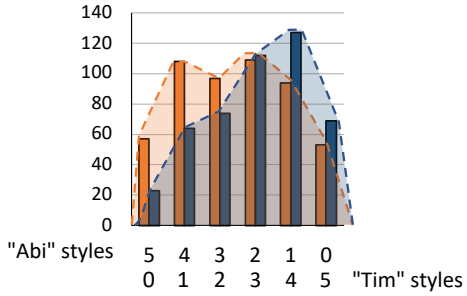
Our investigation into RQ3 will enable leveraging this kind of alignment. For example, if we find that the women participants skewed toward risk aversion, that knowledge would suggest that improving the G18 `Application AI product`'s inclusivity across the risk spectrum could also improve its inclusivity across the gender spectrum.

Thus, to find out how our participants' problem-solving styles aligned with their genders, we counted the number of "Abi"-like and "Tim"-like styles of each participant of all 16 experiments, and then compared the counts by gender. We begin with the two genders for whom enough data are present for inferential statistics—women and men, who provided 98.7% of the data—and then non-statistically present the data for the participants in the LGBTQIA* community[16].

[15] One of the authors of this paper is reminded of the many times she has heard software practitioners say things like, "what am I supposed to do, paint it pink?"

[16] LGBTQIA* used based on Scheuerman et al.'s living document [105].

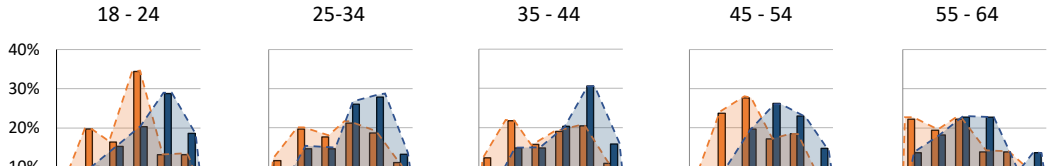
As Figure 8 (left) shows, the women were split almost equally between having three or more "Abi"-like styles (first three orange bars, 50.6%), versus having two or fewer (49.4%). For example, the leftmost pair of bars show that 59 women and 24 men had five Abi-like problem-solving style values (0 Tim-like styles). In contrast, the men skewed heavily toward the right; only 34.5% of the men had three or more "Abi"-like styles (first three blue bars). As Figure 8 (right) shows, these gender skew differences were statistically significant under Fisher's exact test ($p < .0001$)[17]. Vorvoreanu et al. [125] found similar gender skew results while investigating an academic search tool.



	3 or more "Abi"-like styles	less than 3 "Abi"-like styles
Woman	277	249
Man	173	302
Estimate (k)	236	
p-value	$< .0001^{***}$	

Fig. 8. (Left): Counts of women and men (y-axis) in all 16 experiments by the number of Abi-direction or Tim-direction problem-solving styles each participant reported (x-axis). The men (blue) skewed more to the right (i.e., more "Tim" styles) than the women (orange) did. (Right): The Fisher's exact test 2x2 contingency table, revealing that the difference was highly significant.

Adding age demographics into our analysis, an intersectional gender-age analysis showed analogous gender skews in each of the five age groups in our data (Figure 9). The results were significant in the three age groups between ages 25—54.



	18 – 24		25 – 34		35 – 44		45 – 54		55 – 64	
	≥ 3	< 3	≥ 3	< 3	≥ 3	< 3	≥ 3	< 3	≥ 3	< 3
Woman	25	36	101	97	77	70	47	29	23	13
Man	19	40	77	142	37	71	23	38	12	10
Estimate (k)	22		85		66		39		22	
p-value	.3482		$.0015^{**}$		$.0050^{**}$		$.0061^{**}$		.5831	

Fig. 9. (Top): Percentage of participants (y-axes) from Figure 8, divided into age groups. Men in all age groups visually skewed towards having fewer "Abi"-like styles (x-axes) than the women did. (Bottom): The Fisher's exact test 2x2 contingency tables. The middle three categories had significant gender differences.

[17]For this test, we used the threshold that *minimized* the chance of showing significance by maximizing the sum of p-values [101].

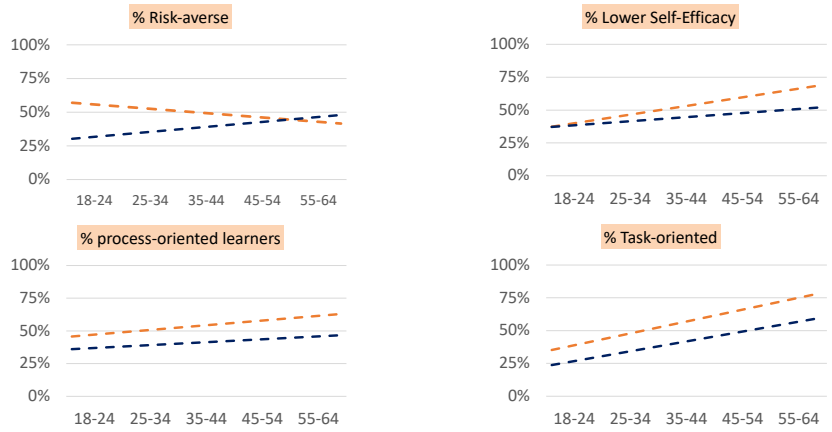


Fig. 10. The percentages (y-axes) across the five age groups (x-axes) of women (orange) and men (blue) who exhibited each of four "Abi"-like problem solving styles—risk-aversion, lower computer self-efficacy compared to peers, process-oriented learning, and task-oriented motivations. The information processing style (not shown) trend lines were horizontal at the 50% mark, indicating no differences between women and men.

We also analyzed the presence of such gender-age intersectional results within each problem-solving style type. As Figure 10 suggests, the gender differences did manifest by age in four of the five style types.

