# "Over-the-Hood" AI Inclusivity Bugs and How 3 AI Product Teams Found and Fixed Them


Andrew A. Anderson
IBM Research
Almaden, California, USA
Andrew.Anderson2@ibm.com

Fatima A. Moussaoui
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, Oregon, USA
moussaof@oregonstate.edu

Jimena Noa-Guevara
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, Oregon, USA
noaguevg@oregonstate.edu

Md Montaser Hamid
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, Oregon, USA
hamidmd@oregonstate.edu

Margaret Burnett
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, Oregon, USA
burnett@eecs.oregonstate.edu



## Abstract

While much research has shown the presence of AI's "under-the-hood" biases (e.g., algorithmic, training data, etc.), what about "over-the-hood" inclusivity biases: barriers in user-facing AI products that disproportionately exclude users with certain problem-solving approaches? Recent research has begun to report the existence of such biases—but what do they look like, how prevalent are they, and how can developers find and fix them? To find out, we conducted a field study with 3 AI product teams, to investigate what kinds of AI inclusivity bugs exist *uniquely* in user-facing AI products, and whether/how AI product teams might harness an existing (non-AI-oriented) inclusive design method to find and fix them. The teams' work revealed 83 instances of 6 AI inclusivity bug types unique to user-facing AI products, their fixes covering 47 bug instances, and a new GenderMag inclusive design method variant, GenderMag-for-AI, that is especially effective at detecting AI inclusivity bugs when the AI's output is not necessarily believed.



## CCS Concepts

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

## Keywords

Intelligent User Interfaces, Human-Computer Interaction

**ACM Reference Format:**

Andrew A. Anderson, Fatima A. Moussaoui, Jimena Noa-Guevara, Md Montaser Hamid, and Margaret Burnett. 2026. "Over-the-Hood" AI Inclusivity Bugs and How 3 AI Product Teams Found and Fixed Them. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3742413.3789068




*IUI '26, Paphos, Cyprus*



ACM ISBN 979-8-4007-1984-4/26/03
https://doi.org/10.1145/3742413.3789068

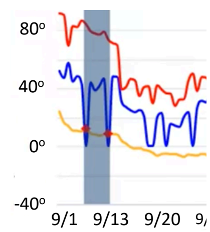




**Figure 1: An AI-powered agricultural tool for predicting when crops are at risk of dying from cold temperatures (y-axis) over time (x-axis). The risky events are where the AI-predictions for this crop (yellow line) intersect with the forecast's low temperature (blue line).**

## 1 Introduction

Suppose an Artificial Intelligence (AI) product team creates a dashboard to help farmers and other agricultural stakeholders apply AI to help find when temperatures sink so low that 50% of their crop will die (Figure 1). Farmers can use the dashboard to decide whether or not to deploy frost-mitigation techniques to keep their crops above this temperature.

However, the AI's user-facing information in Figure 1 may not be useful to all its intended users. Farmers and other agricultural stakeholders have diverse backgrounds, economic resources, gender identities, ages, education levels, and agricultural experience levels. Suppose the farmer does not have much engineering/math background or has limited finances to deploy expensive frost-mitigation techniques [52]. This raises the question of whether a wide range of diverse farmers can decide, using this tool, what to actually *do* with their fields. If not, who would it fail to serve, and how should the tool change to include them?

These are human-centered AI questions (HCAI). The HCI area of inclusive design aims to answer questions like these, so as to create user experiences that are "usable and understandable by as many people as possible, considering users' diverse needs, backgrounds, and experiences" [27]. Usability bugs that disproportionately exclude some groups from receiving a product's intended benefits are sometimes termed "inclusivity bugs" [41].

In the field of AI, some inclusivity bugs are "under-the-hood," such as algorithmic or training data biases. This paper instead considers "over-the-hood" inclusivity bugs in user-facing AI products. Recently, research has begun to report the presence of over-the-hood inclusivity bugs in AI products (e.g., [2, 21]), but has not yet investigated AI inclusivity bugs that are *unique to user-facing AI products*, which we term "*AI inclusivity bugs*." We distinguish AI inclusivity bugs from other usability bugs by the following defining criteria: (1) AI inclusivity bugs exist only in user-facing AI information, making them *AI* usability bugs. (2) AI inclusivity bugs disproportionately disadvantage certain groups of AI product users, making them AI *inclusivity* bugs.

In this paper, we report on a field study with three AI product teams to investigate AI inclusivity bugs: Team Game, Team Weather, and Team Farm. Team Game was working on explaining an AI-powered game that involved sequential decision making. Team Weather was working on AI-powered agriculture, whose prototype (Figure 1) was predicting whether and when to deploy frost-mitigation approaches. Finally, Team Farm was working on AI-powered irrigation scheduling. Through our field study with these AI product teams, we investigated the following research questions:

- **RQ1**: What *types of user-facing AI inclusivity bugs* do AI product teams find? What do they look like? How common are they?
- **RQ2**: How do AI product teams *fix* the AI inclusivity bug instances they find?
- **RQ3**: Is an existing inclusive design method (GenderMag in this paper), which was not designed particularly for AI, "enough" for AI product teams to be effective at finding user-facing AI inclusivity bugs, or is something AI-specific needed?
- **RQ4**: If something AI-specific is needed, how does the inclusive design method need to change? How effective are these changes?

## 2 Background: The Gender Inclusiveness Magnifier (GenderMag) Method

The inclusive design method the AI product teams used was GenderMag (Gender Inclusiveness Magnifier) [9]. The GenderMag method is an inclusive design and evaluation method to assist evaluation teams in improving technology products' inclusivity to diverse users. Multiple empirical studies have shown GenderMag's efficacy at identifying inclusivity bugs and pointing toward fixes [9, 20, 42, 49, 55, 63].

At GenderMag's core are five *problem-solving styles*, shown in Figure 2. Each style has a range of problem-solving style *values*, capturing diverse problem-solving approaches. These five problem-solving style types have repeatedly been shown in research to have strong ties to both problem-solving and gender [2, 9, 59, 63].

Figure 2 shows these five ranges of values, with distinguished endpoints (columns 2 & 4). The set of values in each column are grouped into one of three personas. The five problem-solving style values on the left are assigned to the "Abi (Abigail/Abishek)" persona. The five values on the right are assigned to "Tim (Timara/Timothy)." A mix of values are assigned to "Pat (Patricia/Patrick)."

The principle behind GenderMag is that, when technology *simultaneously* supports the "endpoint" personas Abi and Tim, every problem-solving style value within the ranges those endpoints define is also supported. If those who apply GenderMag identify that something in the technology does not support one of these endpoints, GenderMag defines that occurrence as an *inclusivity bug* [20]. These inclusivity bugs are *problem-solving* inclusivity bugs, since they disproportionately impact people with that problem-solving style value. The inclusivity bugs identified by the five problem-solving style types in Figure 2 are also gender inclusivity bugs. They are gender inclusivity bugs because these five problem-solving style types capture (statistical) gender differences in how people problem-solve [2, 9, 20, 59, 63] (e.g., Figure 3).

Software professionals use these personas in a specialized cognitive walkthrough. As with other cognitive walkthroughs [40], teams apply the GenderMag walkthrough by first choosing a use-case/scenario as an overall goal, and then answering questions about subgoals and actions that the team think the persona "should take" to accomplish the overall goal (Figure 4).

## 3 Related Works

In human-AI interaction, inclusivity and related concepts like fairness can be categorized into two broad focuses: (1) an "under-the-hood" algorithmic or data focus (i.e., detecting/fixing *algorithms or training data* when they cause harm to some groups over others) and (2) an "over-the-hood" focus on *diverse users' usage experiences with user-facing* aspects of AI products). There has been a host of literature for the first (e.g., [5, 8, 10, 18, 22, 31, 36, 69]), but this paper focuses on the second category.

AI usability is distinct from traditional usability, and this is in part because *designing* AI products is different from designing traditional software [70]. Yang et al. posits two sources of AI's distinctive design challenges, that then carry through to users: (1) AI products are nondeterministic on both *what* they do and *how* they do it. This makes AI products more challenging than traditional software for users to understand (e.g., user P06-PostGM from Hamid et al.'s study [21]: "...I don't understand how the probability of winning in move 7 with square F1 wasn't 100%." (2) The *continually evolving sources* of AI products' outputs are inherently complex (depending on not just code but also the latest training data, statistical inferences, this user's current situation, what this user did yesterday, ...). Given these complexities, for designers to design for every possible output is challenging—and when designers are taken by surprise, users can suffer [70].

One direction researchers have taken to address AI usability challenges is taking an analytical approach to assess and improve users' interactions with user-facing AI products. Guidelines are one example of an analytical approach, which provide general advice on how to improve human-AI interaction [1, 26]. Another analytical approach to evaluate human-AI interaction has been to establish frameworks, supporting structures for thinking about and doing human–AI interaction [39]. One instance of these frameworks is

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

**Figure 2: The five GenderMag problem-solving style types (rows), and the range of values for each GenderMag persona (columns). These problem-solving styles have empirically statistically clustered by people's genders (e.g., [2, 9, 20, 59, 63]).**

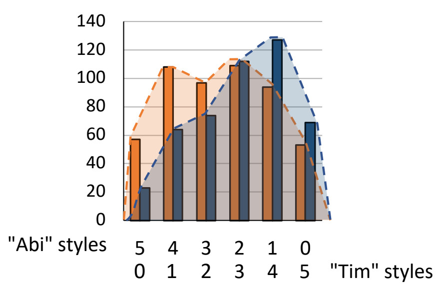


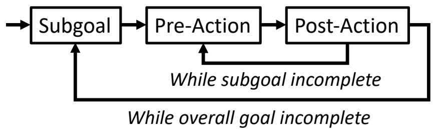

**Figure 3: Number of "Abi" problem-solving style values reported by the women (left bar in each pair, bright orange) and men (right bars, dark blue) in Anderson et al. [2]. X-axis: number of Abi problem-solving style values (Figure 2). Y-axis: number of participants having this number of Abi problem-solving style values. Statistically, men skewed significantly more toward "Tim" problem-solving style values than women did.**

Shneiderman's human-centered AI (HCAI) framework [57]. Their framework's two dimensions assessed (1) how much automation a system has (low ↔ high) and (2) who/what is in control of the system (computer ↔ human). It sought to highlight when designers might encounter negative outcomes for AI-powered systems, like excessive automation/human control. There are myriad examples of both guidelines and frameworks for human-AI interaction (e.g., [51, 60, 64, 68]), but these approaches leave teams to decide how to go about applying them to their AI products. Our approach is neither a set of guidelines nor a framework, but rather a systematic process for teams to apply to their AI products to improve diversity and inclusion.

As with our work, some analytical work uses personas to help evaluate tech products. An abundance of persona work involving AI has focused on how AI can help generate personas (e.g., [11, 24, 56, 61]), rather than evaluating AI products using personas. One notable exception is Joshua Puglisi's thesis [50], which constructed four personas to help generate design requirements, evaluated an AI sentiment analysis tool with these personas, and validated the personas' findings with a user study. Their personas incorporated dimensions of personality (i.e., the Myers-Briggs four personality dimensions [43]), their goals/frustrations, need for cognition, tech savviness, and how social they were. (The "need for cognition" is the dimension most closely aligned with the information processing style problem-solving style presented in our work.) Our work differs from Puglisi's by focusing on AI inclusivity bugs' disproportionate impact on certain diverse problem-solvers, instead of the impact of personality traits.

