“I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust

Sunnie S. Y. Kim*
sunniesuhyoung@princeton.edu
Princeton University
USA

Q. Vera Liao
veraliao@microsoft.com
Microsoft
Canada

Mihaela Vorvoreanu
mihaela.vorvoreanu@microsoft.com
Microsoft
USA

Stephanie Ballard
stephballard@microsoft.com
Microsoft
USA

Jennifer Wortman Vaughan
jenn@microsoft.com
Microsoft
USA

ABSTRACT

Widely deployed large language models (LLMs) can produce convincing yet incorrect outputs, potentially misleading users who may rely on them as if they were correct. To reduce such overreliance, there have been calls for LLMs to communicate their uncertainty to end users. However, there has been little empirical work examining how users perceive and act upon LLMs’ expressions of uncertainty. We explore this question through a large-scale, pre-registered, human-subject experiment (N = 404) in which participants answer medical questions with or without access to responses from a fictional LLM-infused search engine. Using both behavioral and self-reported measures, we examine how different natural language expressions of uncertainty impact participants’ reliance, trust, and overall task performance. We find that first-person expressions (e.g., “I’m not sure, but...”) decrease participants’ confidence in the system and tendency to agree with the system’s answers, while increasing participants’ accuracy. An exploratory analysis suggests that this increase can be attributed to reduced (but not fully eliminated) overreliance on incorrect answers. While we observe similar effects for uncertainty expressed from a general perspective (e.g., “It’s not clear, but...”), these effects are weaker and not statistically significant. Our findings suggest that using natural language expressions of uncertainty may be an effective approach for reducing overreliance on LLMs, but that the precise language used matters. This highlights the importance of user testing before deploying LLMs at scale.

CCS CONCEPTS

• Human-centered computing → Empirical studies in HCI
• Computing methodologies → Artificial intelligence

KEYWORDS

Large language models, Uncertainty expression, Trust in AI, Overreliance, Human-AI interaction

1 INTRODUCTION

Large language models (LLMs) are transforming our daily lives. Today millions of people already incorporate LLMs into everyday tasks like searching for information [56, 66], writing [38, 105], and programming [3, 70, 79]. However, the use of LLMs raises significant risks [11, 14, 99]. Notably, like all models, LLMs are imperfect. They are widely recognized to produce outputs that are fluent and plausible, yet ultimately wrong [41, 42, 52]. This can lead to disastrous outcomes through overreliance [19, 25, 76, 92], when people take actions based on incorrect outputs. This concern garnered much public attention in 2023 when a lawyer included fake judicial opinions generated by ChatGPT in a legal brief presented in court [100]. Such risks have been at the forefront of regulators’ minds when drafting new frameworks for governing AI including the Draft AI Act in the European Union [75] and the NIST AI Risk Management Framework in the United States [87]. In fact, Article 14 of the Draft EU AI Act explicitly requires developing and evaluating approaches to prevent overreliance on AI systems. However, overreliance is notoriously difficult to mitigate, as many mitigations, such as explanations, are found to be ineffective or even can backfire to increase overreliance [7, 76, 77, 97, 108].

As one approach to reduce overreliance, the research community has called for LLMs and LLM-infused applications to express the uncertainty of outputs to end users [6, 50, 51, 67, 91, 110]. The idea of conveying AI uncertainty is not new; in AI-assisted decision-making settings, communicating (un)certainty has been shown to support trust calibration [108], increase vigilance [78], and improve task performance [7]. But because of their open-ended outputs, wide-ranging use cases and user bases, and shifting public perception, LLMs raise new questions about how to
both estimate and express uncertainty [50, 91]. For estimation, “default” approaches are often found to be overconfident [26, 67, 103], and a new line of work has emerged on improving their calibration [4, 26, 31, 47, 51, 53, 67, 88, 110]. For expression, LLMs open up a new design space; instead of presenting uncertainty numerically or visually, LLMs can present natural language expressions of uncertainty — for instance, hedging phrases like “I’m not sure, but...” — embedded in their outputs. Still, there is little understanding about how to effectively express uncertainty in natural language to end users.

To deploy LLMs responsibly, it is necessary to understand how users react to uncertainty expression before implementing approaches at scale since it may have unintended negative consequences — potentially even increasing overreliance if it causes the system to appear more trustworthy than it is. Best practices for uncertainty expression will play a critical role in ensuring that requirements like those in the Draft EU AI Act serve their intended purpose. To that end, we study how people perceive and act upon an LLM’s expression of uncertainty when seeking medical information using a fictional LLM-infused search engine. We choose to study this setting because search (unlike, for example, creative writing) is an application in which the correctness of responses is fundamental — especially for potentially high-stakes medical queries — making overreliance a serious concern. Additionally, LLM-infused search engines are already used by millions of people.

We choose to focus on natural language expressions for several reasons. First, LLM-infused search engines already include hedging language [50, 56, 67]. Second, social science research shows that, in human communication, expressing (un)certainty through natural language is often preferred and perceived as more intuitive than numerical expressions [33, 54, 96, 102, 111]. Third, this allows uncertainty to be expressed seamlessly within the natural language interactions of LLM-infused applications, rather than on the side or in onboarding materials [21, 68, 73] that users might overlook.

Taking inspiration from the uncertainty communication literature — both in the context of AI systems and person-to-person [72, 106] — we also explore the impact of the perspective used to express the uncertainty, comparing expressions in the first person (e.g., “I’m not sure, but...”) with expressions from a general perspective (e.g., “It’s not clear, but...”).

Concretely, we conduct a large-scale, pre-registered, human-subject experiment (N=404) in which participants answer medical questions with or without access to responses from a fictional LLM-infused search engine, referred to as “AI System A.” We randomly vary whether participants have access to the system’s responses as well as the presence (present/not present) and perspective (first-person/general) of uncertainty expressed in these responses. We measure the impact of these experimental conditions on factors including participants’ accuracy, the amount of time they take, their reliance on the system’s responses versus other sources of information, and their self-reported trust in the system.

We find that participants who are shown first-person expressions of uncertainty are less confident in the system’s answers, agree with the system’s answers less often, and submit more correct answers compared with participants who see no expression of uncertainty.