For the four styles shown in Figure 10, the gender-by-age differences in these problem-solving attributes are consistent with other gender- and/or age-difference reports in the literature (e.g., [25, 36, 38, 40, 54, 84, 86, 107]). For information processing style, although our participants did not show these demographic differences, others' research has shown both gender differences [93, 113] and age differences [43, 52, 90, 119]. Such demographic differences in problem-solving style by gender and by age may help explain demographic differences between people's experiences with AI products (e.g., [46, 62, 121]).

Result #9: *Problem-solving styles and gender/age were related.* Participants' problem-solving styles clustered by both gender and age. An implication of this result is that inclusivity gains for certain problem-solving styles, as per the results in Sections 4 and 5, should also translate into inclusivity gains for certain genders and/or age groups.

6.2 The LGBTQIA* Community

The genders "woman" and "man" are only two points on the gender spectrum. Table 13 reports the GenderMag problem-solving style values for the 13 participants who were members of the LGBTQIA* community. Although a data set of 13 participants is small, we hope it will add to literature being populated by other researchers with data sets of LGBTQIA* participants (e.g., [2, 42, 56]), to enable the possibility of future meta-analyses to broaden our understanding of how to inclusively design for users of all gender identities.

7 DISCUSSION

7.1 Inclusivity and equity: Complements in HAI-UX fairness

Ideas about fairness in AI, what it is, and how to achieve it, have recently received substantial attention (e.g., [18, 41, 50, 57]). Research and conversations in this area usually refer to algorithmic or data fairness—but the ideas are also relevant to HAI-UX fairness.

PID	W	M	T	NB	NC	I	RISK	SE	INFO	MOTIV.	LEARN
1176				✓			Averse	Higher	Comprehensive	Tech	by Tinkering
3414				✓			Averse	Lower	Comprehensive	Task	by Process
3931			✓	✓			Tolerant	Lower	Selective	Tech	by Tinkering
3947				✓			Tolerant	Higher	Comprehensive	Tech	by Tinkering
4081				✓			Tolerant	Higher	Selective	Tech	by Tinkering
2718				✓			Tolerant	Higher	Selective	Tech	by Process
3601		✓	✓	✓			Tolerant	Lower	Comprehensive	Task	by Process
3065		✓	✓	✓			Tolerant	Higher	Comprehensive	Tech	by Tinkering
3099		✓		✓	✓	✓	Averse	Higher	Comprehensive	Tech	by Tinkering
1687	✓		✓				Tolerant	Higher	Selective	Tech	by Tinkering
4145				✓	✓		Averse	Higher	Selective	Tech	by Tinkering
1102				✓			Tolerant	Higher	Comprehensive	Tech	by Tinkering
1704				✓			Tolerant	Higher	Comprehensive	Tech	by Tinkering

Table 13. LGBTQIA* facet values: Each row shows one LGBTQIA* participant's problem-solving styles. Total LGBTQIA* participants across all experiments: 13. W = Woman, M = Man, T = Transgender, NB = Non-Binary, NC = Gender Non-Conforming, I = Intersex.

In considering any type of fairness, two concepts often drive the discussion—inclusivity and equity. This paper has considered inclusivity, but not equity.

A way to think about inclusivity in HAI is as an "outcome-oriented" concept that applies *within* a specific group of people. As shown in earlier sections, when an AI product somehow led to disadvantageous outcomes for some particular group of participants (e.g., risk-averse participants), then that product was not inclusive to *that* group.

Although this paper's inclusivity results revealed who the guidelines were helping the most and who was being left out, they do not answer how much more inclusivity progress an AI product still needs to make and for whom. Measuring equity can help to answer this question. Like inclusivity, equity in HAI is also an "outcome-oriented" concept—but it applies to between-group comparisons. For example, if an AI product's user experiences for two groups (e.g., risk-averse participants and risk-tolerant participants) were of the same high—or low—quality, then the product was equitable.

Ideally, one would like the inclusivity gains `Application AI product` achieved to result in a final outcome that is equitable to the two groups. To explore how useful a measure of equity would be to our investigation's results, we measured equity of a dependent variable's outcome for a given product as the absence of a significant difference between the two participant groups. Table 14 shows risk-group equity outcomes by this measure, superimposed on the risk-group inclusivity outcomes.

For example, Table 14's G15 column shows that the G15 `Application AI product` achieved inclusivity gains two times for risk-averse participants (orange cells) only, and three times for risk-averse participants (blue cells) only. The G15 column further shows that those five targeted gains, along with the inclusivity gains experienced by everyone, ultimately led to fully equitable outcomes the risk spectrum ("=" markings). Thus, applying the G15 guideline ended up targeting exactly who it should have targeted in order to bring everyone up to an equitable state.

In total, Table 14 shows that the guidelines' resulting `Application AI products` almost always produced equitable outcomes across the risk spectrum. Specifically, 129/160 outcomes (81%) were equitable, marked by "=" in the table. Of the 31/160 outcomes that were *in*equitable, only 3/160 (2%) favored the risk-averse ("A"), and 28/160 (17%) favored the risk-tolerant ("T").