Other researchers have used analytical evaluation methods such as cognitive walkthroughs while studying AI. One approach has been to incorporate AI *into* a cognitive walkthrough, such as Bisante et al. [6], who embedded Open AI's Generative Pre-trained Transformer (GPT) into their tool for cognitive walkthrough evaluations, named CWGPT. They found that CWGPT independently identified task sequences in the walkthrough, and expert evaluators agreed with 116/128 issues that CWGPT found (93.6%).

In contrast, in our work *human* teams evaluate AI products, and several other AI researchers have also reported human activities with cognitive walkthroughs. For example, De Santana et al. [13] used a cognitive walkthrough as a debriefing tool when users interacted with AI, identifying issues with recency bias, confirmation bias, and trust in the system. Yildirim et al. [71] reported on 21 student teams who applied cognitive walkthroughs to a conversational agent. These teams found a wide range of usability problems,




**Subgoal:** *Will [persona] have formed this sub-goal as a step to their overall goal?*

**Pre-Action:** *Will [persona] take this action at this step? Yes/no/maybe, and why?*

**Post-Action:** If *[persona] does the right thing, will know that they did the right thing and are making progress toward their goal? Yes/no/maybe, and why?*

**Figure 4: (Left): The GenderMag walkthrough graph. Each node represents a step in the walkthrough and is repeated in a loop until the overall goal is complete. (Right): The GenderMag walkthrough questions at each step.**

identifying how to improve conversational interactions with the AI. In contrast, our field study reports on *professional AI development teams* using the full GenderMag method, which integrates personas and a specialized cognitive walkthrough to *find and fix AI inclusivity bugs*.

Finally, a few researchers have focused on users' experiences with user-facing AI through a diversity and inclusion lens. Some works have uncovered gender differences in perceived fairness of an AI product [62], probability of adopting an AI product [46, 47], likability of an AI product [14], and awareness of how an AI product operated [30]. Some investigations have considered the five GenderMag problem-solving style types described earlier in Section 2. For example, Kulesza et al. [34] measured the change in their participants' computer self-efficacy, given a "why"-oriented explanation approach. Jiang et al. [29] found that users with higher self-confidence were less likely to accept an AI's proposed solution. Similar empirical findings have also been found for attitudes toward risk [12, 53], information processing style and learning style [44, 72], and motivations [37, 54, 58]. Vorvoreanu et al. applied GenderMag to an academic search engine, revealing 10 inclusivity bugs, and fixing six of them [63]. Hamid et al. conducted an empirical study of users interacting with a before-GenderMag vs. after-GenderMag AI-powered game, to determine which was more inclusive to problem-solving users like the Abi and Tim personas [21]. They found that the after-GenderMag product was more inclusive, both by persona and by gender. Anderson et al. [2] considered all five of the problem-solving style types discussed in this paper, finding inclusivity and equity differences for all five types.

However, none of these works investigates *how* the generators of these user-facing products, *professional AI teams* evaluating their *own* AI products, go about using such a method to find and fix their AI products' user-facing *AI* inclusivity bugs (i.e., inclusivity bugs communicated from the AI to the user). That is the gap this paper aims to fill.

## 4 Methodology

To answer our research questions, we invited AI product teams who attended a series of AIVO (a virtual organization consisting of the 29 NSF AI Institutes) meetings to evaluate their AI products' inclusivity, and 3 teams stepped forward: Team Game, Team Weather, and Team Farm. We did not offer any monetary compensation.

### 4.1 Team Game

Team Game was a team of computer scientists at a large US university, working on an eXplainable Artificial Intelligence (XAI) approach for sequential decision-making domains. Four team members identified as men and one as a woman. Team Game wanted to investigate inclusivity improvements to their explanation design.

Team Game's domain was sequential decision making for M-N-K games. M-N-K games are played on $M \times N$ sized boards, where players either (1) win by constructing sequences of length $K$ or (2) draw when no more empty squares remain on the board. Tic-Tac-Toe is a well-known 3-3-3 instance of M-N-K games. Team Game investigated 9-4-4 games (board size: $9 \times 4$, win sequence length: 4).

In Team Game's prototype, which was based on Dodge et al.'s source code [15], each game player was an AI agent with a convolutional neural network trained to play 9-4-4 games (Figure 5). The gameboard (Figure 5, left), was how users saw the games progress. For each move, the X-player AI or O-player AI placed one of their game pieces in one of the 36 squares. Each move was labeled with a move number in the top-right of each occupied gameboard square, to help users remember/see which moves had come before which other moves. When one of the players won, the winning sequence of four squares were highlighted in the winning player's color (the X-player AI's blue cells $D2 - G2$ in the gameboard).

The X-player AI decided on its next move by choosing the move with the best score, which it calculated as follows. For each move the X-player AI made, the AI took the current game information as input and calculated three probabilities for each square: the probability it would eventually win the game if it took that square (P(Win)), that it would eventually lose if it took that square (P(Loss)), and that the game would end in a draw (P(Draw)). The X-player AI's "score" for each gameboard square was: $Score = P(Win) - P(Loss)$.

Team Game had three explanations to explain the X-player AI's reasoning to users, updating after each of its moves. The first was Scores Best-to-Worst explanation (Figure 5, top right), which shows how the X-player AI sorted all 36 of its score evaluations in a monotonically decreasing line from the best (left) to worst (right). The second explanation was Scores Through-Time explanation (Figure 5, middle right), where the X-player AI provided distributions of all 36 scores through time (moves); which was designed to show users how "confident" the AI was through time. In the last one was Scores On-the-Board explanation (Figure 5, bottom right), which displayed miniaturized Scores Through-Time explanations for all 36 squares. This provided simultaneous temporal and spatial score information.

### 4.2 Team Weather

Team Weather was a team of professional developers and academic researchers from a different large US university than Team Game.

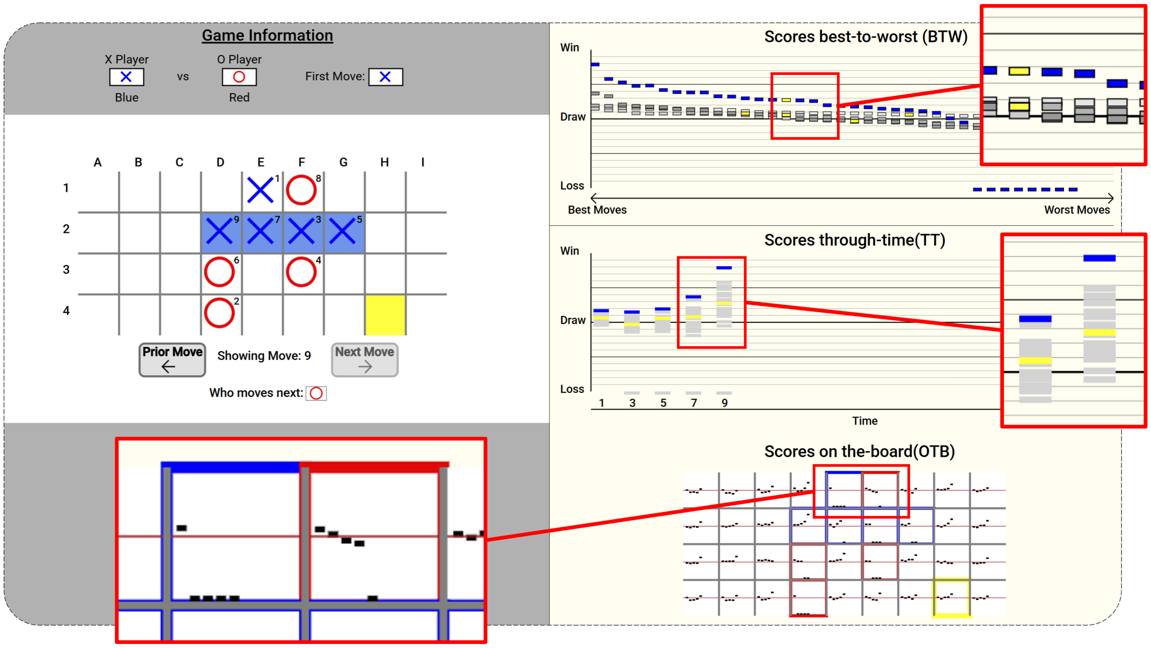



**Figure 5: Team Game's eXplainable AI (XAI) interface. On the left is Team Game's gameboard, on the top-right is the Scores Best-to-Worst explanation, on the middle-right is the Scores Through-Time explanation, and on the bottom-right is the Scores On-the-Board explanation. Call-outs with enlarged portions have been superimposed for readability.**

They had six team members, three of whom had a computer science background, and three with natural sciences background (environmental science, atmospheric science, and meteorology). Four members identified as men, and the remaining two identified as women. Team Weather's application was already in use across the US state of <anonymized>, and the team wanted to improve it for diverse end users.

Team Weather's AI product's recommendations aimed to help agriculturalists protect their wine grapes from grape injuries sustained from exposure to cold temperatures. This ability of grapevines to survive cold temperatures is known as their cold hardiness, using their Low-Temperature Exotherm (LTE). Agriculturalists use these estimated temperature thresholds, which measure the point where they would lose 10% ($LTE_{10}$), 50% ($LTE_{50}$), and 90% ($LTE_{90}$) of their crop yield. To avoid this, agriculturalists deploy expensive, preemptive frost mitigation methods when the weather forecast estimates a drop below one of these thresholds.

Team Weather's AI was a Recurrent Neural Network (RNN), The RNN's input was a table of weather information for field locations over time. Its rows contained air temperature, LTE values, humidity, and dewpoint information, all used to predict when a cold hardiness event might occur. The model's learning goal was predicting a sequence of LTE estimates for multiple wine grape varieties on a given day. That LTE would be compared to the low-temperature forecast for that day, helping agriculturalists decide whether and when to deploy frost mitigation methods.

Figure 6 shows the AI's information in Team Weather's interface. The AI predicted $LTE_{10}$ values (yellow line), and these values were compared against the forecasted low-temperature (blue line) over time (x-axis). The AI's recommendations for when agriculturalists should deploy frost-mitigation techniques occurred at the intersection of these two lines, marked by red diamonds.

## 4.3 Team Farm

Team Farm was a research team working on AI-powered agricultural applications. This team had two members, an engineering technician with no computer science background and a data scientist; one identified as a man, the other a woman. Team Farm wanted to make their AI product more inclusive to agricultural users.

Team Farm's domain was irrigation scheduling for wine grapes. To support growth of wine grapes, agriculturalists have to give different volumes of water to different grape varieties. Agriculturalists have monitored grapes through sensor telemetry, which gather information about the grape vines themselves or field block soil. Agriculturalists develop irrigation schedules from their experiences with historical sensor data on their fields. However, these may not always accurately reflect the needs of wine grape varieties, which can change across growing seasons. One example of this comes towards the end of a growing season, where agriculturalists intentionally reduce irrigation frequency (i.e., *deficit* irrigation [65]) to increase wine grape quality.