An exploratory analysis suggests that the increased accuracy can be attributed to reduced (but not fully eliminated) overreliance on the system’s incorrect answers. While we observe similar effects for uncertainty expressed from a general perspective, these effects are weaker and not statistically significant. These results suggest that expressing uncertainty through natural language can be an effective way to reduce overreliance and (over)trust in LLM-infused search engines. Still, we advocate for teams building and deploying LLMs to evaluate approaches to mitigate overreliance, including language choices, carefully with end users before release and for policymakers to embrace diverse and flexible approaches.

2 RELATED WORK
2.1 Uncertainty Expression

Uncertainty expression has been studied extensively both in the context of AI and in the context of human communication. Estimates of uncertainty can be expressed in different ways, including numerically (e.g., “a probability of 0.2”), visually (e.g., displaying error bars), and through natural language (e.g., “with high uncertainty...”). While numerical expressions and visualizations allow for high precision, they are notoriously difficult for people to understand and are often misinterpreted, even by experts [9, 40, 43, 86]. In contrast, while less precise, natural language expressions of uncertainty are often perceived to be more intuitive and favored by people [33, 54, 96, 102, 111]. Because of this, and since LLMs already produce natural language outputs, we focus on natural language expressions in our study.

Different forms of natural language uncertainty expression have been studied by researchers in disciplines ranging from psychology [30, 95, 101] and human-computer interaction [3, 37, 89] to communication [59] and marketing [36, 72]. Our study design builds on this literature. Most notably, in the context of marketing, Obi and Berger [72] found that different types of hedges — a form of uncertainty expression — have different levels of persuasion, with the most persuasive being those that suggest a high likelihood of occurrence (e.g., “probably” as opposed to “possibly”) and those that take a personal, first-person perspective (e.g., “I feel like...”) as opposed to a general perspective (e.g., “It feels like...”). Indeed, taking a first-person perspective is often found to increase the engagement in the persuasion literature [24]. This motivated our exploration of the effect of AI uncertainty expressions taking a first-person or general perspective, though we explore the negative persuasive effect (i.e., deterring overreliance) of expressing uncertainty.

Many methods have been proposed for estimating and communicating uncertainty in AI systems [13]. Most relevant, several studies have shown that communicating uncertainty can reduce overreliance. In the context of house valuation, Prabhudesai et al. [78] found that visualizing a system’s uncertainty using quantile dot plots forced participants to slow down and think analytically. In the context of sentiment classification and question answering, Bansal et al. [7] found that displaying a system’s numerical confidence improved participants’ task performance.

A few studies have looked specifically at natural language expressions of uncertainty in AI systems. They have been studied in the chatbot literature as a design strategy to prevent user frustration from conversational breakdowns [5, 37]. Radensky et al. [80]...
studied communicating the confidence of a conversational music recommendation system through language and color coding. Experimenting with different levels of calibration, they found that accurate confidence signals led to the greatest increase in trust without encouraging overreliance, though underreliance may still occur. Zhang et al. [106] explored different ways of having an AI system (named “ShapeBot”) express confidence in the reasoning behind its recommendation, examining the effect of point of view (first-person “I think...” vs. third-person “ShapeBot thinks...”) and strength of the belief expressed (“ShapeBot thinks...” vs. “ShapeBot knows...”). They found that both factors affected user reliance, highlighting the importance of carefully considering the language used to express (un)certainty.

### 2.2 Uncertainty in LLMs

Obtaining accurate numerical estimates of uncertainty for LLMs is an active line of research. One way to estimate an LLM’s uncertainty is by the likelihood of generating a specific output given the context. However, this “generation probability” may not reflect what end users expect or want when they think of uncertainty [91]. A more useful notion of uncertainty might be one that captures how likely it is that the LLM’s output is factually correct or correctly meets the user’s needs. This notion of uncertainty may apply to a full output or to sentences, phrases, or words within the output. Many researchers are working on evaluating how calibrated existing uncertainty estimates are and proposing new techniques to improve calibration [4, 26, 31, 47, 51, 53, 67, 88, 110]. Current findings suggest that LLMs are often overconfident [26, 67, 103], which may give a false impression of their capabilities and exacerbate overreliance [34].

There is also a growing interest in LLMs’ ability to directly generate natural language expressions of (un)certainty [51, 67, 103, 110]. Notably, Mielke et al. [67] observed that LLMs regularly express confidence (e.g., “Obviously...”) and doubt (e.g., “I’m not sure, but...”) through the language used in their outputs, but these expressions are poorly calibrated. Zhou et al. [110] “taught” OpenAI’s GPT-3 [18] model to express (un)certainty through prompt engineering, but also found that the generated expressions were not well calibrated, especially those suggesting high certainty.

Despite this active research, there has been little empirical work examining the impact of uncertainty expression on users of LLM-infused systems. Notable exceptions are the works of Vasconcelos et al. [91] and Spatharioti et al. [84], who explored the effect of highlighting uncertain parts of LLM outputs in the context of code completion and search, respectively, and the concurrent work of Zhou et al. [109], who explored the effect of LLMs’ natural language expressions of (un)certainty in the context of trivia question answering. These studies’ results support uncertainty expression as a promising technique to encourage appropriate reliance, particularly when uncertainty estimates are well calibrated. Our work adds empirical knowledge on this topic through a large-scale, pre-registered experiment studying natural language uncertainty expressions in the context of LLM-infused search.

To avoid making assumptions of calibration or tying our experiment to a particular uncertainty estimation approach, we design our study to include both instances in which the system expresses uncertainty when it is incorrect and instances in which it expresses uncertainty when it is correct. By randomly varying whether or not uncertainty is expressed on any particular response, we are able to directly compare participants’ behavior when uncertainty is and is not expressed.

### 2.3 Measuring Reliance and Trust

We hypothesize that whether or not an AI system expresses uncertainty — and if it does, the perspective in which the uncertainty is expressed — impacts user reliance and trust. We note that there are many definitions, measures, and factors of reliance and
trust [27, 46, 90, 93, 107]. We use a combination of dependent variables and a mix of behavioral and self-reported measures to capture aspects most relevant to our research setting of LLM-infused search.

