Causal Perception in Question-Answering Systems

Po-Ming Law
Georgia Institute of Technology
pmlaw@gatech.edu

Leo Yu-Ho Lo
The Hong Kong University of Science and Technology
yhload@cse.ust.hk

Alex Endert
Georgia Institute of Technology
endert@gatech.edu

John Stasko
Georgia Institute of Technology
stasko@cc.gatech.edu

Huamin Qu
The Hong Kong University of Science and Technology
huamin@cse.ust.hk

ABSTRACT

Root cause analysis is a common data analysis task. While question-answering systems enable people to easily articulate a why question (e.g., why students in Massachusetts have high ACT Math scores on average) and obtain an answer, these systems often produce questionable causal claims. To investigate how such claims might mislead users, we conducted two crowdsourced experiments to study the impact of showing different information on user perceptions of a question-answering system. We found that in a system that occasionally provided unreasonable responses, showing a scatterplot increased the plausibility of unreasonable causal claims. Also, simply warning participants that correlation is not causation seemed to lead participants to accept reasonable causal claims more cautiously. We observed a strong tendency among participants to associate correlation with causation. Yet, the warning appeared to reduce the tendency. Grounded in the findings, we propose ways to reduce the illusion of causality when using question-answering systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

correlation and causation, question answering

ACM Reference Format:
Po-Ming Law, Leo Yu-Ho Lo, Alex Endert, John Stasko, and Huamin Qu. 2021. Causal Perception in Question-Answering Systems. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3411764.3445444

1 INTRODUCTION

Root cause analysis is a common task during data analysis. Such analysis provides explanations for events in business processes, observations about human behaviours, and phenomena in society. A business analyst, for instance, may seek explanations for a revenue decrease to identify supply chain bottlenecks and marketing strategies [1]. To help people acquire this important skill, colleges and online learning platforms have offered courses on root cause analysis [11, 15].

The need for root cause analysis skills is not only limited to professional analysts. Open data create opportunities for anyone to engage in personal data projects. Visualization hobbyists, for example, may conduct data analysis on public data and create fascinating visualizations on platforms such as Makeover Monday [50]. Individual citizens might analyze data about the social issues they are concerned about and write a blog post about the analysis [40, 46]. However, root cause analysis could be challenging to these people since they might lack domain knowledge and analysis skills.

Systems with question-answering functionality present a resource that people can utilize to explain data observations even without significant expertise in data analysis. Users of these systems can easily articulate their why questions through natural language [16] or point and click [64]. The systems then employ advanced statistical analysis to infer answers. Some technologists believe that question-answering interfaces will become the norm in analytics platforms [21].

However, causal inference from observational data (as opposed to randomized experiments) is challenging [55]. These question-answering systems often produce unintuitive answers to a user's why questions. Figure 1 shows Explain Data, a question-answering functionality in Tableau [64]. The user observes that Massachusetts has the highest average ACT Math score among all US states. Being curious, she asks Explain Data to provide explanations for the high score. Explain Data infers that the rate of teenage pregnancy is negatively correlated with ACT Math score and that the low rate of teenage pregnancy in Massachusetts may lead to the high ACT Math score. The veracity of the explanation is questionable.

When these systems do not always provide reliable results, a concern is their potential power to persuade people into believing causal claims (e.g., low teenage pregnancy rate in Massachusetts may lead to the high average ACT Math score) that may not be true. Does visualizing correlation (e.g., using a scatterplot) increase the plausibility of a causal claim and user trust in the system even when the claim does not make sense? Does warning users about the potential flaws in the system help them adopt the answers more cautiously? Answering these questions could help understand designs that ensure judicious use of the computational outputs.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445444