Team Farm created their initial dashboard (Figure 7) in a high-fidelity prototype, since they were just starting to gather sensor data to train an AI. Team Farm had deployed six sensors across eight fields (Figure 7's teardrops). Three sensors measured soil water content over time (the three lines in the top-right graph). The other three sensors measured matric potential over time (the three lines in the bottom-right graph). Their vision for their AI product

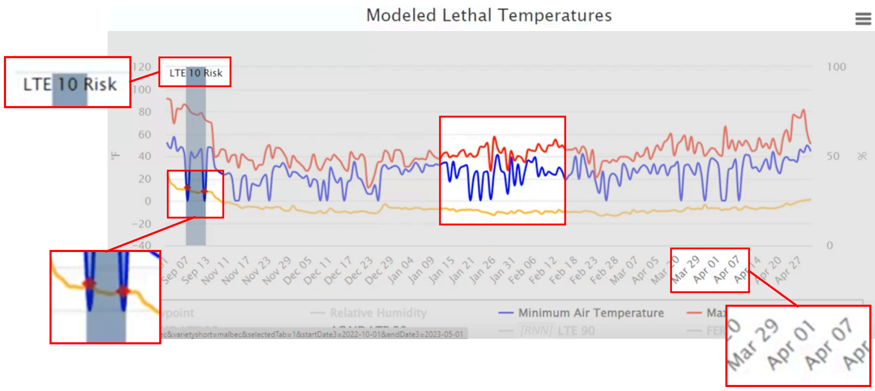



**Figure 6: Team Weather AI's $LTE_{10}$ cold-hardiness predictions (yellow line) for different days (x-axis). The $LTE_{10}$ predictions were compared to minimum forecasted temperature (blue line), and when $LTE_{10}$ > the minimum, the interface marked these events with a red diamond to help agriculturalists decide whether and how to deploy frost-mitigation methods.**

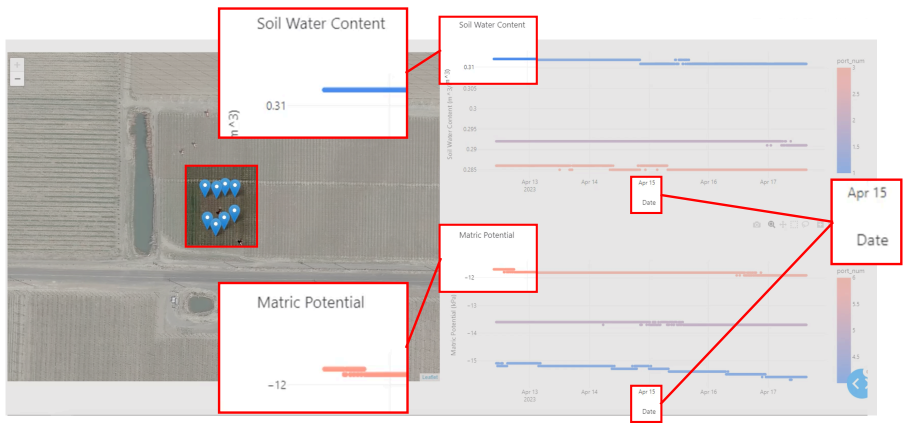



**Figure 7: Team Farm's recreated AI-powered irrigation scheduling prototype. Their AI would use temporal sensor telemetry of soil water content (top-right) and matric potential (bottom-right) from eight fields (left) to provide decision support for agriculturalists of when (and how much) to irrigate each field.**

was to use these two sources of data to predict how much the soil would saturate and how long it would take until the field needed more irrigation. This information would support agriculturalists' irrigation scheduling by explicitly considering wine grape variants' dynamic needs throughout a growing season.

## 4.4 Procedures

We conducted each field study session with one team at a time using the Zoom video conferencing technology. For each team, after obtaining the team's consent using our IRB-approved informed consent form, we began the meeting, which was a pre-evaluation session that introduced the team to GenderMag, the personas and their associated five problem-solving style types, and how GenderMag worked (Section 2). Each team then picked the persona they wanted to evaluate their AI products with; all three teams picked the "Abi" persona (Figure 8). Each team then customized Abi's age, location, pronouns, and background/skills using the persona tool on the GenderMag website [17] to make Abi a good fit to their AI product's target audience. For instance, Team Game made their Abi a 27-year-old mechanical engineer with she/her/hers pronouns who used AI tools to keep organized. Team Farm made their Abi a 49-year-old farm manager who used he/him/his pronouns. All three teams' customized personas can be found in the supplemental documents.

After a team's initial session, subsequent sessions proceeded as per Figure 9. As the figure shows, teams ran a series of "find" sessions, applying a version of GenderMag (at first, this was the "Original GenderMag") to their AI products. Each find session had a facilitator to keep the session on track, a driver to navigate the interface, a recorder to take notes using GenderMag's walkthrough forms, and evaluators to evaluate the system. In these sessions, 1-2 researchers joined the team as participant-observers, to help with facilitating (making sure each team member weighed in on every evaluation step), recording, and evaluating. The result of these find sessions was a set of inclusivity bug instances (i.e., anytime when a team answers "maybe" or "no" and why in GenderMag's

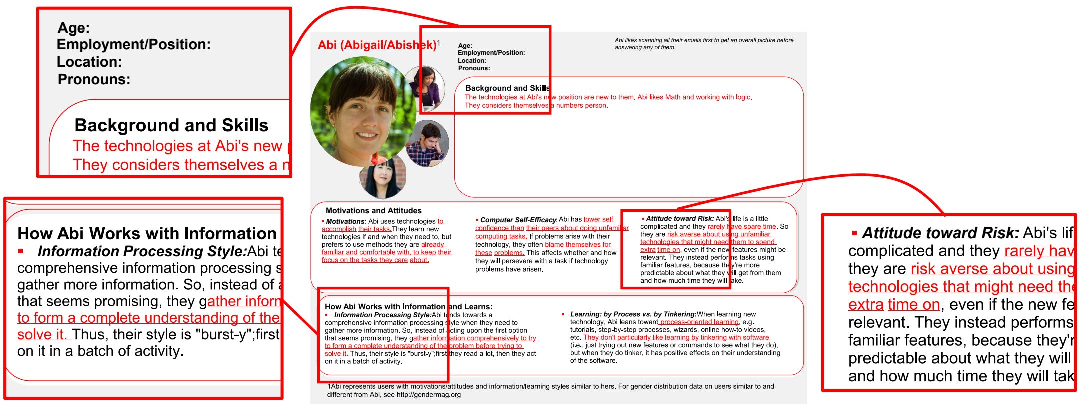



**Figure 8: The "Abigail/Abishek" (Abi) persona template, which the teams customized to represent their users. They could customize the persona's age/employment/location/pronouns, as well as their background and skills (top-left). However, they could not customize any of the five problem-solving style values (bottom half).**

walkthrough form questions (Figure 4)). These sessions enabled us to answer RQ1 and contributed in part to RQ3.

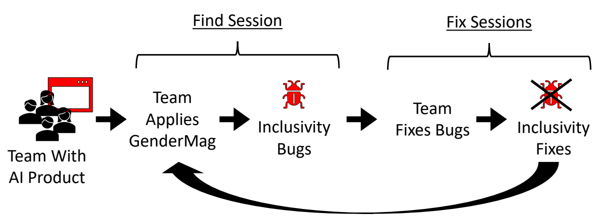



**Figure 9: The field study methodology. Three teams developing AI products iterated through find/fix sessions. Find sessions: teams used a GenderMag walkthrough to evaluate their AI product, producing an inclusivity bug list. Fix sessions: teams fixed inclusivity bugs until they said they were done, prompting another find session.**

Once each team had a set of inclusivity bug instances, they transitioned to "fix" sessions (the right half of Figure 9). Before their first fix session, we gave them a link to the GenderMag design catalog [16] as a potential resource. 1-2 researchers joined the fix sessions, but only as observers, without providing any input as to how to fix the bugs. Thus, the teams alone chose whether and how to fix bug instances, applied their own prioritization processes and design processes, and used whatever technology they wanted to produce the bug fixes. Team Game directly sketched low-fidelity concepts using Zoom's annotation tool, verbally describing their proposed fixes as well. Team Farm modified their high-fidelity PowerPoint prototype directly. Team Weather verbally described their fixes, programming the fixes directly into their AI product. The teams' fixes enabled us to answer RQ2.

After a team had used Original GenderMag on their AI product for one or two sessions, we asked whether they wanted to change the process; this completed the answer to RQ3. We then used the teams' change ideas to create new "Gendermag-for-AI" variants, which the teams began using instead of the Original GenderMag in Figure 9, enabling us to answer RQ4.

## 4.5 Data Analysis

To answer RQ1, we needed to analyze the teams' GenderMag walkthrough forms (described above) to categorize the AI inclusivity bug types that emerged from their work. To do so, we used Hsieh and Shannon's [25] conventional content analysis on the teams' "find" session data. In conventional content analysis, categories are extracted directly from the text data. In our case, the AI product teams had already flagged where these were by answering "maybe" or "no" on their subgoals, pre-action, and post-action GenderMag walkthrough forms they filled out during their evaluation sessions.[1] Through affinity diagramming, we incrementally added teams' walkthrough form responses that flagged bug instances, grouping those which contained similar concepts and labeling those categories. When no more new categories emerged from added data (i.e., data saturation), we stopped adding responses. This revealed six potential AI inclusivity bug types, shown later in Section 5's Table 1, which became our codeset.

We used this codeset to qualitatively code the complete set of evaluation forms. The coding rules and examples used for coding the data are included as a codebook in the supplemental documents. To ensure consistency of our use of these codes, two authors each independently coded 20% of the data, which resulted in 84% agreement under the Jaccard index[2] [28]. Given this level of interrater reliability, one author coded the remaining data. The results answered RQ1, and also contributed to the answer to RQ4.

Answering the remaining research questions did not require qualitative coding, because human judgment was not needed. For RQ2, the teams had already been explicit about which bugs they decided to fix and how with sketches and screenshots of their fixes. For RQ3 and RQ4, the teams explicitly told us changes to the method

[1]We analyzed only the "maybe" and "no" responses because the "yes" response indicates no problem.

[2]Jaccard index between two sets: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$

they would like to see which, combined with the coding results above, completed the answer to RQ4.

# 5 Results RQ1 & RQ2—Six AI Inclusivity Bug Types and the Fixes

We begin with the "bottom line" for RQ1. As Table 1 shows, teams' "find" sessions revealed six AI inclusivity bug types and 83 bug instances. Recall from Section 1 that an "AI inclusivity bug" is a usability bug that also meets two criteria: (1) The teams explicitly identified the bug in the AI's information, which made the usability bug an *AI* usability bug. (2) The teams also tied at least one of the persona's *problem-solving style values* (e.g., risk-aversion, comprehensive information processing style, etc.) to the AI usability bug. An AI usability bug tied with a problem-solving style value means that users with *that* problem-solving style value would be likely to be disproportionately impacted by the bug, which makes an AI usability bug an AI *inclusivity* bug.

As already pointed out in Section 3, two attributes particular to AI products open opportunities for AI usability bugs to arise—(1) AI products' nondeterminism in what they do and how they do it, and (2) the complex, constantly evolving sources of AI products' outputs [70]. If AI usability bugs not only arise, but do so inequitably, these become AI inclusivity bugs. For example, if an AI product is unclear on what it is doing with what data, an *AI: why should I?* bug could result, disproportionately tied with risk-aversion. In fact, Team Weather found exactly this type of bug in their AI product and tied it to Abi's risk-aversion: "Abi doesn't know ... information generally or specific to Abi's farm?" We detail these kinds of AI inclusivity bugs next.