Of course, equity does not always mean success. The G15 `Application AI product` produced entirely equitable HAI user experiences (Table 14), but only moderately positive HAI user experiences (revisit Figure 6). In contrast, the G1 `Application AI product` produced mostly equitable

	G01	G03	G04	G05	G06	G07	G08	G09	G10	G11	G12	G13	G14	G15	G17	G18
Feel In Control	=	=	T	=	=	=	=	=	=	T	=	=	T	=	=	T
Feel Secure	=	=	T	=	=	=	=	=	=	=	=	=	T	=	=	=
Feel Adequate	=	=	T	=	=	=	=	=	=	T	=	T	T	=	=	=
Feel Certain	=	=	T	=	=	=	=	=	=	=	=	T	=	=	=	T
Feel Productive	A	=	=	=	=	=	=	T	=	=	=	=	T	=	=	T
Perceived Useful	=	=	T	=	=	=	=	T	A	=	=	T	=	=	=	T
Not Suspicious	=	=	T	=	=	=	=	=	=	=	=	T	=	=	=	T
Not Harmful	=	=	T	=	=	=	=	=	=	=	=	T	=	=	=	T
Product Reliable	=	=	T	=	=	=	=	=	A	=	=	=	=	=	=	=
Trust Product	=	=	T	=	=	=	=	=	=	=	=	=	=	=	=	=

Table 14. Risk equity of all experiments' `Application AI products`. Equity symbols are superimposed on the Risk inclusivity result colors.
Equity symbols: equitable (=); inequitable favoring risk-Averse (**A**) or risk-Tolerant (**T**).
Inclusivity colors: Inclusivity gains for both risk-Averse and risk-Tolerant; inclusivity gains for risk-Averse only; inclusivity gains for risk-Tolerant only; inclusivity gains for nobody (no color).

but extremely low HAI user experiences, and the G6 `Application AI product` produced entirely equitable and very positive HAI user experiences.

Due to space limitations, we do not present equity results in detail for Risk or for the other four problem-solving styles. Still, our limited exploration here shows the additional value measuring equity can bring, so we advocate for measuring equity as well as inclusivity as a way to fully understand who is being included versus who is being left out.

7.2 Practical implications for HAI practitioners

Measuring inclusivity and equity can bring practical benefits to HAI practitioners. As our results show, incorporating users' problem-solving into AI products' HAI-UX work can sometimes point out where and why mismatches are arising between a group of users and an AI product. Some particularly actionable examples were given in Section 4.4. A way to gather participants' problem-solving styles would be to incorporate the validated survey [55] we used into user testing.

Armed with this new information, HAI practitioners could gain actionable insights in use-cases like the following:

HAI Practice Use-Case 1: To see which problem-solving groups of users are being left behind on an AI product with a problematic HAI-UX, measure *equity state*. To do so, for each problem-solving style, HAI practitioners could compare equity outcomes that are significantly different between the "A" participant group and the "T" participant group.

HAI Practice Use-Case 2: To see who a particular AI product change/new feature has benefited, measure *inclusivity changes*. To do so, HAI practitioners could compare "A" participants before the change versus after the change, and likewise for "T" participants, as in the examples in Sections 4 and 5.

HAI Practice Use-Case 3: After an AI product has changed, complement a measure of its *equity state* (as per Use-Case 1) with a measure of *dependent variable final outcomes* (e.g., as in Figure 6). This combination shows not only final equity state, but also how successful the AI product's HAI-UX is for each group of participants.

Our results suggest that doing measures like these can provide new, valuable information on who is being included, who is being left out, and how a product can improve.

7.3 Threats to Validity & Limitations

As with every empirical study [76, 129], our investigation has limitations and threats to validity.

In any study, researchers cannot ask participants every possible question, having to balance research goals with participant fatigue. As such, the dependent variables we analyzed may not have captured all information about people's reactions. For example, some participants' free-text remarks suggested outcomes that our Likert-style questionnaires did not cover; one example was participants' mentions of privacy concerns while interacting with certain products. Because the study was not designed with a dependent variable about privacy, we cannot be certain if remarks such as these indicated only isolated cases or more prevalent phenomena.

Another threat was how to handle missing data. Since participants had the option to say "I don't know" for any of the questions, we had to decide whether to 1) impute the data or 2) drop the "I don't know" values, costing degrees of freedom in our statistical tests. We chose the latter, because although there are many imputation methods to leverage (e.g., hot-deck, cold-deck, regression), any inferences are then limited to the imputed data, rather than the original data.

Another threat was how to handle the number of statistical tests we ran. As mentioned in Section 3.4, we did not report statistically corrected results in this paper because every test corresponded to a pre-planned hypothesis [8, 9]. That said, we recognize that some readers may not agree with this decision, so we also provide all Holm-Bonferroni corrected results in Appendix D.

Also, we chose to use vignettes vs. a real system. Each approach has its own advantage: Using vignettes allows enough control to genuinely isolate the experimental variation to vary ONLY the independent variable, and this isolation was critical to our statistical power. In contrast, a real system's strength is realism in the external world, but at the cost of controls. Because this was a set of controlled experiments, we chose control, leaving to other studies to investigate external validity questions (faithfulness to real world conditions).

Other threats to validity could arise from the particular pairing of vignette to product, and/or from participants associating a vignette with a specific real product with which they had familiarity. We attempted to avert the latter by randomly assigning generic names (Ione and Kelso) instead of real product names, but participants may have still imagined their favorite productivity software. If this occurred, it would contribute an extra source of variation in these data.

Although the productivity software and GenderMag problem-solving styles have been shown to be viable/useful in countries around the world, the participants in our study were restricted to those who lived in the USA at the time of the study. As such, the results in this paper cannot be generalized to other countries around the world. However, since the methodology is not U.S.-specific, replicating the study with participants from additional countries should be straightforward.

One limitation of this investigation is that its results cannot be generalized to AI-powered systems outside of productivity software. This suggests the need to investigate HAI-UX impacts on diverse problem-solvers across a spectrum of domains, from low-stakes domains (e.g., music recommender systems) to high-stakes domains (e.g., automated healthcare or autonomous vehicles).

Threats and limitations like these can only be addressed through additional studies across a spectrum of empirical methods and situations, in order to isolate different independent variables of study and establish generality of findings across different AI applications, measurements, and populations.