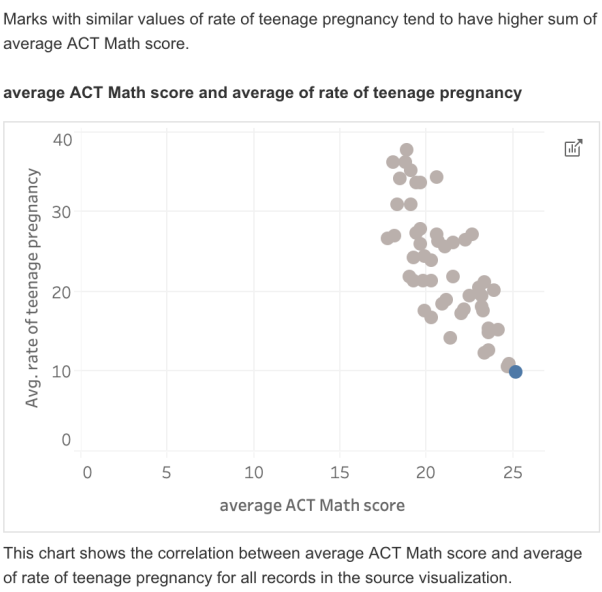

Figure 1: An answer generated by Tableau Explain Data [64]. The user asks about the high ACT Math score in Massachusetts. Explain Data infers that teenage pregnancy rate and ACT Math score are negatively correlated and the low teenage pregnancy rate in Massachusetts might cause the high ACT Math score. It shows the data using a scatterplot in which each dot is a state and the blue dot is Massachusetts.

This paper investigates the impacts of different information (a scatterplot, a description about correlation, and a warning message) shown alongside a causal claim on the perceptions of a question-answering system. We conducted two crowdsourced studies with 200 participants each. In both studies, participants reviewed a series of answers to why questions. These answers were presented with different designs. Across different designs, we compared the perceived plausibility of the causal claims, user trust in the question-answering system, the awareness of the system's flaws, and users' tendency to associate correlation with causation. Whereas the first study presented answers with different degrees of plausibility, the second study presented only reasonable answers.

From the first study, we found that participants tended to disagree less with an unreasonable causal claim when a scatterplot was presented alongside the claim. In contrast, participants appeared to accept a reasonable causal claim more cautiously when they were shown a simple warning about the system's potential confusion of correlation and causation. We further observed a general tendency among participants to associate correlation with causation, but the warning seemed to reduce the tendency. We did not observe these effects in the second study where the system only provided reasonable causal claims.

Question-answering systems often employ data visualizations to provide context for their answers [25, 26]. Our results reveal that these systems could leverage the persuasive power of visualizations to create an illusion of causality: Although scatterplots only provide evidence about correlation, presenting scatterplots next to a causal claim could increase users' tendency to agree with the claim. Based on our findings, we suggest that users should be skeptical when considering answers that are automatically generated and propose design ideas to encourage skepticism.

2 RELATED WORK

Our work intends to understand how the visual design of answers to why questions might influence the perceptions of a question-answering system. We draw on research relating to the impact of visualization design on data interpretation as well as question-answering systems more broadly.

2.1 Impact of Visualization Design on Data Interpretation

Visualization design holds significant power to shape data interpretation [51]. Researchers have investigated a wide range of factors such as knowledge, perceptual biases, and cognitive biases that influence the messages communicated to viewers.

Knowledge external to visualizations often affects how we interpret the visualizations. As users look at a visualization, they often apply their domain knowledge [56]. Xiong et al. [69] showed that this prior knowledge could prime a viewer to obtain a particular message from a visualization and lead the viewer to believe that other viewers would receive the same message. Besides prior knowledge, social information also affects data interpretation. Kim et al. [32] found that seeing others' expectations about the data influenced people's trust in the accuracy of the data.

Moreover, perceptual biases play a role in manipulating data interpretation [14, 52]. For example, distorting the aspect ratio of a line chart can lead to an inaccurate assessment of trends in the data [27]; truncating the y-axis in a bar chart exaggerates effect sizes [13]; the neighborhood of a bar in a bar chart can change the perceptions of the bar's height [71]. However, these biases could be mitigated through judicious design. For instance, Ritchie et al. [54] showed that an animated transition from an untruncated bar chart to a truncated one could avoid misinterpretation.

Cognitive biases can further change the lens through which we interpret visualizations [18]. An example is priming and anchoring effects. Calero Valdez et al. [65] conducted experiments to show that the judgment of class separability in scatterplots depended on the scatterplots users saw before. Biases in data interpretation can also have consequences on decision making. Dimara et al. [17] provided evidence that the presence of dominated data points in a scatterplot influenced the judgement of which points were dominating.

Besides knowledge and biases, subtle design choices also matter to data interpretation: Titles can have a misleading impact on visualization interpretation [6, 31, 35, 36]; visual embellishments can affect the insights we gain from visualizations [7, 49].