In the AI-assisted decision-making literature, many experiments follow the judge-advisor paradigm [15], presenting a participant with an AI-generated answer to a question and then asking the participant to provide their own answer, a set-up we adopt in our work. In this set-up, agreement between a participant’s answer and that of the AI system is a commonly studied behavioral measure of reliance and trust [19, 22, 49, 55, 61, 69, 104, 108]. While this does not capture reliance or trust directly — the participant may have come up with the same answer on their own even without the AI system — comparing how often participants agree with the AI system’s work [22, 29, 45, 61, 77], we ask participants to report their whether they rely on the system differently across conditions. We use this measure in our work. We note that in some prior work, the participant is asked to provide an initial answer first before seeing that of the AI system [57, 58, 77, 81]. In these cases, other metrics like weight of advice can be used to more directly capture reliance. We do not adopt this set-up because we use questions we do not expect participants to be able to answer on their own.

To complement agreement, we examine participants’ confidence, source usage, trust intentions, and trust beliefs. First, as in prior work [22, 29, 45, 61, 77], we ask participants to report their confidence both in the answer output by the AI system and in their own answer. Second, as an indirect measure of reliance and trust, we capture participants’ source usage by both tracking whether or not participants click on the linked sources in the system’s responses and asking them to self-report the resources that they based their final answer on. Finally, using responses collected in an exit questionnaire, we measure participants’ trust intentions and trust beliefs using the scales developed by McKnight et al. [39]. Trust intentions refer to a participant’s desire to use the system, while trust beliefs refer to their perceptions about the system’s trustworthiness such as the system’s perceived ability, benevolence, and integrity [64]. In general, the two are positively related [39], but Radensky et al. [80] found that they can be differently impacted by a system’s expressed confidence. We measure both in our experiment to better understand the impact of the system’s uncertainty expression.

We also measure two system facets that are known to impact trust: perceived anthropomorphism and perceived transparency. Participants may view the expression of uncertainty (especially first-person) as an inherently human behavior, leading to increased anthropomorphism. Recent work has expressed concern around anthropomorphism leading to over-trust [1, 83], a potential path for uncertainty expression to backfire. Uncertainty expression can also increase the system’s perceived transparency [13], which is generally shown to enhance trust, whether or not appropriate [50].

In addition to reliance and trust, we also consider task performance as a dependent variable, measured as correctness of participants’ answers and time on task. Both of these have been studied in prior work on AI-assisted decision making [48], as well as specifically for studying the effect of uncertainty expression [92].

We formally define the dependent variables in Section 3.2 and articulate hypotheses in Section 3.3.

3 METHODS

As described in Section 2.3, our experiment is designed to measure the impact of natural language expressions of an LLM’s uncertainty on user reliance and trust. We do this in the context of information seeking in the medical domain. We pre-registered our experimental design, hypotheses, analysis plan, and data collection procedures before collecting data. To complement our pre-registered analyses, we include exploratory analyses and a qualitative analysis of participants’ free-form responses. The study was approved by our internal Institutional Review Board (IRB).

3.1 Procedure and Experimental Conditions

We designed a between-subjects experiment with some within-subjects comparisons, which we conducted on Qualtrics. Participants complete a set of information-seeking tasks. Each task involves determining the correct yes-or-no answer to a challenging, factual question in the medical domain with or without access to responses from a fictional LLM-infused search engine, “AI system A.” The presence and form of system responses provided to participants depend on their experimental condition. Specifically, participants are randomly placed into one of four experimental conditions:

- CONTROL: Participants see AI responses without any expression of uncertainty.
- UNCERTAINIST: Participants see AI responses and half of the time these responses include uncertainty expressed in the first person, with personal pronouns (e.g., “I’m not sure, but it seems...”).
- UNCERTAINGENERAL: Participants see AI responses and half of the time these responses include uncertainty expressed in a general perspective, without personal pronouns (e.g., “There is uncertainty, but it seems...”).
- No-AI: Participants are not told about the AI system and do not see AI responses.

CONTROL is a baseline to which we compare the conditions UNCERTAINIST and UNCERTAINGENERAL to understand the impact of uncertainty expressions. No-AI is a second baseline to understand the impact of access to the AI system.

The experiment is divided into three components. In the first, participants are introduced to the study and to AI system A (if applicable). They are given several task comprehension questions and are asked to complete an example task.

In the second component, participants answer a total of eight questions (details in Section 3.4). They are told they can use any resources they want as in natural settings. For each question, participants, except for those in the No-AI condition, are provided with responses from AI system A (Figure 1). The system’s yes-or-no answers within their responses are correct for only half the questions. In the UNCERTAINIST and UNCERTAINGENERAL conditions, the AI system expresses uncertainty in its answers for half the questions — we refer to them as uncertain answers versus not uncertain answers. We chose this breakdown to have sufficient data for each of the four possible scenarios of correct/incorrect answers with/without uncertainty expression. We randomize the order in which questions are presented, as well as the set of questions for which the AI system expresses uncertainty. However, since we based the AI system’s
answers on real responses from Copilot in Bing (see Section 3.4), the set of answers and their correctness are fixed.

In the final component, participants fill out an exit questionnaire about their experience with and perception of the AI system (if applicable), their background on LLMs, and basic demographic information; see Appendix D. Lastly, participants are debriefed and reminded that some of the AI responses they saw may have contained inaccurate information.

3.2 Dependent Variables
We now formally define the dependent variables (DVs) that we measured, motivated in Section 2.3. First, for each of the eight questions, we measured the following DVs based on participants’ observed behavior:

- **Agree**: TRUE if the participant’s final answer is the same as the AI system’s answer; FALSE otherwise.
- **Correct**: TRUE if the participant’s final answer is correct; FALSE otherwise.
- **Time**: Number of minutes from when the participant saw the task to when they clicked next.
- **LinkClick**: TRUE if the participant clicks on one or more links in the system’s answer; FALSE otherwise.

We additionally measured the following DVs based on participants’ self-reported ratings:

- **UseAI**: TRUE if the participant selected “AI system A’s answer” in the question “What is your final answer based on?” (Select all that apply); FALSE otherwise.
- **UseInternet**: TRUE if they selected “Your own reading of the linked sources in AI system A’s answer” in the above question; FALSE otherwise.
- **ConfidenceAI**: Rating on the question “How confident are you in AI system A’s answer?” on a 5-point scale.
- **ConfidenceAnswer**: Rating on the question “How confident are you in your final answer?” on a 5-point scale.