## 5.1 Three AI Inclusivity Bug Types & Risk-Aversion

Considering risk-averse users was a powerful AI inclusivity bug-finding aid to the teams. In their evaluation sessions, the teams referred to risk-averse users more often than any other problem-solving style, and for three of the AI inclusivity bug types—*Interpret AI?*, *AI input↔output?*, and *AI: why should I?*—risk was the top (or tied for top) reason.

### 5.1.1 *The Interpret AI? Inclusivity Bug.*

"What does this even mean?" *Interpret AI?* was the most frequent AI inclusivity bug type. In total, the teams identified 27 instances of it. In these bug instances, the teams decided that the AI's information could be difficult for their populations to interpret, and they expected this AI inclusivity bug to particularly impact risk-averse users (10/27 instances). They also expected this bug to affect comprehensive information-processing users, citing it 10 times as well, sometimes in combination with risk-aversion.

The attitudes toward risk problem-solving style is nuanced. It can include not only well-known technology risks tied with privacy/security, but also risk of producing low-quality work, of wasting too much time, of failing to succeed in harnessing the product's hoped-for benefits, and more [32, 59]. In their evaluations, the teams spoke frequently of the latter two aspects. For example, the risk of wasting time without obtaining benefits figured prominently in Team Game's evaluation of Figure 10. The figure shows the "before" state (left) and the "after" state (right) for one of Team Game's *Interpret AI?* inclusivity bug instances. The after state (right) is where Team Game identified the *Interpret AI?* instance.

In the before state (left), both the X-player AI and O-player AI had already made the moves shown on the gameboard. The Scores Best-to-Worst explanation (bottom left) showed why the X-player AI made that move: the X-player AI had evaluated all 36 squares before it moved, which it explains by drawing a line of blue rectangles from "best" move (far left) to "worst" (far right). When the O-player AI responded with the move shown left, the X-player AI re-calculated scores and moved as shown right, which it explains by the updated explanation on the bottom right. The blue rectangles are the new move's calculations, and the gray rectangles are the previous move's calculations, to show what has changed. But Team Game decided the Scores Best-to-Worst explanation was not particularly interpretable, and decided that Abi's perceived risk of time-wasting would outweigh Abi's interest in learning the AI's "process":

> Team Game: "*...wants to go through the process but does not have the context to forge the relationship. The time investment... to understand what's going on [vs.] the perceived benefits...*"
> • Bug Type(s): *Interpret AI?*
> • Problem-Solving Style Value(s): *Risk-Aversion, Process-Oriented Learning.*

Team Game's reasoning is consistent with Blackwell's Model of Attention Investment [7]. According to this model, technology users decide whether to spend their attention the same way they decide to spend money: using their own expectations of the cost, of the benefits they can expect, and of the risk (probability) that they will spend the cost but not gain any benefits. Team Game's focus on Abi's expection of wasted time (high probability of risk), despite her interest in the hoped-for benefits, fits Blackwell's model well.

Fortunately, once *Interpret AI?* bugs were found, they were often straightforward to fix using standard HCI techniques, here by following the Blackwell model. The Blackwell model suggests that the key to resolving *Interpret AI?* bugs is increasing/clarifying the benefits and/or reducing risks/costs in learning how to interpret the AI's output. And that is what Team Game did, via the risk route: they added a legend (Figure 11) to reduce the risk of failing to learn what the explanation meant. In fact, all the teams fixed the *Interpret AI?* bug in ways consistent with the Blackwell model, by increasing the benefits or reducing the attention costs and/or risks (e.g., via adding/improving legends, info boxes, labels) of spending time on uninterpretable AI output the team had spotted.

### 5.1.2 *The AI input↔output? Inclusivity Bug.*

"What does this (AI-input) have to do with that (AI-output)?" In the same moment of play shown above in Figure 10, Team Game found another AI inclusivity bug, *AI input↔output?*. This type of bug is a lack of clarity about whether/how the AI's input(s) relate to its output(s):

> Team Game: "*The relationship between squares on the board and the Best-to-Worst explanation isn't clear...*"
> • Bug Type(s): *AI input↔output?*
> • Problem-Solving Style Value(s): *Risk-Aversion*

Teams seemed to regard most instances of *AI input↔output?* as specialized instances of *Interpret AI?*, with their verbalizations suggesting both on 7/9 of the *AI input↔output?* instances. Thus, their

| AI Inclusivity Bug Type | Definition | Example | # Bugs (%) | Section# |
|---|---|---|---|---|
| Interpret AI? What does this even mean? | Persona had difficulty interpreting AI's output, left wondering "why" it was that way, or had difficulty understanding what was going on. | Team Weather: "*Confused about the line that is there, where you have LTE10 at the top, but we are working with LTE50.*" | 27 (33%) | 5.1 Risk |
| AI input ↔ output? What does this (AI-input) have to do with that (AI-output)? | The relationship between the AI's inputs & outputs is unclear/confusing to the persona. | Team Weather: "*... might understand the information ... but doesn't know how they tie together.*" | 9 (11%) | 5.1 Risk |
| AI: why should I? Why even look at this? | Why persona should interact with the AI's information is unclear. | Team Farm: "*...wouldn't know that spending time on the graph is going to give them anything.*" | 19 (23%) | 5.1 Risk |
| AI: more info! Need more info! | Insufficient detail in AI information for persona to make use of it, or persona needed more explanation for why information appeared. | Team Farm: "*There is not enough information...*" | 9 (11%) | 5.2 Info. Proc. |
| AI: actionable? So? What should I DO? | Not clear to persona how to access AI's information or actionable steps. | Team Game: "*...would not quite know what to do at this point...*" | 12 (14%) | 5.2 Info. Proc. |
| AI changes? What's changed? | Unclear to persona how an AI's output changed. | Team Game: "*Nothing has changed from the previous explanation...*" | 7 (8%) | 5.3 Self-Eff |
| Total | | | 83 (100%) | |

**Table 1: The six AI inclusivity bug types (rows). Together, the teams found 83 instances of these bug types.**


**Figure 10: An *Interpret AI?* instance (Team Game). Left: Before Bug—the gameboard (top) and Scores Best-to-Worst explanation (bottom, enlarged for readability). Right: Discovered Bug—Team Game noted that the explanation changes (the two lines of rectangles in the explanation) were not straightforward to interpret.**

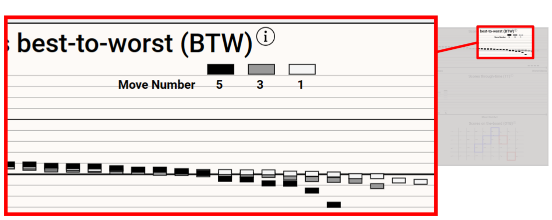



**Figure 11: An *Interpret AI?* fix (Team Game). To improve the interpretability of the Scores Best-to-Worst explanation, Team Game added a legend showing what each color means.**

solutions to *AI input↔output?* tended to be specialized instances of solutions to *Interpret AI?*.

Specifically, they again added information to increase the user's perceivable benefits or reduce their apparent attention costs and/or risks (e.g., via adding/improving legends, info boxes, labels)—but for *AI input↔output?*, the new information explicitly connected the AI's input to the AI's output. For example, Team Game's solution was to show the connection dynamically. Whenever users highlighted a score rectangle in the explanation, the corresponding gameboard square became highlighted and vice-versa, and a tooltip (Figure 12) would appear to connect the gameboard state (the AI's inputs) to the score rectangles on the Scores Best-to-Worst explanation (its output).

*5.1.3 The AI: why should I? Inclusivity Bug.* "Why even look at this?" *AI: why should I?* was the third AI inclusivity bug type associated with risk-aversion particularly frequently (10/19 instances). These bugs differ from the first two: the first two show at least some user

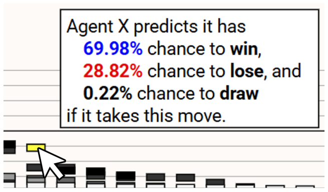



**Figure 12: A fix for one of Team Game's *AI input↔output?* instances. The purpose of the fix explicitly mapped how the X-player AI used its input (moves on the gameboard) to calculate its output (win, lose, draw probabilities).**

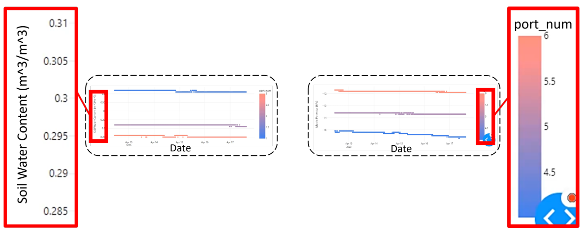



**Figure 13: An *AI: why should I?* instance (Team Farm). The AI used six real-time sensor readings to make irrigation recommendations, three sensors each for soil water content (left) and soil matric potential.[3] The colored lines show each sensor's readings, buried at different depths. The port_num (right) identifies each sensor.**

interest in engaging with the AI's information, whereas this bug depicts some risk-averse users not even seeing the point of trying.

Figure 13 shows an example of this bug type, which Team Farm found in a pair of irrigation recommendation visualizations based on soil measurements. These soil measurements are critical in measuring whether soil can support crop growth, but Team Farm raised concerns that Abi would not know this, and might give up instead of wasting time trying to understand the information:

Team Farm: "*...[Abi] wouldn't know that looking at this graph for prolonged periods of time is going to help them understand the irrigation issues.*"

- Bug Type(s): *AI: why should I?*
- Problem-Solving Style Value(s): *Risk-Aversion*

In a later version of their graphs (Figure 14, left), Team Farm fixed a bug like this by indicating that interaction was available. They first added a blue rectangle to indicate that it was clickable (Figure 14, right). For comprehensive information processors like Abi, these clickable instances hint at the possibility of acquiring more information. Once Abi interacted with this area, Team Farm provided Abi with information about how much and when to irrigate their field:

Team Farm: "*We can make [the graph] look more clickable. We know that Abi doesn't like to take risks... There's no clear indication that hovering over water content is going to lead to irrigation decisions...*"

[3]"*Soil matric potential represents the change in energy state of soil water relative to a reference, and it is the component of water potential attributed to the effects of capillary and adsorptive forces acting between liquid, gas, and solid phases*" [48].

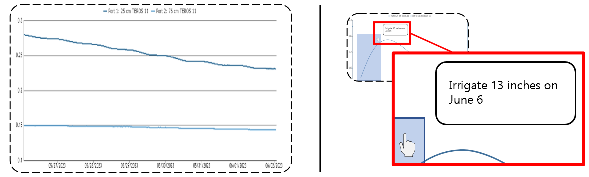



**Figure 14: An *AI: why should I?* fix (Team Farm). (Left): Pre-fix, the information did not make clear why Abi should engage with the decreasing sensor readings (risk-aversion). (Right): Team Farm's fix aimed to reduce Abi's time cost to scan the graph (blue rectangle) by making clear why they should—to irrigate the field (the tooltip).**

This kind of fix can be regarded as an instantiation of the Surprise-Explain-Reward strategy [66]. That strategy, inspired in part by Loewenstein's work on curiosity [38], attempts to pique the user's curiosity by surprising them "just enough" to inspire them to engage with a particular feature if/when they need to without interrupting them. The surprise's job is to deliver the user to a suitable explanation that hints at the benefits that ensue if they further engage, and the "reward" is a real-world benefit that the product feature delivers (here, showing exactly when to irrigate). Team Farm and the other teams may not have known about the Surprise-Explain-Reward strategy, but many of them used it by adding information and interactions to give users like Abi more reason to engage with the AI product feature.