8 CONCLUSION

This paper has presented a new empirical approach for measuring an AI product's user experience inclusivity, to answer questions like these: *what kinds of users does this AI-powered product support and who does it leave out? And what changes could make it more inclusive?*

The essence of the approach is to empirically measure participants' values in each of five problem-solving style types, and then to empirically analyze how well users with different values in those style types were supported by the AI product. The paper demonstrates the approach on an empirical investigation of 16 AI-powered products, and those products' outcomes on a total of 1,106 human participants.

Among the results of the empirical investigation were:

- *Actionable*: Many of the empirical results of applying the approach pointed to directions that HAI practitioners could take to make the AI products more inclusive (**Result #6**).
- *Risk—impacts on all of control, suspicion, trust, and certainty*: When an AI product had risk implications, four variables' values varied in tandem: participants' feelings of control, their (lack of) suspicion, their trust in the product, and their certainty while using the product (**Result #2**).
- *User control mattered*: The more control an AI product offered users, the more inclusive it was for both risk attitudes (**Results #3 & 4**).
- *Stay away from my data!* When an AI product was learning from "my" data, risk-averse participants rarely experienced inclusivity gains (**Result #5**).
- *The Amershi HAI guidelines usually helped inclusivity*: Although the Amershi guidelines were not designed with inclusivity in mind, they usually helped with inclusivity for at least one group of participants, for all five GenderMag problem-solving style spectra (**Results #1 & 7**).
- *Problem-solving diversity meets demographic diversity*: The participants' problem-solving styles showed alignments with their intersectional gender-and-age demographic diversity. These alignments suggest that improving an AI product's inclusivity to diverse problem-solving styles (e.g., attitudes toward risk) is likely to improve the product's demographic (e.g., gender, age) inclusivity as well (**Result #9**).
- *Problem-solving styles mattered*: Which participants were most advantaged/disadvantaged depended on both the AI product and the particular problem-solving style. This suggests that measuring each problem-solving style separately is key to finding an actionable fix to pinpoint supporting that particular style (**Result #8**).

Abstracting above these results, this work directly relates to one of Shneiderman's stances for Human-Centered Artificial Intelligence [112]. Shneiderman advocates a "shift from emulating humans" to "empowering people". This paper provides an approach for carrying out this point, and taking it one step further: from "empowering people" to empowering *diverse* people.

ACKNOWLEDGMENTS

We thank Rupika Dikkala, Catherine Hu, Jeramie Kim, Elizabeth Li, Caleb Matthews, Christopher Perdriau, Sai Raja, and Prisha Velhal for their help with this paper. We are grateful to the editors and reviewers for their encouragement and constructive engagement, which greatly helped to improve this paper. This work was supported in part by Microsoft, by NSF #1901031 and #2042324; and by USDA-NIFA/NSF #2021-67021-35344. Any opinions, findings, conclusions, or recommendations expressed are the authors' and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Bryan Abendschein, Chad Edwards, and Autumn Edwards. 2021. The influence of agent and message type on perceptions of social support in human-machine communication. *Communication Research Reports* 38, 5 (2021), 304–314.

[2] Dane Acena and Guo Freeman. 2021. "In My Safe Space": Social Support for LGBTQ Users in Social Virtual Reality. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.

[3] Tamara Adlin and John Pruitt. 2010. *The essential persona lifecycle: Your guide to building and using personas*. Morgan Kaufmann.

[4] Puja Agarwal, Jeramie Kim, Elizabeth Li, Margaret Burnett, and Anita Sarma. 2023. Designing for Inclusive National Digital ID Platform: A MOSIP Case Study. In *International Conference on Trustworthy Digital ID*.

[5] Rangina Ahmad, Dominik Siemon, Ulrich Gnewuch, and Susanne Robra-Bissantz. 2022. Designing personality-adaptive conversational agents for mental health care. *Information Systems Frontiers* (2022), 1–21.

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 3.

[7] Taif Anjum, Steven Lawrence, and Amir Shabani. 2021. Augmented Reality and Affective Computing on the Edge Makes Social Robots Better Companions for Older Adults.. In *ROBOVIS*. 196–204.

[8] Richard A Armstrong. 2014. When to use the B onferroni correction. *Ophthalmic and Physiological Optics* 34, 5 (2014), 502–508.

[9] Richard A Armstrong and Anthony C Hilton. 2011. *Statistical analysis in microbiology: statnotes*. Wiley Online Library.

[10] Katrin Auspurg, Thomas Hinz, and Stefan Liebig. 2009. Complexity, learning effects and plausibility of vignettes in the factorial survey design. *methods, data, analyses* 3, 1 (2009), 38.

[11] Tae Hyun Baek and Minseong Kim. 2023. Is ChatGPT scary good? How user motivations affect creepiness and trust in generative artificial intelligence. *Telematics and Informatics* 83 (2023), 102030.

[12] Albert Bandura. 1986. The explanatory and predictive scope of self-efficacy theory. *Journal of social and clinical psychology* 4, 3 (1986), 359–373.

[13] Laura Beckwith and Margaret Burnett. 2004. Gender: An important factor in end-user programming environments?. In *2004 IEEE symposium on visual languages-human centric computing*. IEEE, 107–114.

[14] 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.

[15] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[16] Eileen Bridges. 2018. Hedonic and utilitarian shopping goals: a decade later. *Journal of Global Scholars of Marketing Science* 28, 3 (2018), 282–290.

[17] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[18] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[19] Margaret Burnett, Scott D Fleming, Shamsi Iqbal, Gina Venolia, Vidya Rajaram, Umer Farooq, Valentina Grigoreanu, and Mary Czerwinski. 2010. Gender differences and programming environments: Across programming populations. In *Proceedings of the 2010 ACM-IEEE international symposium on empirical software engineering and measurement*. 1–10.

[20] 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.

[21] Margaret M Burnett, Laura Beckwith, Susan Wiedenbeck, Scott D Fleming, Jill Cao, Thomas H Park, Valentina Grigoreanu, and Kyle Rector. 2011. Gender pluralism in problem-solving software. *Interacting with computers* 23, 5 (2011), 450–460.