Finally, based on responses to the exit questionnaire, we calculated the following indexes (all on a 5-point scale):

- **TrustBelief**: Average rating on six statements adapted from the trust scale by McKnight et al. [39].
- **TrustIntention**: Average rating on four statements adapted from the trust scale by McKnight et al. [39].
- **Anthropomorphism**: Average rating on four items from the Godspeed Questionnaire Series [8].
- **Transparency**: Average rating on two statements: “I feel I had a good understanding of what AI system A’s answers were based on” and “I feel I had a good understanding of when AI system A’s answers might be wrong.”

Full details are in Appendix D. Note that some DVs were not applicable for the No-AI condition, where we measured only Agree, Correct, Time, UseInternet, and ConfidenceAnswer. Here we made one (and only one) minor deviation from our pre-registration by including Agree. Although participants in this condition do not see the AI system’s answers, this gives us a baseline for how often participants would arrive at the same answer on their own.

3.3 Hypothesis & Analysis
We expected the presence and perspective of uncertainty expression to impact participants’ reliance, trust, and performance. Formally, for each DV, we hypothesized that condition affects DV. For each repeatedly measured DV, we additionally hypothesized that whether or not uncertainty was expressed in a particular AI response affects DV. We tested our hypotheses with the following pre-registered, confirmatory analyses, for which we present results in Section 4.

We first test the main effect of the conditions with a between-condition analysis. For repeatedly measured DVs, we fit the model $DV \sim Condition + (1|participant) + (1|question)$ with Control as the reference level for Condition. Then to compare the effects of the two conditions with uncertainty, we used a Wald test to test the equality of the corresponding coefficients. For DVs measured once in the exit questionnaire, we use analysis of variance (ANOVA) to compare means across the conditions. If significant, we conduct pairwise comparisons with a post-hoc Tukey test.

Next, we test the effect of uncertainty being expressed or not in a particular response with a within-condition analysis for conditions with uncertainty expression. For repeatedly measured DVs, we fit the model $DV \sim AIUncertain + (1|participant) + (1|question)$, where AIUncertain is TRUE if the AI response is uncertain and FALSE otherwise. We fit this model once for data from the Uncertain1st condition and once for data from UncertainGeneral.

We complement the confirmatory analyses with two additional analyses. First, we conduct an exploratory analysis of the effect of AI’s uncertainty expression on over- and underreliance by separately analyzing cases where the AI system gave correct versus incorrect answers. Analysis details and results are presented in Section 5.1. Second, we conduct a thematic analysis [16, 17] of free-form responses from participants in the Uncertain1st and UncertainGeneral conditions describing their experience with and perception of the AI system. The first author drafted the codebook and conducted the initial coding, then discussed the results with all authors and refined the coding together. We describe how the AI system’s uncertainty expression affected participants in Section 4 along with the quantitative results, and describe how participants interpreted the system’s uncertainty expression in Section 5.2.

3.4 Questions and AI Responses Used
We selected a set of factual questions for participants to answer according to the following criteria: (1) most lay people should not know the answer; (2) the question and answer should not directly show up when using popular search engines; and (3) the answer can be objectively and automatically assessed. To satisfy the criteria, we constructed a set of yes/no medical questions. We began with questions from the MedQuAD dataset [10] and made minor modifications to some to increase the difficulty of finding an answer. We verified that each question does not show up as it is and can not be immediately answered using popular search engines, and consulted multiple sources to confirm the correct answer.

To create AI responses that are realistic and reflect the state-of-the-art in LLM-infused search, we input the selected questions into Microsoft’s Copilot in Bing. All responses were obtained in July 2023. To keep the fluency, style, and content of responses
as realistic as possible, we made only minor modifications such as presenting in-line citations using square brackets instead of superscripts (see Figure 1) and starting each response with “Yes” or “No” for consistency (most responses from Copilot in Bing did this already). We did not make substantive changes to the content.

To create the uncertain versions of the responses, we drew on Oba and Berger [72]. For Uncertain1st, we replaced the leading “Yes” or “No” with an expression of the form “I’m not [certain, sure], but [it seems to me, I would guess, I’d guess that],” selecting one phrase from each set of brackets. Similarly, for UncertainGeneral, we used an expression of the form “[It’s unclear, It’s not clear, There is uncertainty], but it seems like.” Current LLMs already output expressions such as “I’m not sure, but” and “I’m not sure, but my guess is” [67], so we believe these insertions preserve the realism of responses. The only difference between the three versions of the AI system’s responses is the presence and perspective of uncertainty expression; their information content is otherwise identical.

We selected the final eight questions such that: (1) four questions have a correct answer of “Yes” and four have a correct answer of “No” so that always selecting “Yes” is no better or worse than random guessing; (2) the AI answers are correct for four and incorrect for four, so that always agreeing with the AI system is no better or worse than random guessing; (3) questions are not too easy to answer without access to the AI system (determined via piloting). The final set of questions used is provided in Appendix E along with the original responses from Copilot in Bing and our modified responses.

3.5 Data Collection and Participants

We conducted our experiment on Amazon Mechanical Turk (MTurk), a crowdsourcing platform widely used for human-subject experiments. Research has shown data from MTurk workers is comparable to data from other pools (e.g., commercial panels, social media, colleges) [12, 20, 23, 32, 63, 74], but recently there has been a decrease in data quality [28, 44, 60, 62, 94, 98]. Indeed, in pilot studies we found that a strong requirement on qualification was necessary to obtain meaningful data, with the highest quality data obtained when requiring a “Masters” qualification (granted by Amazon based on past performance). Since the pool of available participants with a Masters qualification is limited, we pre-registered a recruitment plan in which we would initially aim to recruit 432 (determined via a power analysis) U.S.-based MTurk workers with a Masters qualification, 99% or higher approval rating, and at least 2000 completed human intelligence tasks (HITs), removing the Masters requirement after 7 days if we were unable to meet our target sample size. In parallel, we implemented best practices to mitigate the use of bots and improve data quality, including CAPTCHAs, honeypot questions, speed checks, attention checks, and open-ended questions, which we used to define data exclusions.

We collected 656 complete responses over the course of two weeks in September 2023, of which we excluded 252 (38.4%) based on five pre-registered exclusion criteria. Our final sample consists of 404 responses: 104 in Control, 92 in Uncertain1st, 94 in UncertainGeneral, and 114 in No-AI. See Appendix B for more on our data collection procedures.