## 5.2 Two AI Inclusivity Bug Types & Information Processing Style

*5.2.1 The AI: more info! Inclusivity Bug.* "Need more info!" Teams found nine instances of *AI: more info!*. Over half of the time (5/9 instances), they decided that these would particularly affect comprehensive information processors like Abi.

Figure 15 (left) shows one of Team Farm's *AI: more info!* instances. The AI's assessment (tooltip) of Plot A's soil status was "OK," so Abi would not have to irrigate this plot of Cabernet Sauvignon grapes. However, Team Farm questioned whether this summary was oversimplified, insufficient for Abi's comprehensive information processing style:

Team Farm: "*There may not be enough information in the tooltip to give the impression that the 'OK' is OK.*"

- Bug Type(s): *AI: more info!*
- Problem-Solving Style Value(s): *Comprehensive Info. Proc.*

This was reminiscent of Kulesza et al. [35], who found that if an AI's explanation is too simple, it could lead to increased mental demand and decreased trust in the explanation.

Team Farm regarded this AI inclusivity bug as critical, because it prevented users like Abi from going to Plot B (Figure 15, right), which was "DRY" and *needed* Abi to irrigate it. Since the overview was insufficiently detailed, Team Farm decided that Abi would instead try to validate the irrigation status by investigating the sensor readings from Figure 13:

Team Farm: "*Abi still may want to know if the status is correct by checking the actual graphs thoroughly..*"

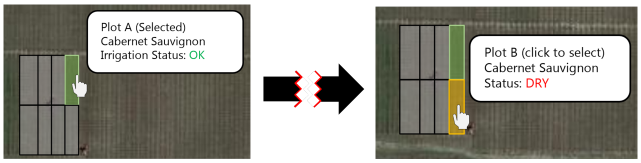



**Figure 15: An *AI: more info!* inclusivity instance (Team Farm). The team wanted Abi to see an important need for action by transitioning from the "OK" Plot A (left) to the DRY Plot B (right). However, the information was insufficient (i.e., just "OK"), preventing Abi from forming this subgoal (broken arrow).**

- Bug Type(s): *AI: more info!*
- Problem-Solving Style Value(s): *Comprehensive Info. Proc.*

While Team Farm did not fix this *AI: more info!* bug, teams fixed four other instances by adding more information in the AI's inputs and/or output where teams (through Abi's problem-solving styles) decided it was not enough. The teams tied all four of these bug instances to Abi's lower self-efficacy or comprehensive information processing style.

*5.2.2 The AI: actionable? Inclusivity Bug.* "So? What should I DO?" Teams found 12 instances where it was not clear how users should act upon an AI's information. Almost half of the time (5/12 instances), these teams decided this would particularly affect comprehensive information processors like Abi.

When Team Farm later evaluated their solution to the graph in Figure 14 (right), they decided that their solution was not enough to hint what Abi should do:

> Team Farm: "*Will Abi click? No. Will Abi move the cursor? Probably. Exactly on top of the rectangle? Probably not. If I (Abi) don't really understand what that rectangle is for...* "
> - Bug Type(s): *AI: actionable?*
> - Problem-Solving Style Value(s): *Comprehensive Info. Proc., Risk-Aversion*

Figure 16 shows how Team Farm tried to make it clearer for Abi what to do about the AI's information. They replaced the rectangular object with an icon of a faucet with a water droplet to make it more apparent what to do and when. Other teams made similar fixes by giving more hints/directions on what users should do through instructions, changing button/icons, or clarifying language.

> Team Farm: "*Is Abi going to hover over the rectangle? There's no rectangle anymore. Now we have an indication point. Is Abi going to hover over the indication point? Yes... I wanted an icon that's representative not just in computer science, but in anything that you have.*"

## 5.3 An AI Inclusivity Bug Type & Self-Efficacy

"What's changed?" These teams associated problem-solving values only 8 times with the *AI changes?* bug, but *six* of them were Abi's lower self-efficacy. Teams identified such instances when it was not clear how the AI's information changed through time. Teams decided that this deficit in the visualization made those with lower self-efficacy think they had done something wrong or would blame themselves:

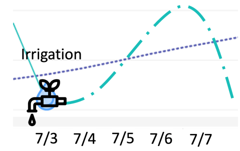



**Figure 16: An *AI: actionable?* fix (Team Farm). Team Farm added information that told the user what to *do* with the AI's information. They added an icon to reflect *when* Abi should irrigate (x-axis) and showed the outcomes of irrigating (dashed line, y-axis).**

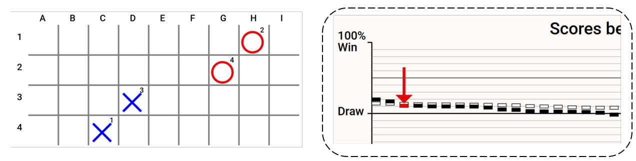



**Figure 17: An *AI changes?* fix (Team Game). Each time a move is made by the AI players, the score of that move is temporarily highlighted with an arrow in the explanations. On the right, Scores Best-to-Worst explanation shows how O-player AI scored on move 4.**

> Team Farm: "*Abi would struggle to differentiate between the graphs (previous vs. now) and would blame themselves, wondering if they selected the previous plot or not.*"
> Team Farm: "*When Abi clicks on the thing, it's going to lead them to believe they've done something wrong because nothing changes even upon clicking.*"
> Team Game: "*The game has progressed. There was no update for the other player. Perhaps she may consider she may have done something wrong.*"
> - Bug Type(s): *AI changes?*
> - Problem-Solving Style Value(s): *Lower Self-Eff.*

The literature provides evidence to support these potential consequences of failing to support users with lower self-efficacy [3]. For example, research has shown that people with lower self-efficacy, such as Abi, may be less willing to explore features new to them [19]. Others have found that such people may be more likely to abandon the system when they think barriers are "too high" [59], something Team Farm also identified:

> Team Farm: "*Because of Abi's self-efficacy, clicking is going to lead Abi to believe it's the wrong move and abandon the system. It's become more trial-and-error...*"
> - Bug Type(s): *AI changes?*
> - Problem-Solving Style Value(s): *Lower Self-Eff.*

When fixing this bug type, teams generally changed or added elements in the interface to grab attention when an action was taken. Team Game fixed their above bug (no changes for other player) with two additions (1) a legend clarifying how changes appear in the Scores Best-to-Worst explanation (recall Figure 11) and (2) adding colored arrows to show how the latest move the X-player AI or O-player AI picked now scored in each explanation (Figure 17). The temporary appearance of arrows along with the explanation legend aimed to point out what changed in the interface.

## 5.4 Did the fixes help? An external empirical investigation

In total, the teams fixed 47 of the 83 AI inclusivity bugs (Table 2). But did the fixes actually help end users of the AI products?

| AI Inclusivity Bug Type | # Bugs Found (%) | # Bugs Fixed (%) | % Bugs Fixed (%) |
|---|---|---|---|
| *Interpret AI?* | 27 (33%) | 14 (30%) | 14/27 (52%) |
| *AI input↔output?* | 9 (11%) | 7 (15%) | 7/9 (78%) |
| *AI: why should I?* | 19 (23%) | 10 (21%) | 10/19 (53%) |
| *AI: more info!* | 9 (11%) | 4 (9%) | 4/9 (44%) |
| *AI: actionable?* | 12 (14%) | 8 (17%) | 8/12 (67%) |
| *AI changes?* | 7 (8%) | 4 (9%) | 4/7 (57%) |
| Totals: | 83 | 47 | 47/83 (57%) |

**Table 2: The AI inclusivity bug instances teams found and fixed, by type. The teams fixed 47/83 bug instances, fixing each bug type at similar rates.**

Team Game worked with us to empirically find out the answer for their AI product. That empirical study has been presented elsewhere [21], but we briefly summarize it here because of its direct pertinence to Team Game's fixes. The study was a between-subjects lab experiment, in which 69 participants with no formal AI background worked with Team Game's prototype. Half worked with the original (pre-GenderMag) prototype and the other half worked with the fixed version [21]. The results showed that participants using Team Game's Post-GenderMag version had significantly better mental models (conceptual understanding) of the AI's reasoning than participants who used the Original version (Figure 18). A second measure was participants' ability to predict the AI's next move, and for this measure, there was no significant difference between groups. A third measure was gender equity, and here the difference was again significant. In fact, the fixes in the Post-GenderMag version improved the gender equity of participants' mental model scores by 45%.

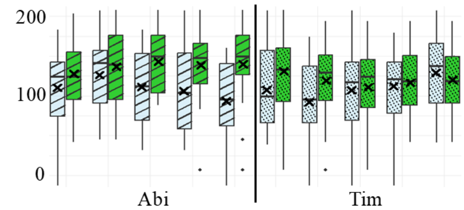



**Figure 18: From "Inclusive Design of AI's Explanations: Just for Those Previously Left Out?" by Hamid et al. [21], used under CC BY 4.0, Mental Model Concept Score split by prototype version and participants with an Abi-like or Tim-like problem-solving style. Post-GenderMag Abi (striped) and Tim (dotted) participants had higher (better) mental model concepts scores than their Original counterparts. The five problem-solving styles (from left to right): Info, Learn, Motiv, Risk, and SE.**

# 6 Results RQ3 & RQ4: Was GenderMag "Enough"?

## 6.1 RQ3: The Original GenderMag and AI

As the previous section shows, using GenderMag enabled the AI product teams to find AI inclusivity bugs, and fix them in ways that significantly improved at least one of the products (Team Game's). But only some of these effects came from teams' use of the GenderMag variant we term Original GenderMag, and Table 3 considers just these. As the table shows, the teams' use of Original GenderMag did enable them to be effective—they found 54 AI inclusivity bug using it. However, the table also shows an important omission in the teams' reasoning. Although the teams found 54 AI inclusivity bugs with the Original GenderMag, fixing 34 of them, their transcripts showed that *none* of them considered the possibility of the persona not believing the AI's outputs.

RQ3 asked whether Original GenderMag was "enough" for the teams' effectiveness, or whether something AI-specific is needed. The table suggests that something more AI-specific is needed, because the teams consistently overlooked situations in which the user was unconvinced by the AI. Given this omission, we invited teams to suggest whether and how to adapt the Original GenderMag to better fit AI products. Teams gave five ideas.

The teams had time to try only two of them, which we describe in the upcoming subsections; the rest of the ideas are enumerated in the supplemental documents. We refer to all five adaptations as GenderMag-for-AI variants.