[22] John T Cacioppo, Richard E Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of personality assessment* 48, 3 (1984), 306–307.

[23] Jeffrey Carver, Rafael Capilla, Birgit Penzenstadler, Alexander Serebrenik, and Alejandro Valdezate. 2018. Gender, sentiment and emotions, and safety-critical systems. *IEEE Software* 35, 6 (2018), 16–19.

[24] Justine Cassell et al. 2002. Genderizing hci. *The Handbook of Human–Computer Interaction. Mahwah, NJ: Erlbaum* (2002), 402–411.

[25] Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren G Terveen. 2014. Specialization, homophily, and gender in a social curation site: Findings from Pinterest. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 674–686.

[26] Suzanne Chapman. 2022. A Quick Guide to Inclusive Design. (2022). https://medium.com/the-u-s-digital-service/a-quick-guide-to-inclusive-design-be4931ef2c

[27] Gary Charness and Uri Gneezy. 2012. Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization* 83, 1 (2012), 50–58.
[28] Robin Cohen, Rishav Raj Agarwal, Dhruv Kumar, Alexandre Parmentier, and Tsz Him Leung. 2020. Sensitivity to Risk Profiles of Users When Developing AI Systems. In *Advances in Artificial Intelligence: 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13–15, 2020, Proceedings 33*. Springer, 138–150.
[29] Savia Coutinho, Katja Wiemer-Hastings, John J Skowronski, and M Anne Britt. 2005. Metacognition, need for cognition and use of explanations during ongoing learning and problem solving. *Learning and Individual Differences* 15, 4 (2005), 321–337.
[30] Sally Jo Cunningham, Annika Hinze, and David M Nichols. 2016. Supporting gender-neutral digital library creation: A case study using the GenderMag Toolkit. In *Digital Libraries: Knowledge, Information, and Data in an Open Access Society: 18th International Conference on Asia-Pacific Digital Libraries, ICADL 2016, Tsukuba, Japan, December 7–9, 2016, Proceedings 18*. Springer, 45–50.
[31] Rachel Curry. 2023. Recent data shows AI job losses are rising, but the numbers don't tell the full story. *CNBC Technology Executive Council* (2023). https://www.cnbc.com/2023/12/16/ai-job-losses-are-rising-but-the-numbers-dont-tell-the-full-story.html
[32] Jenny L Davis, Daniel B Shank, Tony P Love, Courtney Stefanik, and Abigail Wilson. 2022. Gender Dynamics in Human-AI Role-Taking. (2022).
[33] Maartje Ma De Graaf, Somaya Ben Allouch, and Tineke Klamer. 2015. Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in human behavior* 43 (2015), 1–14.
[34] Douglas C Derrick and Gina Scott Ligon. 2014. The affective outcomes of using influence tactics in embodied conversational agents. *Computers in Human Behavior* 33 (2014), 39–48.
[35] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3301275.3302310
[36] Thomas Dohmen, Armin Falk, Bart HH Golsteyn, David Huffman, and Uwe Sunde. 2017. Risk attitudes across the life course.
[37] Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the european economic association* 9, 3 (2011), 522–550.
[38] Bireswar Dutta, Mei-Hui Peng, and Shu-Lung Sun. 2018. Modeling the adoption of personal health record (PHR) among individual: the effect of health-care technology self-efficacy and gender concern. *Libyan Journal of Medicine* 13, 1 (2018).
[39] Simon Eisbach, Markus Langer, and Guido Hertel. 2023. Optimizing human-AI collaboration: Effects of motivation and accuracy information in AI-supported decision-making. *Computers in Human Behavior: Artificial Humans* 1, 2 (2023), 100015.
[40] Priska Flandorfer. 2012. Population ageing and socially assistive robots for elderly persons: the importance of sociodemographic factors for user acceptance. *International Journal of Population Research* 2012 (2012).
[41] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1918–1921.
[42] Guo Freeman, Divine Maloney, Dane Acena, and Catherine Barwulor. 2022. (Re) discovering the Physical Body Online: Strategies and Challenges to Approach Non-Cisgender Identity in Social Virtual Reality. In *CHI Conference on Human Factors in Computing Systems*. 1–15.
[43] Linda Geerligs, Karen L Campbell, et al. 2018. Age-related differences in information processing during movie watching. *Neurobiology of Aging* 72 (2018), 106–120.
[44] Dimitrios Giakoumis, Konstantinos Votis, Efthymios Altsitsiadis, Sofia Segkouli, Ioannis Paliokas, and Dimitrios Tzovaras. 2019. Smart, personalized and adaptive ICT solutions for active, healthy and productive ageing with enhanced workability. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 442–447.
[45] Omri Gillath, Ting Ai, Michael S Branicky, Shawn Keshmiri, Robert B Davison, and Ryan Spaulding. 2021. Attachment and trust in artificial intelligence. *Computers in Human Behavior* 115 (2021), 106607.
[46] Jessica Gish, Brenda Vrkljan, Amanda Grenier, and Benita Van Miltenburg. 2017. Driving with advanced vehicle technology: A qualitative investigation of older drivers' perceptions and motivations for use. *Accident Analysis & Prevention* 106 (2017), 498–504.
[47] Google. 2019. The UX of AI - Library. https://design.google/library/ux-ai/
[48] M Graham, A Milanowski, and J Miller. 2011. Measuring and promoting inter-rater reliability of teacher and principal performance ratings.

[49] Catarina Gralha, Miguel Goulão, and João Araujo. 2020. Are there gender differences when interacting with social goal models? *Empirical Software Engineering* 25, 6 (2020), 5416–5453.

[50] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[51] Chiara Grosso and Cipriano Forza. 2021. Exploring Configurator Users' Motivational Drivers for Digital Social Interaction. In *Intelligent Systems in Industrial Applications*. Springer, 118–138.