Participants were paid $5 USD. The payment was determined based on the expected experiment duration of 20 minutes (estimated from pilot studies) and target hourly wage of $15. The actual median experiment duration was 20.5 minutes, so on average, participants were paid $14.80 per hour. See Appendix A for more information about participants.

4 RESULTS: CONFIRMATORY ANALYSIS

We now present the results of our pre-registered, confirmatory analyses. Tables 1 and 2 contain the between-condition and within-condition analysis results, respectively. We refer the reader to Section 3.2 for definitions of all DVs. We present the estimated means (and standard errors) from the fitted models, calculated without conditioning on the random effects, and represent binary variables in percentages. We use significance to refer to statistical significance at the level of $p < 0.05$.

4.1 Agreement with AI: Agree

We begin with the results of agreement, a commonly used behavioral measure of reliability and trust. Our first finding is that people tend to agree with the AI system when its responses are provided. Our between-condition analysis suggests that participants with access to the AI system are significantly more likely to submit the same answer as the system than those who do not have access (80.9% Control vs. 58.4% No-AI). We next find that AI’s uncertainty expression decreases agreement with the AI system.

Compared to Control (80.9%), Agree is significantly lower in Uncertain1st (74.8%). It is also lower in UncertainGeneral (77.6%), although the difference is not significant. Our within-condition analysis suggests AI’s uncertainty expression decreases agreement at the instance level as well. The estimated means of Agree for not uncertain vs. uncertain AI responses are 84.7% vs. 70.9% for UncertainGeneral (significantly different) and 79.5% vs. 73.4% for Uncertain1st (not significantly different). We find support for this finding in the qualitative data as well. 10 participants (out of 186 in Uncertain1st and UncertainGeneral) stated that when they disagreed with the system it was, as one put it, “because of the uncertainty of the answers provided by AI.”

Other reasons participants mentioned for disagreement included the system’s answer being different from their own knowledge (e.g., “A couple of the AI’s answers didn’t make sense so from my own common sense I had to make my own judgement”) or the information in other resources (e.g., “The answer seemed to contradict the links given, or I could not find how they came to that answer with the information from the links given”) and having lower trust in the system.

4.2 Confidence in Answers: ConfidenceAI, ConfidenceAnswer

Looking at participants’ self-reported confidence in answers, we find that AI’s uncertainty expression decreases people’s confidence in its answer. Compared to Control (3.95 on a 5-point scale), ConfidenceAI is significantly lower in Uncertain1st (3.66). It is also lower in UncertainGeneral (3.80), although the difference is not significant. Our within-condition analysis provides further evidence for this finding and suggests that AI’s uncertainty expression decreases people’s confidence in their final answer.
Table 1: Between-condition analysis. We compare DVs across conditions. We report the model-estimated means (and standard errors) from our confirmatory analysis. The rightmost column shows pairs of conditions with statistically significant differences with significance marked as * (p < 0.05) or ** (p < 0.01).

<table>
<thead>
<tr>
<th>Sec.</th>
<th>DV</th>
<th>Control (%)</th>
<th>Uncertain1st (%)</th>
<th>UncertainGeneral (%)</th>
<th>No-AI (%)</th>
<th>Significant differences†</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Agree (1-5)</td>
<td>80.9% (5.5)</td>
<td>74.8% (6.7)</td>
<td>77.6% (6.2)</td>
<td>58.4% (8.5)</td>
<td>Control &lt;** No-AI</td>
</tr>
<tr>
<td></td>
<td>ConfidenceAI (1-5)</td>
<td>3.95 (0.17)</td>
<td>3.66 (0.17)</td>
<td>3.80 (0.17)</td>
<td></td>
<td>Uncertain1st &lt;** Control</td>
</tr>
<tr>
<td>4.2</td>
<td>ConfidenceAnswer (1-5)</td>
<td>4.30 (0.08)</td>
<td>4.34 (0.08)</td>
<td>4.27 (0.08)</td>
<td>4.22 (0.08)</td>
<td>Control &lt;** No-AI</td>
</tr>
<tr>
<td></td>
<td>LinkClick (%)</td>
<td>2.7% (2.1)</td>
<td>7.2% (4.8)</td>
<td>3.9% (3.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UseAI (%)</td>
<td>77.3% (5.1)</td>
<td>64.8% (6.9)</td>
<td>72.3% (6.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>UseLink (%)</td>
<td>74.7% (5.5)</td>
<td>85.5% (4.0)</td>
<td>81.7% (4.8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UseInternet (%)</td>
<td>19.1% (5.6)</td>
<td>27.0% (7.4)</td>
<td>23.2% (6.7)</td>
<td>92.9% (2.5)</td>
<td>Control &lt;** No-AI</td>
</tr>
<tr>
<td></td>
<td>TrustBelief (1-5)</td>
<td>3.90 (0.06)</td>
<td>3.86 (0.07)</td>
<td>4.00 (0.07)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>TrustIntention (1-5)</td>
<td>3.25 (0.10)</td>
<td>2.91 (0.10)</td>
<td>3.36 (0.10)</td>
<td></td>
<td>Uncertain1st &lt; Control</td>
</tr>
<tr>
<td></td>
<td>Anthropomorphism (1-5)</td>
<td>3.07 (0.10)</td>
<td>3.00 (0.11)</td>
<td>3.13 (0.11)</td>
<td></td>
<td>Uncertain1st &lt; UncertainGeneral</td>
</tr>
<tr>
<td></td>
<td>Transparency (1-5)</td>
<td>4.04 (0.06)</td>
<td>3.93 (0.07)</td>
<td>4.01 (0.07)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>Correct (%)</td>
<td>63.9% (8.6)</td>
<td>72.8% (7.4)</td>
<td>67.9% (8.1)</td>
<td>74.2% (7.1)</td>
<td>Control &lt;** Uncertain1st</td>
</tr>
<tr>
<td></td>
<td>Time (min)</td>
<td>2.13 (0.22)</td>
<td>2.10 (0.23)</td>
<td>2.03 (0.22)</td>
<td>1.57 (0.21)</td>
<td>Control &lt;** No-AI</td>
</tr>
</tbody>
</table>

at the instance level. For both Uncertain1st and UncertainGeneral, ConfidenceAI and ConfidenceAnswer are significantly lower on instances with uncertain (vs. not uncertain) AI responses. Indeed, one participant in the Uncertain1st condition stated, “If the AI didn’t seem confident, I would like [sic] on the links. If the AI seemed confident I assumed he was correct most of the time.”