## 6.2 RQ4: Seeds of Change—Team Game's Trees

After Session 5, we asked Team Game how they would change the Original GenderMag for AI products. They discussed what made AI products different, concluding that the walkthrough should consider when the AI is wrong (further supporting RQ3):

> Team Game: "*With a [traditional] UI, when something goes wrong, it's either the user's fault or the interface's fault, and the job of this walkthrough is essentially to parse that out... With AI systems, there is a third possibility, and that is that the AI is wrong. ... we need to think about a bit, and I don't know that we've thought that through...*"
> • Walkthrough Version: *Original GenderMag*

Team Game suggested changing the Original GenderMag's structure to consider the above:

> Team Game: "*...the [GenderMag] document is linear, [but] the goal structure is a tree...*"
> • Walkthrough Version: *Original GenderMag*

Figure 19 shows a conceptual view of Team Game's suggestions. This tree's two forks could provide teams the opportunity to explore different scenarios. The left fork would explore what happens when the persona believes the AI. The right fork explores the opposite, when the persona doubts the AI. We considered 3 possible places to introduce the forking structure—at the subgoal, the pre-action, and/or the post-action steps (recall Figure 4). The teams explored two of the possibilities, as we describe next.

| Session Number | Team | Which GenderMag? | Considered "Doubts the AI" | # Bugs Found | # Bugs Fixed |
|---|---|---|---|---|---|
| 1 | Weather | Original GenderMag | No | 3 | 3 |
| 2 | Farm | Original GenderMag | No | 13 | 8 |
| 3 | Farm | Original GenderMag | No | 8 | 5 |
| 4 | Game | Original GenderMag | No | 5 | 3 |
| 5 | Game | Original GenderMag | No | 25 | 15 |
| | | | Totals: | 54 | 34 |

**Table 3: Teams' find sessions applying the Original GenderMag. All teams successfully found and fixed AI inclusivity bug instances, but *none* of the teams considered whether the persona may doubt the AI.**

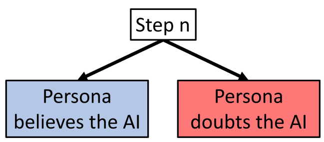



**Figure 19: Team Game's suggestion of adding a tree structure.**

## 6.3 RQ4: The Post-Action Fork GenderMag

We began by implementing Team Game's suggestions at the last step of the Original GenderMag, creating the Post-Action Fork GenderMag with the workflow shown in Figure 20. Like the other GenderMag walkthrough questions, we framed the new Post-Action Fork GenderMag questions in terms of the *persona's* beliefs—here, of whether the AI is right or wrong—not the *developer's* belief.

Team Farm tested the Post-Action Fork GenderMag, and they found it confusing. Their primary feedback to this adaptation was that they thought that this was not the right question to ask at this stage, instead, suggesting adding a question between the pre-action and post-action questions:

> Team Farm: "*...I feel like that's the wrong question to ask. Would it be more pertinent to know if Abi understands what they're seeing first and then whether or not they believe it's correct?*"
> • Walkthrough Version: *Post-Action Fork GenderMag*

The team reverted back to Original GenderMag. Although they still did not consider whether users may doubt the AI, Team Farm found an additional 12 more AI inclusivity bugs and completing their goal. In discussing possible GenderMag walkthrough changes for AI products, the topic of potential cognitive taxes came up [23]. Team Farm acknowledged this potential tax, reconsidering their initial suggestion in light of these taxes:

> Team Farm: "*...make it less taxing on the people evaluating...I think it is okay to skip the step of 'does Abi understand that information'...I don't really like the question. I get the point. I just don't know how to change it...*"
> • Walkthrough Version: *Post-Action Fork GenderMag*

## 6.4 RQ4: The Pre-Action Fork GenderMag

Given that the fork in the post-action did not work for Team Farm, we went back to Team Game's initial suggestions by forking at the pre-action step. This resulted in the Pre-Action Fork GenderMag with the workflow shown in Figure 21.

A difference in the Pre-Action Fork GenderMag from the other variants was in how many actions the teams could evaluate (Figure 22). With the Original GenderMag (left), teams selected only a single action to evaluate in the pre-action step. However, the Pre-Action Fork GenderMag (right) prodded the teams to set *two* actions: one when Abi believed the AI's previous output (Pre-Action) and one when Abi did not (Pre-Action). Once the teams evaluated both actions, they chose which one to evaluate in the post-action step.

Both Team Weather and Team Game tested the Pre-Action Fork GenderMag, and noticed that the Believes the AI fork performed similarly to the Original GenderMag pre-action step. Because of this, teams' ability to find AI inclusivity bugs as they had with Original GenderMag did not change when the persona did not doubt the AI's output. But now that they also considered situations when the persona did doubt the AI's output, teams also found new, different bugs compared to the Believes the AI fork.

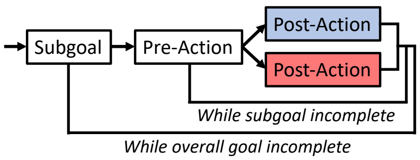
For example, Team Game wanted Abi to take the same action as the Believes the AI fork, revisiting the X-player AI's previous move, but they found *different* AI inclusivity bug instances than those in the Believes the AI fork (Figure 23). By considering both forks, teams found AI inclusivity bugs that they could not have uncovered with the Original GenderMag.

> Team Game: "*Something has gone wrong, and Abi will want to validate why something went wrong before losing the context...*"
> • Bug Type(s) Found: *Interpret AI?*
> • Problem-Solving Style Value(s): *Comprehensive Info. Proc.*
> • Walkthrough Version: *Pre-Action Fork Version*

Team Game chose to always keep Abi's action the same when evaluating both forks, but they gave different answers for whether Abi would take the action or not in each fork. For instance, when evaluating the Believes the AI, Team Game decided that Abi would *not* take the intended action, because Abi's comprehensive information processing style was satiated, so they would not know why they should:

> Team Game: "*The reason I say [no] is, if Abi is already happy with the thing, there's no real reason to be clicking on this button... they have no reason to want this information if they're happy.*"
> • Bug Type(s) Found: *AI: why should I?*
> • Problem-Solving Style Value(s): *Comprehensive Info. Proc.*
> • Walkthrough Version: *Pre-Action Fork Version*

However, while evaluating the other fork (Doubts the AI), Team Game also gave a different answer. They said that since Abi did *not*



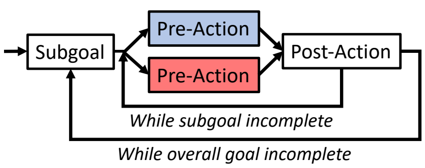

**Subgoal:** *Will [persona] have formed this sub-goal as a step to their overall goal?*

**Pre-Action:** Will [persona] take this action at this step? Yes/no/maybe, and why?

**Post-Action:** Believes the AI—*If [persona] believes that the previous action's output is correct, will [persona] know that they did the right thing and are making progress toward their goal? Yes/no/maybe, and why?*

**Post-Action:** Doubts the AI—*If [persona] does not believe that the previous action's output is correct, will [persona] know that they did the right thing and are making progress toward their goal? Yes/no/maybe, and why?*

**Figure 20: (Left): The Post-Action Fork GenderMag graph, forking the post-action step for Believes the AI and Doubts the AI. (Right): Question wording for each node in the graph. Gray text: Unchanged from the Original.**




**Subgoal:** *Will [persona] have formed this sub-goal as a step to their overall goal?*

**Pre-Action:** Believes the AI—*If [persona] believes that the previous action's output is correct, will [persona] take this action at this step? Yes/no/maybe, and why?*

**Pre-Action:** Doubts the AI—*If [persona] does not believe that the previous action's output is correct, will [persona] take this action at this step? Yes/no/maybe, and why?*

**Post-Action:** (Picking one pre-action) — *Will [persona] know that they did the right thing and are making progress toward their goal? Why?*

**Figure 21: Pre-Action Fork GenderMag. (Left): Graphical depiction, forking the pre-action step into 2 circumstances: when the persona believes the AI (Believes the AI), and when they do not (Doubts the AI). (Right): Question wording for each node. Gray text: Unchanged from the Original.**

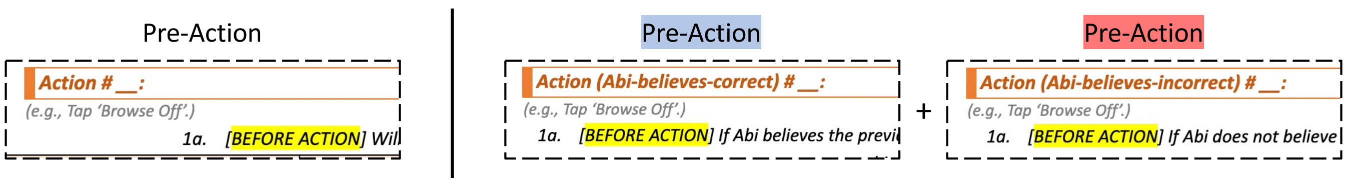



**Figure 22: The pre-action fork form. (Left): For Original GenderMag, teams choose *one* action to evaluate. (Right): In the Pre-Action Fork GenderMag, teams choose *two* actions to evaluate: one when Abi believes the AI's output (Believes the AI) and another when Abi does not (Doubts the AI).**

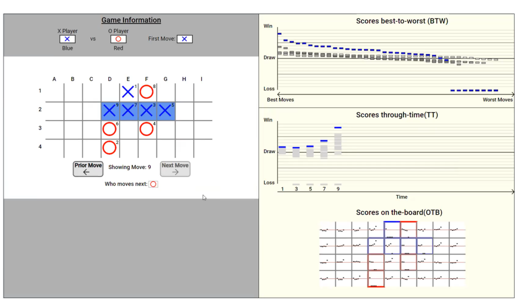



**Figure 23: Interface state when Team Game was evaluating in the Doubts the AI fork of the Pre-Action Fork GenderMag. The game ended with X-player AI winning and the intended action was for Abi to go back to previous moves.**

believe the AI's previous output, they would now take the action, providing a different problem-solving style value to explain it:

> Team Game: "*...I'm a yes here... when things go south, Abi is going to be task-oriented and create a task for themselves ...to figure out what went differently from expectation.*"
> - Bug Type(s) Found: *None*
> - Problem-Solving Style Value(s): *Task-oriented motivations*
> - Walkthrough Version: *Pre-Action Fork Version*

For Team Weather, Pre-Action Fork GenderMag revealed a different kind of bug. It revealed that Team Weather's prototype designers had overlooked the possibility that Abi might doubt the AI's output, Abi's doubts brought about an abrupt dead-end—there was no action in the interface for Abi to take in such a scenario:

> Team Weather: "*If Abi is not satisfied with the output, I don't know what I would expect from Abi...That's not– I haven't looked at it that way before. It's not the point of view that we would, or that I would typically try to*

*put myself into. The assumption is that you believe me.*"
• Walkthrough Version: *Pre-Action Fork GenderMag*

Because the teams started trying Pre-Action Fork GenderMag well into the study period, it did not produce as much data as did their uses of Original GenderMag. Even so, the evidence so far of Pre-Action Fork GenderMag's greater effectiveness over Original GenderMag for AI products is very consistent. First, as Table 4 shows, the *only* times any team considered the persona doubting the AI was when they used the Pre-Action Fork GenderMag. Second, as Table 5 shows, 9 sources of evidence over time, over different teams, and over different data sources (bugs found vs. verbalizations) cross-corroborate the teams' effectiveness when using Pre-Action Fork GenderMag.

## 7 Discussion

Several theoretical lenses provide perspectives on the 6 AI inclusivity bug types the teams found. In this section, we consider how these theoretical lenses may be of practical use to AI product designers.