While correlation does not imply causation, it is easy to confuse them, leading to an illusion of causality [47]. Xiong et al.'s [68] found that this illusion would increase with the aggregation level of data visualizations. Instead of studying data aggregation, we investigated the effects of different information on perceived causality when using question-answering systems. Specifically, we studied

whether two forms of correlational evidence (scatterplot and textual description about correlation) could create causal illusion and whether a simple warning could reduce the illusion.

2.2 Question-Answering Systems and User Perception

Technologists have developed question-answering systems to meet users' information needs in various domains including sports [72], work settings [44], and data science [20]. These systems exhibit a wide variety of designs. Some (e.g., conversational agents or chatbots) mimic natural human conversations and can understand a rich diversity of topics [3]. Others resemble web search and focus only on a small set of tasks [37, 38].

In data visualization, researchers have developed natural language interfaces to facilitate visual data analysis [58, 61, 70]. Many of these systems aim to address specific usability challenges as users employ natural language for data analysis. For example, users' utterances are often ambiguous. Datatone utilizes ambiguity widgets to expose the ambiguity and let users correct the system's decisions [24]. Moreover, conversations happen in some context on which utterance semantics depend [30]. To address this, Evizeon provides pragmatics support to retain contextual information and infer a user's meaning based on the context [30].

Another line of research focuses on understanding the impact of system behaviors and information presentation on the perceptions of these systems. Liao et al. [43] investigated how agent sociability influences user interactions with conversational agents. Ashktorab et al. [4] studied preferences for different strategies to handle conversational breakdowns. Hearst et al. [25, 26] investigated the visual designs of answers provided by a natural language interface and how users perceive these designs. In a similar vein, we intend to provide insights into how the visual design of answers to why questions might affect user perceptions of a question-answer system.

3 PRE-STUDY: COLLECTING CAUSAL STATEMENTS

As a starting point to understand the appropriate presentation of answers to why questions, we focus on why questions about extremum (i.e., an extreme value). An example is why students in Massachusetts have high ACT Math scores on average (Fig. 1). Finding extremum is a common task during data analysis [2]. Also, functionality to answer such questions has emerged in commercial systems such as Tableau [64]. Findings from our studies could offer design guidelines in practice.

In study 1, we showed participants a series of answers to why questions. We created answers with different visual designs and assessed user perceptions of the system given the designs. Due to the inherent challenges in causal inference [55], question-answering systems occasionally provide unreasonable answers to why questions. To emulate these systems, we selected causal claims with different levels of plausibility as answers presented to participants. To select these causal claims, we conducted a pre-study.

3.1 Methods

3.1.1 Datasets. We planned to generate causal claims that were backed up by observational data and considered using synthetic data. However, our goal was to study user perceptions of a system, and the credibility of the data might affect user perceptions. To control for the potential experimental confounds, we used real-world data instead.

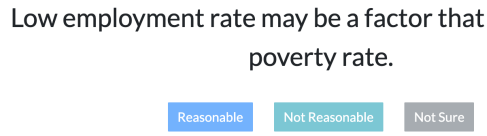

Figure 2: Interface used in the pre-study.

We first curated a dataset about states in the US from sources including US Census Bureau [8], National Center for Education Statistics [22], and Kaiser Family Foundation [23]. The curated dataset has 258 attributes about demographics, healthcare, and education for each US state. We chose these topics because they are accessible to laypeople. This enabled participants to judge the plausibility of the generated causal claims based on common sense.

With the curated data table, we computed the Pearson correlation for all attribute pairs. To find attribute pairs with a potential causal relationship, we collected the ones with a high correlation (above 0.7 or below -0.7). For each of the 1522 attribute pairs with a high correlation (e.g., employment rate and poverty rate), we found a state (e.g., Mississippi) that has an extreme value for *both* attributes and omitted the attribute pairs where such a state did not exist.

Based on the collected attribute pairs and states, we generated causal claims (e.g., low employment rate in Mississippi may be a factor that leads to the high poverty rate in Mississippi). The plausibility of this claim can be affected by the plausibility of the causal relationship (e.g., employment rate affects poverty rate) and that of the information about the state (e.g., Mississippi has a high poverty rate). Since we intended to assess the plausibility of the causal relationship, we removed the states from the causal claims (see Fig. 2).