[52] Duncan Guest, Christina J Howard, Louise A Brown, and Harriet Gleeson. 2015. Aging and the rate of visual information processing. *Journal of vision* 15, 14 (2015), 10–10.

[53] Mariam Guizani, Igor Steinmacher, Jillian Emard, Abrar Fallatah, Margaret Burnett, and Anita Sarma. 2022. How to Debug Inclusivity Bugs? A Debugging Process with Information Architecture. In *ACM/IEEE International Conference on Software Engineering, Software Engineering in Society Track (ICSE-SEIS'22)*. ACM, 1–12.

[54] Mustafa Serkan Gunbatar and Halit Karalar. 2018. Gender differences in middle school students' attitudes and self-efficacy perceptions towards mBlock programming. *European Journal of Educational Research* 7, 4 (2018), 925–933.

[55] Md Montaser Hamid, Amreeta Chatterjee, Mariam Guizani, Andrew Anderson, Fatima Moussaoui, Sarah Yang, Isaac Escobar, Anita Sarma, and Margaret Burnett. To Appear. *How to Measure Diversity Actionably*.

[56] Jean Hardy and Stefani Vargas. 2019. Participatory design and the future of rural LGBTQ communities. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. 195–199.

[57] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 392–402.

[58] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[59] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[60] Weimin Hou, Manpreet Kaur, Anita Komlodi, Wayne G Lutters, Lee Boot, Shelia R Cotten, Claudia Morrell, A Ant Ozok, and Zeynep Tufekci. 2006. "Girls don't waste time"' pre-adolescent attitudes toward ICT. In *CHI'06 extended abstracts on Human factors in computing systems*. 875–880.

[61] Yaou Hu and Hyounae Kelly Min. 2023. The dark side of artificial intelligence in service: The "watching-eye" effect and privacy concerns. *International Journal of Hospitality Management* 110 (2023), 103437.

[62] Lynn M Hulse, Hui Xie, and Edwin R Galea. 2018. Perceptions of autonomous vehicles: Relationships with road users, risk, gender and age. *Safety science* 102 (2018), 1–13.

[63] Walter Hyll and Maike Irrek. 2015. *The impact of risk attitudes on financial investments*. Technical Report. IWH Discussion Papers.

[64] Apple Inc. 2019. Machine learning. https://developer.apple.com/design/human-interface-guidelines/technologies/machine-learning/introduction/

[65] Spencer W JaQuay. 2023. *Exploring the Emotional Basis of Need for Cognition in Collaborative Problem-Solving*. Ph.D. Dissertation. UC Irvine.

[66] 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.

[67] JJ Jiang, Gary Klein, and RG Vedder. 2000. Persuasive expert systems: the influence of confidence and discrepancy. *Computers in Human Behavior* 16, 2 (2000), 99–109.

[68] Josef Jönsson. 2021. AI acceptance and attitudes: people's perception of healthcare and commercial AI applications. *Linköpings University LIU-IDA/KOGVET-G–21/003–SE* (2021).

[69] Ofem Usani Joseph, Iyam Mary Arikpo, Ovat Sylvia Victor, Nwogwugwu Chidirim, Anake Paulina Mbua, Udeh Mary-rose Ify, and Otu Bernard Diwa. 2024. Artificial Intelligence (AI) in academic research. A multi-group analysis of students' awareness and perceptions using gender and programme type. *Journal of Applied Learning and Teaching* 7, 1 (2024).

[70] Büşra Kartal and Uğur Başarmak. 2022. Preservice computer science teachers' beliefs, motivational orientations, and teaching practices. *Educational Studies* (2022), 1–24.

[71] Jared Katzman, Angelina Wang, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. 2023. Taxonomizing and Measuring Representational Harms: A Look at Image Tagging. *arXiv preprint arXiv:2305.01776* (2023).

[72] Jack Kelly. [n.d.]. Goldman Sachs Predicts 300 Million Jobs Will Be Lost Or Degraded By Artificial Intelligence. *Forbes* ([n. d.]). https://www.forbes.com/sites/jackkelly/2023/03/31/goldman-sachs-predicts-300-million-jobs-will-be-lost-or-degraded-by-artificial-intelligence/?sh=632106f782b4

[73] Aqdas Khan. 2023. Psychological influences on problem-solving following lab-induced learned helplessness. (2023).

[74] Doug Kim, Margaret Price, Christina Mallon, Nathan Kile, Anna Cook, Keira Xu, Anna Tendera, Andres Pacheco, Tenille Lively, Catherine Ekonomou, and Dona Sarkar. 2023. Microsoft Inclusive Design for Cognition Guidebook. (2023).

[75] Yus Kirillova and DA Malykh. 2017. Gender accessories recognition of user web-applications by classifiers. *Alley of Science* 4, 9 (2017), 854–857.

[76] A J Ko, T D Latoza, and M M Burnett. 2015. A practical guide to controlled experiments of software engineering tools with human participants. *Empirical Software Engineering* 20, 1 (2015), 110–141.

[77] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *ACM Intl. Conf. on Intelligent User Interfaces (IUI '15)*. ACM, 126–137.

[78] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–10.

[79] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M. Burnett, Stephen Perona, Amy J. Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of Naive Bayes text classification. *ACM Transactions on Interactive Intelligent Systems* 1, 1 (2011). https://doi.org/10.1145/2030365.2030367

[80] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[81] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2022. Towards Efficient Annotations for a Human-AI Collaborative, Clinical Decision Support System: A Case Study on Physical Stroke Rehabilitation Assessment. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 4–14. https://doi.org/10.1145/3490099.3511112

[82] Qing Li, Sharon Chu, Nanjie Rao, and Mahsan Nourani. 2020. Understanding the Effects of Explanation Types and User Motivations on Recommender System Use. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1, 83–91. https://doi.org/10.1609/hcomp.v8i1.7466

[83] Tianyi Li, Mihaela Vorvoreanu, Derek DeBellis, and Saleema Amershi. 2022 (to appear). Assessing Human-AI Interaction Early through Factorial Surveys: A Study on the Guidelines for Human-AI Interaction. *ACM Transactions on Computer-Human Interaction (ToCHI)* (2022 (to appear)).