First and most obvious, the bug instances tie directly to the foundational theories behind GenderMag's problem-solving style types [9]. Recall that, whenever a team noticed an AI inclusivity bug, they tied that bug to any problem-solving styles that they expected the bug to particularly impact. For example, they often tied instances of *Interpret AI?* ("What does this even mean?") to risk-averse users and to those who are comprehensive information processors. These ties provide the "why"s behind the AI inclusivity bugs—why users with those problem-solving attributes might experience that particular bug. For these teams, these why's often pointed the way toward fixes. Continuing our example, when they tied an instance of *Interpret AI?* to comprehensive information processing, they tended to fix the bug by making more information available. Thus, these ties between an AI inclusivity bug and the problem-solving styles behind it sometimes pointed the way toward how to fix the bug.

Another theory that relates to these AI inclusivity bug types is Blackwell's model of attention investment [7], already alluded to in Section 5. Strictly speaking, this model's units are units of attention (similar to time): attention costs or investments (similar to "time I must spend now") vs. attention benefits ("time I save later"), modified by the risk ("probability that if I spend the time I still won't receive the benefits"). However, for this discussion, we relax this constraint to allow *any* kind of cost/benefit.

This relaxed model of attention investment enables unifying the 6 AI inclusivity bugs as obstructions to receiving the benefits. For this discussion, we use Team Weather as a running example. Since Team Weather's intended users are agricultural growers, the hoped-for benefit they gain from using these AI products is to produce higher quality/quantity of crops. Here, *AI: more info!* could stop some growers from having the information they need to decide whether to spend hard-earned dollars on frost mitigation today to salvage some percentage of their crops, versus avoiding the expenditure while leaving crop survival rate more to chance. The other five AI inclusivity bugs—*Interpret AI?*, *AI input↔output?*, *AI: why should I?*, *AI: actionable?*, and *AI changes?*—can similarly be seen as barriers to some growers receiving the hoped-for benefit of more financially viable farms.

Under this reasoning, the attention investment lens could facilitate an AI product designer using a cost-benefit perspective when ideating potential fixes. Specifically, it suggests an AI-inclusivity debugging approach driven by: "how can I fix this in a way that brings to diverse users greater benefit or lower cost and/or a lower probability of failing to receive this benefit?"

A third theory lens is Norman's Gulf of Execution and Gulf of Evaluation [45]. A Gulf of Execution describes barriers to *doing* something, and a Gulf of Evaluation describes not knowing whether what a user just *did* made an impact (and if so, what). This theory brings an actionability perspective to the 6 AI inclusivity bugs.

Using this perspective, the *AI: more info!* example above could create an execution gulf, because lack of information could stop some growers from knowing a suitable way to take action on their farm. *AI: actionable?* directly describes a barrier to taking suitable action. *AI: why should I?* suggests a barrier between the goal of a more productive farm and engaging in *any* way with the AI. *AI input↔output?* and *AI changes?* raise gulfs of evaluation, and *Interpret AI?* could produce gulfs of either execution or evaluation.

Using the Norman gulfs could enable an AI product designer to think about fixes from an actionability perspective. For example, it suggests an inclusivity debugging approach driven by: "how can I fix this in a way that enables diverse users to take the most appropriate action for their particular farm?"

It remains an open question whether and which of the unifying directions these theories encourage will be useful abstractions beyond the 6 AI inclusivity bug types this paper identified. We look forward to future researchers' discoveries of new AI inclusivity bugs types that will enable exploration of this question.

## 8 Limitations

Every empirical study has limitations [33, 67], and ours is no exception. One limitation of our results is that all three teams evaluated their AI products from the perspective of only the Abi persona. Ideally, the teams would have also evaluated from the perspective of the Tim persona, so as to work on addressing inclusivity bugs at both endpoints of each problem-solving style. The teams elected to use only Abi to save time, but this shortcut missed an opportunity to find AI inclusivity bugs across the full range of problem-solving style values.

A limitation in experimenting with different variants of GenderMag-for-AI was the availability of teams' time to try out all the suggestions to change the Original GenderMag. Between sessions, we only had time to incorporate some of the teams' suggestions. As previously stated, time was available to try out only two versions. Thus, although the Pre-Action Fork GenderMag demonstrated the most promise for two teams, there may be a variant that fits AI products better. Still, the suggested-but-untried variants are shown in the supplemental documents, and we invite interested researchers to try out the others.

Further, the list of AI inclusivity bugs we identified in our data is unlikely to be a complete list. For example, in eXplainable AI (XAI) alone, there are many explanation types that our data do not cover (e.g., counterfactual, feature-importance, saliency maps). Also, the bug types in these systems may not apply to all AI domains, and the fixes' effectiveness was evaluated for only one of the three AI

| Session Number | Team | Which GenderMag? | Considered Doubts the AI | # Bugs Found | # Bugs Fixed |
|---|---|---|---|---|---|
| 1–5 | All 3 | Original GenderMag | No | 54 | 34 |
| 6 | Farm | Post-Action Fork GenderMag | No | 12 | 6 |
| 7 | Weather | Pre-Action Fork GenderMag | Yes | 9 | 5 |
| 8 | Game | Pre-Action Fork GenderMag | Yes | 8 | 2 |
| | | | *Totals*: | 83 | 47 |

**Table 4: Teams Weather and Game successfully found and fixed AI inclusivity bug instances with the Pre-Action Fork GenderMag. With this version, teams considered when the persona doubted the AI.**

| | Bug in Believes the AI | Bug in Doubts the AI | Unique bug in fork | Verbal support | Total |
|---|---|---|---|---|---|
| Team Game | ✓✓ | ✓✓ | ✓✓ | ✓ | 7 |
| Team Weather | ✓ | | | ✓ | 2 |
| | 3 | 2 | 2 | 2 | 9 |

**Table 5: Triangulation Table—9 sources of evidence of teams' reactions to the Pre-Action Fork GenderMag.**

product teams. Because this was a field study of particular teams in particular contexts with particular AI products, we expect more AI inclusivity bug types and fixes to emerge as other researchers begin to find them in other types of AI products and explanations.

Although having only 3 teams may also seem to limit generality, Baskerville and Lee distinguish *inductive generalizing* from *deductive generalizing* [4]. Lab studies use inductive generalizing, where larger n is desirable. They take as input empirical observations and produce some kind of general principle (e.g.,: "more experienced programmers produce better code"). In contrast, qualitative field studies like ours use deductive generalizing, taking as input a general principle and empirical observations in a new real-world situation, and producing evidence for and/or against the principle's applicability to that situation [4]. Our study produced evidence of a boundary on Original GenderMag's effectivness: although the AI practitioners' use of Original GenderMag was effective at finding AI inclusivity bugs, it also had a blind spot, with the AI practitioners overlooking possibilities of a user doubting the AI's output. The study also produced some evidence that GenderMag-for-AI can address that issue, but more investigations in a variety of contexts are needed to further explore boundaries of GenderMag-for-AI's effectiveness.

## 9 Conclusion

This paper set out to investigate AI inclusivity bugs, and how to find and fix them. To investigate this question, we conducted a field study, in which three AI product teams used several variants of the GenderMag inclusive design method to evaluate their own products. This paper's first contribution is defining the concept of AI inclusivity bugs.

AI inclusivity bugs are a new concept. They exist *only* in an AI's communications with users and *disproportionately impact* some group(s) of AI product users. Given bugs like this, our second contribution is the answer to RQ1, which asks what AI inclusivity bugs these AI product teams were able to find in their own products, what these bugs looked like, and how commonly they arose. The result was 6 AI inclusivity bug type categories, which arose 83 times:

- *Interpret AI?* (27 instances): "What does this even mean?" Particularly tied with risk-aversion and/or comprehensive information processing.
- *AI input↔output?* (9 instances): "What does this (input) have to do with that (output)?" Particularly tied with risk-aversion.
- *AI: why should I?* (19 instances): "Why even look at this?" Particularly tied with risk aversion.
- *AI: more info!* (9 instances): "Need more info!" Particularly tied with comprehensive information processing.
- *AI: actionable?* (12 instances): "So? What should I DO?" Particularly tied with comprehensive information processing.
- *AI changes?* (7 instances): "What's changed?" Particularly tied with self-efficacy relatively low compared to peers.

Our third contribution was the answer to RQ2: how teams' fixed these AI inclusivity bug types. None of the teams had noticed any of the AI inclusivity bug types before this field study, which suggests that finding such bugs was a challenge that needed to be overcome, such as by using the methods in this paper. However, once found, it turned out that these AI inclusivity bug instances were often straightforward for the teams to fix using techniques common in HCI practice. For example, a common fix to *Interpret AI?* was adding clarifications to features that triggered it (e.g., Figure 11). A common fix to *AI input↔output?* was adding explicit ties between the AI's inputs and its outputs, as in Figure 12 For *AI: why should I?*, teams often instantiated the Surprise-Explain-Reward strategy, to subtly use surprising results to entice users toward an actionable explanation (Figures 13 and 14). *AI: more info!*'s fix was especially obvious: provide more information. Finally, the teams' fixes to both

*AI: actionable?* and *AI changes?* tended to add explicitness: making possible actions explicit for the former (Figure 16), and being explicit about what changed for the latter (Figure 17).

Our fourth contribution lies in what the AI product teams' work revealed for RQ3, which asks if a current inclusive design method (GenderMag) is "enough" for evaluating AI-powered systems. The teams' work showed that, although the original method did allow them to be effective at finding AI inclusivity bugs, it had a blind spot—it did not prod them to consider users being doubtful of the AI's recommendations/decisions.

Our final contribution is the answer to RQ4, which asked how the method should change to better accommodate AI-powered systems. When we worked with the teams to explore this question, several GenderMag-for-AI variants emerged. The teams were unanimous on the importance of these variants to consider both when users believed the AI and when they did not. Of the methods the teams tried, Pre-Action Fork GenderMag was the most successful, uncovering AI inclusivity bugs in situations that the original method overlooked. As Team Weather put it, before using Pre-Action Fork GenderMag...

> Team Weather: "*...if the user doesn't trust it?...it really isn't something that had ever occurred to me at all.*"

We hope other researchers will join us in investigating the effectiveness of GenderMag-for-AI variants, and more generally of any approach that can improve AI products' inclusiveness for all users.

## Acknowledgments

We thank the reviewers and meta-reviewer for their supportive engagement with our paper's quality and clarity. We are also very grateful to the AI product teams for their participation and hard work. This work has been supported in part by the USDA-NIFA and NSF #2021-67021-35344.

## Generative AI Usage Disclosure

No generative AI tools were used in this research or writing process.

## References


[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[2] Andrew Anderson, Jimena Noa Guevara, Fatima Moussaoui, Tianyi Li, Mihaela Vorvoreanu, and Margaret Burnett. 2024. Measuring User Experience Inclusivity in Human-AI Interaction via Five User Problem-Solving Styles. *ACM Trans. Interact. Intell. Syst.* 14, 3, Article 21 (Sept. 2024), 90 pages. https://doi.org/10.1145/3663740

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

[4] Richard Baskerville and Allen S. Lee. 1999. Distinctions among Different Types of Generalizing in Information Systems Research. In *New Information Technologies in Organizational Processes: Field Studies and Theoretical Reflections on the Future of Work*, Ojelanki Ngwenyama, Lucas D. Introna, Michael D. Myers, and Janice I. DeGross (Eds.). Springer US, Boston, MA, 49–65. https://doi.org/10.1007/978-0-387-35566-5_5

[5] 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.