An author carefully picked 30 reasonable claims, 30 unreasonable claims and 30 claims that were hard to tell if they were reasonable (hereafter, *hard-to-tell claims*). We verified and ranked the plausibility of these claims through a study on Amazon Mechanical Turk (MTurk).

3.1.2 Participants. We randomly segmented the 90 claims into five batches of 18 claims and recruited 20 workers on MTurk to rate each batch (100 unique workers in total). We limited the tasks to workers in the United States and had an acceptance rate of 95% or above. During data analysis, we omitted participants who failed to pass attention checks (but compensated them for participation). We recruited participants until reaching the target sample size for each batch. Participants were compensated $1 for the study that took approximately 5-10 minutes.

Among the 100 participants, 55 were male, and 45 were female. They aged 22-64 (M=35.5, SD=11.2). Participants reported their educational attainment to be high school (8 participants), professional school (18), college (49), graduate school (17), PhD (7), and postdoctoral (1).

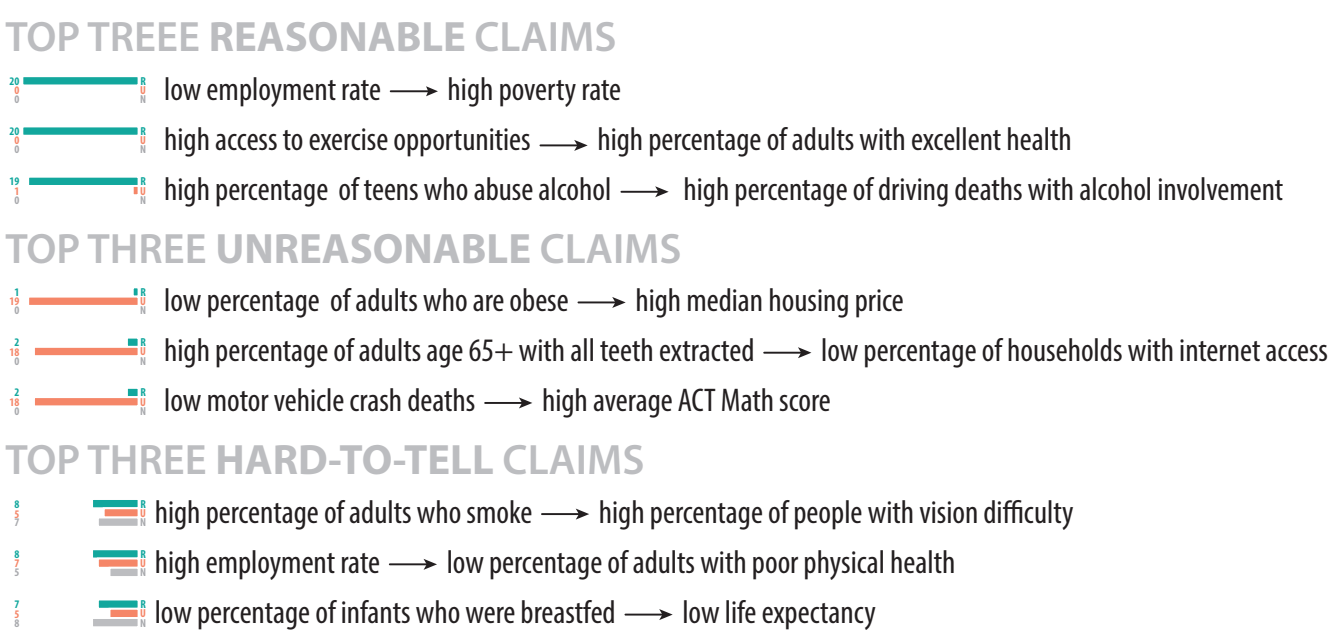

Figure 3: The top three reasonable, unreasonable, and hard-to-tell causal claims. Each row shows a causal claim (right) and a bar chart that visualizes the votes (left). The green, red, and gray bars represent the votes for *Reasonable, Unreasonable,* and *Not Sure* respectively. For example, the most reasonable claim was "a low employment rate may be a factor that leads to a high poverty rate." It got 20/20 votes for *Reasonable* (R), 0/20 vote for *Unreasonable* (U), and 0/20 vote for *Not Sure* (N).