4.3 Source Usage: LinkClick, UseAI, UseLink, UseInternet

Looking at source usage, we find that people with access to the AI system conduct their own Internet search less frequently than those without access. UseInternet is notably lower in Control (19.1%) than No-AI (92.9%). However, we see no significant differences in source usage between the three conditions in which AI responses are present, meaning there is no evidence that the presence and perspective of AI’s uncertainty expression affect people’s source usage behavior. On the other hand, from our within-condition analysis, we find that AI’s uncertainty expression has a significant effect on the self-reported DVs at the instance level. On instances with uncertain AI responses (vs. not uncertain responses), UseAI is significantly lower in both Uncertain1st and UncertainGeneral, and UseInternet is significantly higher in Uncertain1st. These results suggest that at the instance level, AI’s uncertainty expression decreases the use of the system’s answer and increases the use of other resources. In their free-form responses, 11 participants stated that the system’s uncertainty motivated them to verify information using the links provided in the answer (e.g., “I clicked on the links provided when the AI was uncertain of the answers”) while 13 mentioned it motivated them to perform their own search (e.g., “When the AI wasn’t certain, I searched on Google”).

We note that both LinkClick and UseLink are intended to capture whether or not participants read the linked sources provided in the AI system’s responses, but the estimated means of LinkClick (2.7% – 7.2%) from the between-condition analysis are much lower than those of UseLink (74.7% – 85.5%). There are several factors that might contribute to this discrepancy. First, these estimated means are from different models that include participants and questions as random effects. The intercepts for random effects in these models are quite high. Looking at the raw data, there is still a gap, but not as large: the actual means are 30.6% – 34.9% for LinkClick and 64.2% – 71.6% for UseLink. Second, there could have been measurement error from self-report bias for UseLink [35]. Finally, some participants could have considered their answers to be based on their “own reading of the linked sources” if they read the list of links, even if they didn’t click to open them. For example, one participant wrote, “If the link title gave the same answer as the AI answer, then I assumed it was the right answer.” This discrepancy emphasizes the value of including both behavioral and self-reported measures.
Table 2: Within-condition analysis. For each of the two conditions with uncertainty, we compare DVs measured on instances with AI answers that are not uncertain vs. uncertain. We report the model-estimated means (and standard errors) from our confirmatory analysis. > and < note statistically significant differences with significance marked as * (p < 0.05) or ** (p < 0.01).

<table>
<thead>
<tr>
<th>Sec.</th>
<th>DV</th>
<th>UncertainNot Uncertain</th>
<th>UncertainNot Uncertain</th>
<th>UncertainUncertain</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Agree (%)</td>
<td>79.5% (7.2)</td>
<td>73.4% (8.6)</td>
<td>84.7% (5.0)</td>
</tr>
<tr>
<td></td>
<td>ConfidenceAI (1-5)</td>
<td>3.88 (0.20)</td>
<td>3.44 (0.20)</td>
<td>3.95 (0.16)</td>
</tr>
<tr>
<td></td>
<td>ConfidenceAnswer (1-5)</td>
<td>4.42 (0.09)</td>
<td>4.26 (0.09)</td>
<td>4.36 (0.07)</td>
</tr>
<tr>
<td>4.2</td>
<td>LinkClick (%)</td>
<td>14.3% (6.9)</td>
<td>9.9% (5.1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UseAI (%)</td>
<td>73.2% (7.2)</td>
<td>57.4% (8.9)</td>
<td>79.7% (4.7)</td>
</tr>
<tr>
<td></td>
<td>UseLink (%)</td>
<td>86.5% (4.2)</td>
<td>84.3% (4.7)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UseInternet (%)</td>
<td>23.0% (6.3)</td>
<td>34.1% (7.8)</td>
<td>19.2% (6.5)</td>
</tr>
<tr>
<td>4.3</td>
<td>Correct (%)</td>
<td>73.6% (8.9)</td>
<td>75.7% (8.4)</td>
<td>69.4% (10.1)</td>
</tr>
<tr>
<td></td>
<td>Time (min)</td>
<td>2.00 (0.29)</td>
<td>2.19 (0.29)</td>
<td>1.84 (0.25)</td>
</tr>
</tbody>
</table>

4.4 Trust and Perception of AI: TrustBelief, TrustIntention, Anthropomorphism, Transparency

Moving onto trust and perception of AI, we find that the presence and perspective of uncertainty expression neither affect people's trust beliefs nor the perceived anthropomorphism and transparency of the system. There are no significant differences in TrustBelief, Anthropomorphism, and Transparency between Control and the two conditions with uncertainty. Overall, participants had somewhat positive trust beliefs about the system (TrustBelief is around “4: Somewhat agree” for all conditions); reported that they had a somewhat good understanding of what the AI system’s answers were based on and when they might be wrong (Transparency is around “4: Somewhat agree”); and had neutral perceptions of anthropomorphism (Anthropomorphism is around “3: Neutral”). In contrast, we find that first-person expressions of uncertainty decrease trust intentions while expressions from a general perspective do not. TrustIntention is significantly lower in Uncertain1st (2.91) compared to both Control (3.25) and UncertainGeneral (3.36), indicating a lower desire to use the system. Illustrative of this, one participant in the Uncertain1st condition stated, “[The AI system] was very non-committal in its answers so I didn't feel I could trust it.” This suggests that frequent first-person expressions of uncertainty can lead people to view the system as less trustworthy and decrease their desire to use it. More about the used scales and item-level results is in Appendix C.2.

4.5 Task Performance: Correct, Time

Finally, we analyze participants’ task performance. From our between-condition analysis of Correct, we first find that having access to the AI system decreases people’s accuracy. Participants with access to the system have significantly lower accuracy than those without (63.9% Control vs. 74.2% No-AI). However, this result should be interpreted in the context of the AI system’s low overall accuracy (50.0% in our experimental setup). Second, we find that AI’s uncertainty expression increases people’s accuracy. Correct is significantly higher in Uncertain1st (72.8%) than Control (63.9%). It is also higher in UncertainGeneral (67.9%), but the difference is not significant.