[6] Alba Bisante, Venkata Srikanth Varma Datla, Emanuele Panizzi, Gabriella Trasciatti, and Stefano Zeppieri. 2024. Enhancing Interface Design with AI: An Exploratory Study on a ChatGPT-4-Based Tool for Cognitive Walkthrough Inspired Evaluations. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces (AVI '24)*. Association for Computing Machinery, New York, NY, USA, Article 41, 5 pages. https://doi.org/10.1145/3656650.3656676

[7] A.F. Blackwell. 2002. First steps in programming: a rationale for attention investment models. In *Proceedings IEEE 2002 Symposia on Human Centric Computing Languages and Environments*. IEEE, Washington DC, USA, 2–10. https://doi.org/10.1109/HCC.2002.1046334

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, Stockholm, Sweden, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[9] 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.

[10] Nicole Chi, Emma Lurie, and Deirdre K. Mulligan. 2021. Reconfiguring Diversity and Inclusion for AI Ethics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 447–457. https://doi.org/10.1145/3461702.3462622

[11] Seung Jin Chung. 2024. Revolutionizing Persona Design with Generative AI: Insights from Experts. *The International Journal of Design Management and Professional Practice* 18, 2 (2024), 109.

[12] 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*, Cyril Goutte and Xiaodan Zhu (Eds.). Springer International Publishing, Cham, 138–150.

[13] Vagner Figueredo de Santana, Larissa Monteiro Da Fonseca Galeno, Emilio Vital Brazil, Aliza Heching, and Renato Cerqueira. 2023. Retrospective End-User Walkthrough: A Method for Assessing How People Combine Multiple AI Models in Decision-Making Systems. arXiv:cs.HC/2305.07530 https://arxiv.org/abs/2305.07530

[14] 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.

[15] Jonathan Dodge, Andrew A. Anderson, Matthew Olson, Rupika Dikkala, and Margaret Burnett. 2022. How Do People Rank Multiple Mutant Agents?. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 191–211. https://doi.org/10.1145/3490099.3511115

[16] GenderMag. 2021. *GenderMag design catalog*. The GenderMag Project. http://gendermag.org/dc/

[17] GenderMag. 2025. *The GenderMag Project: The Personas*. The GenderMag Project. https://gendermag.org/custom_persona.php

[18] 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 (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 90–99. https://doi.org/10.1145/3287560.3287563

[19] Valentina Grigoreanu, Jill Cao, Todd Kulesza, Christopher Bogart, Kyle Rector, Margaret Burnett, and Susan Wiedenbeck. 2008. Can feature design reduce the gender gap in end-user software development environments?. In *2008 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, New York, NY, USA, 149–156. https://doi.org/10.1109/VLHCC.2008.4639077

[20] 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 *Proceedings of the 2022 ACM/IEEE 44th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS '22)*. Association for Computing Machinery, New York, NY, USA, 90–101. https://doi.org/10.1145/3510458.3513009

[21] Md Montaser Hamid, Fatima A. Moussaoui, Jimena Noa Guevara, Andrew Anderson, Puja Agarwal, Jonathan Dodge, and Margaret Burnett. 2026. Inclusive Design of AI's Explanations: Just for Those Previously Left Out? *ACM Trans. Interact. Intell. Syst.* 16, 1, Article 8 (Jan. 2026), 44 pages. https://doi.org/10.1145/3772074

[22] 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 (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 392–402. https://doi.org/10.1145/3351095.3372831

[23] Claudia Hilderbrand, Christopher Perdriau, Lara Letaw, Jillian Emard, Zoe Steine-Hanson, Margaret Burnett, and Anita Sarma. 2020. Engineering gender-inclusivity into software: ten teams' tales from the trenches. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE '20)*. Association for Computing Machinery, New York, NY, USA, 433–444. https://doi.org/10.1145/3377811.3380371

[24] Andreas Holzinger, Michaela Kargl, Bettina Kipperer, Peter Regitnig, Markus Plass, and Heimo Müller. 2022. Personas for artificial intelligence (AI) an open source toolbox. *IEEE Access* 10 (2022), 23732–23747.

[25] Hsiu-Fang Hsieh and Sarah Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research* 15, 9 (2005), 1277–1288.

[26] Apple Inc. 2019. Machine learning. https://developer.apple.com/design/human-interface-guidelines/technologies/machine-learning/introduction/
[27] Design Foundation Interaction. 2024. What is inclusive design? - updated 2024. https://www.interaction-design.org/literature/topics/inclusive-design#:~:text=Inclusive%20design%20is%20an%20approach,diverse%20needs%2C%20backgrounds%20and%20experiences.
[28] Paul Jaccard. 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* 44 (1908), 223–270.
[29] 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.
[30] Ofem Usani Joseph, Iyam Mary Arikpo, Ovat Sylvia Victor, Nworgwugwu Esther Chidirim, Anake Paulina Mbua, Udeh Maryrose 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 & Teaching* 7, 1 (2024), 76–92. https://search.informit.org/doi/10.3316/informit.T2025091900005900080440812
[31] 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. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (Jun. 2023), 14277–14285. https://doi.org/10.1609/aaai.v37i12.26670
[32] 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.
[33] 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.
[34] 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 SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2207676.2207678
[35] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, New York, NY, USA, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235
[36] 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
[37] 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 (Oct. 2020), 83–91. https://doi.org/10.1609/hcomp.v8i1.7466
[38] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75.
[39] John Long. 2021. *Approaches and Frameworks for HCI Research.* Cambridge University Press, Trinity Lane, Cambridge, UK.
[40] Thomas Mahatody, Mouldi Sagar, and Christophe Kolski. 2010. State of the art on the cognitive walkthrough method, its variants and evolutions. *Intl. Journal of Human–Computer Interaction* 26, 8 (2010), 741–785.
[41] Mark C Miller. 2021. Inclusivity bugs and the language we use. *Computing in Science & Engineering* 23, 6 (2021), 69–71.
[42] 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: An Multi-Year Case Study of Suggest Edit, a Code Review Feature. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE '24)*. Association for Computing Machinery, New York, NY, USA, Article 189, 12 pages. https://doi.org/10.1145/3597503.3639097
[43] IB Myers. 1962. The Myers-Briggs Type Indicator. , 109 pages. https://doi.org/10.1037/14404-000
[44] Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an LLM to Help With Code Understanding. arXiv:cs.SE/2307.08177 https://arxiv.org/abs/2307.08177
[45] Donald A Norman. 1986. Cognitive engineering. In *User centered system design*. CRC Press, Boca Raton, FL, USA, 31–62.
[46] Rasha Mohammad Nouraldeen. 2023. The impact of technology readiness and use perceptions on students' adoption of artificial intelligence: the moderating role of gender. *Development and Learning in Organizations: An International Journal* 37, 3 (2023), 7–10.
[47] Kingsley Ofosu-Ampong. 2023. Gender differences in perception of artificial intelligence-based tools. *Journal of Digital Art & Humanities* 4, 2 (2023), 52–56.
[48] Dani Or, Markus Tuller, and Jon M Wraith. 2005. Soil water potential. *Encyclopedia of soils in the environment* 3 (2005), 270–277.
[49] Hema Susmita Padala, Christopher Mendez, Felipe Fronchetti, Igor Steinmacher, Zoe Steine-Hanson, Claudia Hilderbrand, Amber Horvath, Charles Hill, Logan Simpson, Margaret Burnett, Marco Gerosa, and Anita Sarma. 2022. How Gender-Biased Tools Shape Newcomer Experiences in OSS Projects. *IEEE Transactions on Software Engineering* 48, 1 (2022), 241–259. https://doi.org/10.1109/TSE.2020.2984173
[50] Joshua Giuseppe Puglisi. 2023. *Developing Personas for Explainability Requirements by Means of a User Study.* Bachelor's Thesis. Gottfried Wilhelm Leibniz Universität Hannover.
[51] Jeba Rezwana and Mary Lou Maher. 2023. Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–28.
[52] Aseem Saxena, Paola Pesantez-Cabrera, Rohan Ballapragada, Kin-Ho Lam, Markus Keller, and Alan Fern. 2024. Grape Cold Hardiness Prediction via Multi-Task Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 13 (Jul. 2024), 15717–15723. https://doi.org/10.1609/aaai.v37i13.26865
[53] 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*, Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl (Eds.). Springer International Publishing, Cham, 431–449.
[54] 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.
[55] 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 (GenderIT '18)*. Association for Computing Machinery, New York, NY, USA, 191–198. https://doi.org/10.1145/3196839.3196869
[56] Joongi Shin, Michael A. Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding Human-AI Workflows for Generating Personas. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 757–781. https://doi.org/10.1145/3643834.3660729
[57] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.
[58] Marita Skjuve, Petter Bae Brandtzaeg, and Asbjørn Følstad. 2024. Why do people use ChatGPT? Exploring user motivations for generative conversational AI. *First Monday* 29, 1 (Jan. 2024), 23. https://doi.org/10.5210/fm.v29i1.13541
[59] 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.
[60] S Shyam Sundar. 2020. Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication* 25, 1 (2020), 74–88.
[61] Hao Tan, Shenglan Peng, Jia-Xin Liu, Chun-Peng Zhu, and Fan Zhou. 2022. Generating personas for products on social media: a mixed method to analyze online users. *International Journal of Human–Computer Interaction* 38, 13 (2022), 1255–1266.
[62] 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. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 245, 13 pages. https://doi.org/10.1145/3411764.3445365
[63] 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 (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300283
[64] Chathurika S. Wickramasinghe, Daniel L. Marino, Javier Grandio, and Milos Manic. 2020. Trustworthy AI Development Guidelines for Human System Interaction. In *2020 13th International Conference on Human System Interaction (HSI)*. IEEE, New York, NY, USA, 130–136. https://doi.org/10.1109/HSI49210.2020.9142644
[65] Larry Williams. 2019. Deficit irrigation of wine grape vineyards. , 11 pages. https://ucanr.edu/repository/fileaccess.cfm?article=177652&p=CYZFUA
[66] Aaron Wilson, Margaret Burnett, Laura Beckwith, Orion Granatir, Ledah Casburn, Curtis Cook, Mike Durham, and Gregg Rothermel. 2003. Harnessing curiosity to increase correctness in end-user programming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, New York, NY, USA, 305–312. https://doi.org/10.1145/642611.642665
[67] 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.
[68] Wei Xu, Zaifeng Gao, and Marvin Dainoff. 2025. An HCAI Methodological Framework (HCAI-MF): Putting It Into Action to Enable Human-Centered AI. arXiv:cs.HC/2311.16027 https://arxiv.org/abs/2311.16027

[69] 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 (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 547–558. https://doi.org/10.1145/3351095.3375709

[70] 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 (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[71] Seyma N Yildirim-Erbasli, Okan Bulut, Carrie Demmans Epp, and Ying Cui. 2023. Conversation-Based Assessments in Education: Design, Implementation, and Cognitive Walkthroughs for Usability Testing. *Journal of Educational Technology Systems* 52, 1 (2023), 27–51.

[72] Ke Zhang. 2021. How information processing style shapes people's algorithm adoption. *Social Behavior and Personality: an international journal* 49, 8 (2021), 1–13.