[84] Matthew J Liberatore and William P Wagner. 2022. Gender, performance, and self-efficacy: a quasi-experimental field study. *Journal of Computer Information Systems* 62, 1 (2022), 109–117.

[85] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing systems*.

[86] Omar López-Vargas, Leydy Duarte-Suárez, and Jaime Ibáñez-Ibáñez. 2017. Teacher's computer self-efficacy and its relationship with cognitive style and TPACK. *Improving Schools* 20, 3 (2017), 264–277.

[87] Jane Margolis and Allan Fisher. 2002. *Unlocking the clubhouse: Women in computing*. MIT press.

[88] Nicola Marsden and Maren Haag. 2016. Evaluation of GenderMag personas based on persona attributes and persona gender. In *HCI International 2016–Posters' Extended Abstracts: 18th International Conference, HCI International 2016, Toronto, Canada, July 17-22, 2016, Proceedings, Part I 18*. Springer, 122–127.

[89] Juan Martínez-Miranda, Humberto Pérez-Espinosa, Ismael Espinosa-Curiel, Himer Avila-George, and Josefína Rodríguez-Jacobo. 2018. Age-based differences in preferences and affective reactions towards a robot's personality during interaction. *Computers in Human Behavior* 84 (2018), 245–257.

[90] Jennifer McIntosh, Xiaojiao Du, Zexian Wu, Giahuy Truong, Quang Ly, Richard How, Sriram Viswanathan, and Tanjila Kanij. 2021. Evaluating Age Bias In E-commerce. In *2021 IEEE/ACM 13th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*. IEEE, 31–40.

[91] Kevin McKee, Xuechunzi Bai, and Susan Fiske. 2021. Understanding human impressions of artificial intelligence. (2021).

[92] Christopher Mendez, Lara Letaw, Margaret Burnett, Simone Stumpf, Anita Sarma, and Claudia Hilderbrand. 2019. From GenderMag to InclusiveMag: An inclusive design meta-method. In *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 97–106.

[93] Joan Meyers-Levy and Barbara Loken. 2015. Revisiting gender differences: What we know and what lies ahead. *Journal of Consumer Psychology* 25, 1 (2015), 129–149.

[94] Martijn Millecamp, Robin Haveneers, and Katrien Verbert. 2020. Cogito ergo quid? the effect of cognitive style in a transparent mobile music recommender system. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 323–327.

[95] Emerson Murphy-Hill, Alberto Elizondo, Ambar Murillo, Marian Harbach, Bogdan Vasilescu, Delphine Carlson, and Florian Dessloch. 2024. GenderMag Improves Discoverability in the Field, Especially for Women. In *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)*. IEEE Computer Society, 973–973.

[96] Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2023. In-IDE Generation-based Information Support with a Large Language Model. *arXiv preprint arXiv:2307.08177* (2023).

[97] Heather Lynn O'Brien. 2010. The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. *Interacting with computers* 22, 5 (2010), 344–352.

[98] Susmita Hema Padala, Christopher John Mendez, Luiz Felipe Dias, Igor Steinmacher, Zoe Steine Hanson, Claudia Hilderbrand, Amber Horvath, Charles Hill, Logan Dale Simpson, Margaret Burnett, et al. 2020. How gender-biased tools shape newcomer experiences in OSS projects. *IEEE Transactions on Software Engineering* (2020).

[99] Thomas V Perneger. 1998. What's wrong with Bonferroni adjustments. *Bmj* 316, 7139 (1998), 1236–1238.

[100] Peter Pirolli and Stuart Card. 1995. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 51–58.

[101] Fred Ramsey and Daniel Schafer. 2012. *The statistical sleuth: a course in methods of data analysis*. Cengage Learning.

[102] 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. https://books.google.com/books?id=e8jhiYjQrU0C

[103] Lara Riefle, Patrick Hemmer, Carina Benz, Michael Vössing, and Jannik Pries. 2022. On the Influence of Cognitive Styles on Users' Understanding of Explanations. In *Proceedings of the Forty-Third International Conference on Information Systems (ICIS)*.

[104] Julia Rudolph, Samuel Greiff, Anja Strobel, and Franzis Preckel. 2018. Understanding the link between need for cognition and complex problem solving. *Contemporary Educational Psychology* 55 (2018), 53–62.

[105] Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. 2020. HCI guidelines for gender equity and inclusivity. *UMBC Faculty Collection* (2020).

[106] Philipp Schmidt and Felix Biessmann. 2020. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*. Springer, 431–449.

[107] Günther Schreder, Michael Smuc, Karin Siebenhandl, and Eva Mayr. 2013. Age and computer self-efficacy in the use of digital technologies: an investigation of prototypes for public self-service terminals. In *International conference on universal access in human-computer interaction*. Springer, 221–230.

[108] Suleman Shahid, Emiel Krahmer, and Marc Swerts. 2014. Child–robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend? *Computers in Human Behavior* 40 (2014), 86–100.

[109] Esha Shandilya and Mingming Fan. 2022. Understanding Older Adults' Perceptions and Challenges in Using AI-enabled Everyday Technologies. *arXiv preprint arXiv:2210.01369* (2022).

[110] Chun Shao and K Hazel Kwon. 2021. Hello Alexa! Exploring effects of motivational factors and social presence on satisfaction with artificial intelligence-enabled gadgets. *Human Behavior and Emerging Technologies* 3, 5 (2021), 978–988.