From our between-condition analysis of Time, we find that having access to the AI system increases task time. Time is significantly higher in Control (2.13 min) than No-AI (1.57 min). Together with the results on correctness, this suggests that having access to the AI system decreases overall performance in our experimental setup. Between Control and the two conditions with uncertainty, there are no significant differences in task time. However, our within-condition analysis suggests that AI’s uncertainty expression increases task time at the instance level. The estimated means of Time for not uncertain vs. uncertain AI responses are 1.84 min vs. 2.23 min for UncertainGeneral (significantly different) and 2.00 min vs. 2.19 min for Uncertain1st (not significantly different). An explanation of this result is that AI’s uncertainty expression slows people down to use more caution when completing the task.

5 RESULTS: ADDITIONAL ANALYSES

5.1 Effect of Uncertainty Expression on Over- and Underreliance

In Section 4.1, we analyzed participants’ agreement with the AI system as a measure of reliance. Agreement can be appropriate or inappropriate, depending on the correctness of the AI system’s answers; agreeing with the system when it is incorrect is a sign of overreliance, whereas disagreeing with it when it is correct is a sign of underreliance. To better understand the extent to which uncertainty expression leads to either of these phenomena, we take inspiration from the analysis of Chen et al. [25] and separately analyze participants’ agreement on questions the AI system answers correctly and questions it answers incorrectly. Note that, having conditioned on the (in)correctness of the AI system, analyzing agreement is equivalent to analyzing correctness of people’s final answers. We present the results here in terms of correctness to focus attention on whether reliance is appropriate (i.e., beneficial to the user) or not.
Figure 2: Exploratory analysis of over- and underreliance (Section 5.1). We analyze Correct (%) separately for questions the system answered correctly vs. incorrectly. We show the model-estimated means and standard errors for each condition (Figure 2a) and for not uncertain vs. uncertain responses in the conditions with uncertainty (Figure 2b).

5.2 Participants’ Interpretations of AI’s Uncertainty Expression

While our quantitative results shed light on whether expressions of uncertainty impact reliance and trust, they cannot tell us why. We next explore participants’ interpretations of the expressed uncertainty via a thematic analysis of free-form responses to the question “When and why do you think AI system A expresses uncertainty?” in the exit questionnaire.

The majority of participants (N=102 of the 186 in the conditions with uncertainty) attributed the system’s expressed uncertainty to its inability to answer a particular question, for example because it could not find an answer, found conflicting or unreliable answers, or could not understand the information it found. Three suggested the system was programmed to express uncertainty, as in “I would guess there’s some sort of certainty variable and if the score is below a level, an uncertainty message is included in the result.” Another ten suggested the expressed uncertainty is due to the inherent difficulty of the question. As one put it, “It could be a question that is very hard to come up with a simple yes or no answer.”

These interpretations are all in line with the goal of reducing overreliance: if the system is unable to answer a question or the question is inherently difficult, users should verify the answer for themselves. Five participants explicitly interpreted the expressed uncertainty as a way of encouraging users to check their answers, for example, “I suppose the AI wanted us to do further research in those cases where it could not be 100% sure of the answer.”

A small number of participants attributed the uncertainty to other reasons, such as impression management (“to appear more human, encourage confidence, and appear thoughtful”), maintaining credibility (“it doesn’t want to risk being wrong on something and ruining its credibility”), avoiding liability (“it expresses uncertainty to absolve it of responsibility in the event it is wrong”), or restrictions on answering medical questions (“the AI is programmed not to dispense medical advice which could potentially be harmful without a caveat”). These interpretations may not necessarily reduce overreliance.
One might ask whether participants interpreted the two types of uncertainty expression differently. Indeed, we found that participants in the UncertainGeneral condition were more likely than those in Uncertain1st (51.5% vs. 41.3%) to attribute the uncertainty to the AI system finding conflicting or unreliable information or the question being inherently hard, whereas those in Uncertain1st were more likely (20.7% vs. 7.4%) to attribute it to limitations of the AI system itself.

6 DISCUSSION

Our results suggest that expressing uncertainty through natural language can be an effective way to reduce overreliance and (over)trust in LLM-infused search engines. Expressions of uncertainty led to more cautious behaviors, from taking longer to arrive at an answer to reporting more relevance on outside sources. However, it did not fully eliminate overreliance; the participants with the highest task performance were those without access to AI responses.

We find that perspective matters: uncertainty expressions in first-person show stronger effects than general perspective. This is consistent with prior findings that first-person messages increase recipients’ involvement and engagement [24, 106] compared with general or third-person messages. This persuasive effect of first-person expressions should be interpreted with caution: while it helps heighten the warning effect of a negative message such as uncertainty, it might amplify a positive message, even if unjustified. For example, one may want to avoid first-person expressions of confidence because they may exacerbate overreliance and overtrust, as found in prior work [109]. There are also concerns around harms from anthropomorphism of AI systems that may stem from over-trust, deception, threats to human agency, and propagation of stereotypes [1]. While we did not observe that first-person uncertainty expression increases perceived anthropomorphism, people can start assigning social attributes to machines without conscious awareness [71]. Future research and practices should further explore the long-term effects of interacting with AI systems expressing uncertainty in a first-person perspective and consider other potential negative effects of anthropomorphism.

Our research has implications both for those building and deploying LLMs and LLM-infused applications and for policymakers regulating the use of AI. Most critically, any approach to reducing overreliance should be validated through empirical research. There may also be tradeoffs when balancing over- and underreliance. In our study, the most successful approach to reducing overreliance was to use first-person uncertainty expression, but this also decreased participant trust in the AI system, which may be undesirable in settings where people already under-trust the AI system. We believe there is no one-size-fits-all approach to implementing natural language uncertainty expression. For these reasons, given that the issue of how to manage overreliance is of particular importance to regulators, we advocate for raising awareness of the complexities of mitigating overreliance and for customized, evidence-based solutions, rather than universal ones.

There are limitations to our research. The widespread deployment of LLMs is still relatively new and the human-computer interaction and broader research communities are still grappling with the question of how to design effective studies to understand how end users perceive and interact with them. There is always a tradeoff between the controllability of the experiment and the generalizability of the conclusions to user behaviors in their day-to-day tasks [65]. In order to be able to measure agreement and correctness, we adopted an experimental set-up inspired by the AI-assisted decision-making literature in which study participants provide simple yes/no answers to questions. This approach does not allow exploring how the expression of uncertainty would impact people’s behavior when completing more complex tasks, like writing an article or planning a trip. Further, our measurements of time and source usage are less reliable than they would have been had we opted for an in-person lab study. We chose questions from the medical domain, where overreliance is particularly concerning, but people may behave differently when seeking information about their own medical symptoms rather than answering a pre-defined set of questions. People may also behave differently when given the chance to interact with the system repeatedly as opposed to in a single session. The AI system in our study exhibited low accuracy and expressed uncertainty often, in a poorly calibrated manner. These design choices may have impacted our results — particularly the lower task performance when given access to the AI system. Also, there may be differences across cultures and languages in how people interpret or react to uncertainty. Our study was conducted in English with U.S.-based participants and results may not generalize to other cultural and linguistic contexts.

For all of these reasons, while our findings suggest that natural language expressions of uncertainty could be an effective approach to reducing overreliance, we caution against overgeneralizing from our study. Instead, we view our results as evidence that language choices matter in how people perceive and act on the outputs of LLMs, and teams building and deploying LLMs should therefore evaluate them carefully with end users before release.

7 ETHICAL CONSIDERATIONS AND POSITIONALITY

We conclude with a reflection on the ethical considerations of our work and our positionality.

Mitigating harms to human subjects. We recruited U.S.-based participants on MTurk, which many people rely on as a primary source of income. As discussed in Section 3.5, we aimed to provide an hourly wage of $15 USD. We came close to this goal, with participants receiving an estimated $14.80 per hour on average. (This is likely an underestimate of average wage, since we have no way to know if workers spent time on other activities between accepting the task and completing it.) This is substantially higher than the U.S. federal minimum wage of $7.25 per hour, though a few states have recently adopted a minimum wage of $15/hour or higher. As discussed in the FaCT 2023 panel “The Humans Behind the Intelligence: Speaking with Data Workers,” our choice to limit participation to workers with a 99% or higher approval rating, at least 2000 completed tasks, and in some cases, a Masters qualification, prevented workers who are new to MTurk from participating. We made the decision to include these qualifications after piloting several versions of the study with less restrictive qualifications and
finding that the data quality was too poor to use. We paid and approved the work of everyone who completed the study, regardless of whether their responses passed our quality checks. At the end of the study, we debriefed participants, reminding them that the medical information output by the AI system was sometimes incorrect. We did not collect personally identifiable information except for MTurk IDs, which were used to ensure that workers who participated in pilots of our study did not participate in the main study. These were deleted when no longer needed. Our procedure was reviewed and approved by our internal IRB and we obtained participant consent.

**Potential negative societal impact.** While our results provide evidence for the effectiveness of natural language expressions of uncertainty for reducing overreliance, generalizing too heavily from our findings could lead to potential harms. Given the limitations of our research (see Section 6), teams deploying LLMs or LLM-infused applications should not make decisions about how to express uncertainty to end users without extensive user testing in their own contexts. They also should not assume that they have addressed overreliance by expressing uncertainty. (Indeed, in our study we see that participants still have higher task performance with no access to the AI system.) Likewise, regulators should avoid making blanket requirements on uncertainty expression, at least until more research has been done.

Separate from these potential unintentional misuses of our research, there is a possibility that bad actors could strategically incorporate the expression of uncertainty into an LLM’s output to make them more persuasive, regardless of whether they represent objective fact, potentially contributing to the spread of misinformation.

**Positionality.** Our research questions and design were influenced by our position as employees of a U.S.-based technology company. Members of our research team have first-hand experience observing and participating in discussions regarding the responsible development and deployment of LLM-infused applications, which has shaped our understanding of gaps in knowledge and other practical challenges that are in attempting to meet responsible AI principles and proposed regulatory requirements. We had access to sufficient contexts. They also should not assume that they have addressed overreliance by expressing uncertainty. (Indeed, in our study we see that participants still have higher task performance with no access to the AI system.) Likewise, regulators should avoid making blanket requirements on uncertainty expression, at least until more research has been done.

Separate from these potential unintentional misuses of our research, there is a possibility that bad actors could strategically incorporate the expression of uncertainty into an LLM’s output to make them more persuasive, regardless of whether they represent objective fact, potentially contributing to the spread of misinformation.

**Positionality.** Our research questions and design were influenced by our position as employees of a U.S.-based technology company. Members of our research team have first-hand experience observing and participating in discussions regarding the responsible development and deployment of LLM-infused applications, which has shaped our understanding of gaps in knowledge and other practical challenges that are in attempting to meet responsible AI principles and proposed regulatory requirements. We had access to sufficient contexts. They also should not assume that they have addressed overreliance by expressing uncertainty. (Indeed, in our study we see that participants still have higher task performance with no access to the AI system.) Likewise, regulators should avoid making blanket requirements on uncertainty expression, at least until more research has been done.

Separate from these potential unintentional misuses of our research, there is a possibility that bad actors could strategically incorporate the expression of uncertainty into an LLM’s output to make them more persuasive, regardless of whether they represent objective fact, potentially contributing to the spread of misinformation.

Our view that there are potential benefits of responsibly deployed LLM-infused applications is likely influenced by our experience in industry, yet we acknowledge some fundamental limitations of using LLMs for information retrieval [82]. We encourage future research on overreliance from research teams in academia and civil society.

**ACKNOWLEDGMENTS**

We thank the participants for their time and effort and the anonymous reviewers for helpful comments. We also thank Jake Hofman, Dan Goldstein, Sophie Spatharioti, Amy Heger, Samir Passi, Harsh Kumar, Wesley Deng, Snial Prabhudesai, Abhraneel Sarma, Yeji Park, Nina Grigić-Hláca, Julia Stoyanovich, Jessica Hullman, and the Microsoft Research New York City lab and FATE group for thoughtful discussion and feedback. SK acknowledges support from the NSF Graduate Research Fellowship and the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award to her PhD advisor.

**REFERENCES**


[10] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmatc...


[98] Margaret A. Webb and June P. Tangney. 0. Too Good to Be True: Bots and Bad Data From Mechanical Turk. Perspectives on Psychological Science 0, 0 (0), 17459616221120027. https://doi.org/10.1177/17456916221120027 PMID: 36534213