[111] Arun Shekhar and Nicola Marsden. 2018. Cognitive Walkthrough of a learning management system with gendered personas. In *Proceedings of the 4th Conference on Gender & IT*. 191–198.

[112] Ben Shneiderman. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12, 3 (2020), 109–124.

[113] Dilruba Showkat and Cindy Grimm. 2018. Identifying gender differences in information processing style, self-efficacy, and tinkering for robot tele-operation. In *2018 15th international conference on ubiquitous robots (UR)*. IEEE, 443–448.

[114] Steven John Simon. 2000. The impact of culture and gender on web sites: an empirical study. *ACM SIGMIS Database: The Database for Advances in Information Systems* 32, 1 (2000), 18–37.

[115] Marita Skjuve, Petter Bae Brandtzæg, and Asbjørn Følstad. 2024. Why do people use ChatGPT? Exploring user motivations for generative conversational AI. *First Monday* 29, 1 (2024).

[116] Thomas F Stafford and Marla R Stafford. 2001. Identifying motivations for the use of commercial web sites. *Information Resources Management Journal (IRMJ)* 14, 1 (2001), 22–30.

[117] Simone Stumpf, Anicia Peters, Shaowen Bardzell, Margaret Burnett, Daniela Busse, Jessica Cauchard, and Elizabeth Churchill. 2020. Gender-Inclusive HCI Research and Design: A Conceptual Review. *Foundations and Trends in Human-Computer Interaction* 13, 1 (2020), 1–69.

[118] Adrian Tchaikovsky. 2023. Attack of the 50-foot A.I. *New York Times* (2023).

[119] Anna Torrens-Burton, Claire J Hanley, Rodger Wood, Nasreen Basoudan, Jade Eloise Norris, Emma Richards, and Andrea Tales. 2020. Lacking pace but not precision: Age-related information processing changes in response to a dynamic attentional control task. *Brain Sciences* 10, 6 (2020), 390.

[120] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. (2021).

[121] Margot J van der Goot and Tyler Pilgrim. 2019. Exploring age differences in motivations for and acceptance of chatbot communication in a customer service context. In *International Workshop on Chatbot Research and Design*. Springer, 173–186.
[122] Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8 (2018), 1080–1088.
[123] Boele Visser. [n.d.]. Gender Appearance, Collaboration and Disclosure on Motivation. ([n. d.]).
[124] Cristian Voica, Florence Mihaela Singer, and Emil Stan. 2020. How are motivation and self-efficacy interacting in problem-solving and problem-posing? *Educational Studies in Mathematics* 105 (2020), 487–517.
[125] 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 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[126] Gert G Wagner, Richard V Burkhauser, and Friederike Behringer. 1993. The English language public use file of the German Socio-Economic Panel. *Journal of Human resources* (1993), 429–433.
[127] Elke U Weber, Ann-Renee Blais, and Nancy E Betz. 2002. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of behavioral decision making* 15, 4 (2002), 263–290.
[128] Steve Whittaker. 2011. Personal information management: from information consumption to curation. *Annual review of information science and technology* 45, 1 (2011), 1.
[129] Claes Wohlin, Per Runeson, Martin Höst, Magnus Ohlsson, Björn Regnell, and Anders Wesslén. 2000. *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers, Norwell, MA, USA.
[130] Austin P Wright, Zijie J Wang, Haekyu Park, Grace Guo, Fabian Sperrle, Mennatallah El-Assady, Alex Endert, Daniel Keim, and Duen Horng Chau. 2020. A comparative analysis of industry human-AI interaction guidelines. *arXiv preprint arXiv:2010.11761* (2020).
[131] Wei Xu, Marvin J Dainoff, Liezhong Ge, and Zaifeng Gao. 2021. From human-computer interaction to human-AI Interaction: new challenges and opportunities for enabling human-centered AI. *arXiv preprint arXiv:2105.05424* 5 (2021).
[132] Mustafa Yağcı. 2016. Effect of attitudes of information technologies (IT) preservice teachers and computer programming (CP) students toward programming on their perception regarding their self-sufficiency for programming Bilişim teknolojileri (BT) öğretmen adaylarının ve bilgisayar programcılığı (BP) öğrencilerinin programlamaya karşı tutumlarının programlama öz yeterlik algılarına etkisi. *Journal of Human Sciences* 13, 1 (2016), 1418–1432.
[133] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 547–558.
[134] 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*. 1–13.
[135] Ziyue Yang, Fengye Sun, Lingrui Zhao, Tingwei Hu, Xin Lin, and Yufang Guo. 2023. Self-efficacy and well-being in the association between caregiver burden and sleep quality among caregivers of elderly patients having multiple chronic conditions in rural China: a serial multiple mediation analysis. *BMC nursing* 22, 1 (2023), 424.
[136] Chengmin Zhou, Wenjing Zhan, Ting Huang, Hanxiao Zhao, and Jake Kaner. 2023. An empirical study on the collaborative usability of age-appropriate smart home interface design. *Frontiers in Psychology* 14 (2023).
[137] Dilawar Shah Zwakman, Debajyoti Pal, and Chonlameth Arpnikanondt. 2021. Usability evaluation of artificial intelligence-based voice assistants: The case of Amazon Alexa. *SN Computer Science* 2, 1 (2021), 1–16.

ONLINE APPENDICES

The following electronic appendices accompany this paper in the ACM Digital Library. In addition, we have provided local copies of each:

- Appendix A provides the problem-solving style survey, along with the rules for discerning participants' problem-solving style values for each of the problem-solving style types.
- Appendix B shows the vignettes for both the `Violation AI product` and `Application AI product` for all 16 experiments.
- Appendix C provides the demographic data for all participants.
- Appendix D shows *all* of the statistical tests for *all* experiments and *all* problem-solving style types, along with whether they were significant under Holm-Bonferroni correction.