3.1.3 Procedure. Each participant was randomly assigned to rate one of the five batches of 18 claims. Participants first filled out a demographic survey on their gender, age, and highest education level. They then saw a series of 18 causal claims that were presented on separate pages (Fig. 2). We randomized the presentation order of these claims to prevent order effects. Based on the plausibility of each claim, participants selected one of the three options: *Reasonable*, *Unreasonable*, and *Not Sure*. As each participant rated more than a dozen causal claims, we used the three options rather than a Likert scale with five options or more to keep the study short. During the study, participants also answered two attention check questions asking them to directly select one of the three options.

3.2 Results

For each causal claim, we computed the probabilities that participants selected *Reasonable*, *Unreasonable*, and *Not Sure*. We then calculated the entropy for each claim. A low entropy implies that participants mostly voted for the same option, whereas a high entropy means that participants' votes tended to distribute across the three options. Within each bucket of the 30 reasonable claims, 30 unreasonable claims, and 30 hard-to-tell claims, we ranked the claims by entropy.

For the 30 reasonable claims, we ranked them in increasing order of entropy. The top claims had a low entropy because participants mostly voted for *Reasonable*. For the 30 unreasonable claims, we again ranked them in increasing order of entropy. Participants mostly selected *Unreasonable* for the top claims. For the 30 hard-to-tell claims, we expected that the claims where the plausibility was the most difficult to judge had a high entropy score. This is because participants likely struggled to choose among the options. We sorted these claims in decreasing order of entropy.

Figure 3 shows the top three reasonable, unreasonable, and hard-to-tell claims based on the above ranking. We provide a ranked list of all the 90 claims as a supplementary material. Study 1 confirmed the validity of our results: In study 1, when participants rated the claims on a 7-point Likert scale, they tended to agree with the top reasonable claims, disagree with the top unreasonable claims, and be neutral about the top hard-to-tell claims.

4 STUDY 1: PROVIDING ANSWERS WITH DIFFERENT PLAUSIBILITY

With the ranked lists of reasonable, unreasonable, and hard-to-tell claims, we designed a between-subject experiment during which participants reviewed a series of answers to why questions.

When designing the presentation of answers, we considered its complexity to typical end users. One way to answer why questions (e.g., why Mississippi has a high poverty rate) is causal graph [55, 66], a technique often employed in statistics literature for visualizing complex causal relationships. However, causal graphs may require more advanced statistics training to understand.

We also considered showing multiple factors in the answer but were concerned about introducing experimental confounds. For example, the number, the perceived plausibility, and the underlying causal relationship of the factors could potentially alter the perception of system performance. Yet, using real-world data implied that these variables could be difficult to control for.

We therefore adopted a simplified design where the system responded to a why question by stating a factor that could answer

Figure 4: The four experimental conditions. A user asks about the high poverty rate in Mississippi (a). The system answers only a causal claim (b), shows a scatterplot next to the claim (c), adds a description about the correlation (d), and warns about the system's flaws besides showing the previous information (e).

the question. In each task, participants saw a why question (e.g., why Mississippi has a high poverty rate) and the system's answer (e.g., low employment rate in Mississippi may be a factor that leads to the high poverty rate). Across conditions, the answers had different designs (Fig. 4b-d). We provide screenshots of the experiment interface as a supplementary material.

4.1 Methods

4.1.1 Conditions. Building on prevailing system designs and the research literature, we focused on two types of correlational evidence (scatterplot and textual description about correlation) to investigate whether they created an illusion of causality. We further studied the effectiveness of warning in reducing the illusion. Here, we describe these three types of information:

Scatterplot. Scatterplots are common for showing the relationship between two numerical variables [57]. They have also been applied in question-answering functionality in commercial systems for showing the relationship between cause and effect (Fig. 1).

Textual description about correlation. While the causal claim (e.g., low employment rate in Mississippi may be a factor that leads to the high poverty rate) describes a single state in the US, a description about correlation (e.g., as employment rate decreases, poverty rate tends to increase) depicts the overall trends for all the states. To facilitate interpretation, visualization systems often provide such descriptions next to a chart [39, 60].

Warning message. Although scatterplots and the textual descriptions only reveal correlation, they might induce an illusion of causality [68]. A mitigation strategy is to use a message to warn users that correlation is not causation. While such warnings are less common

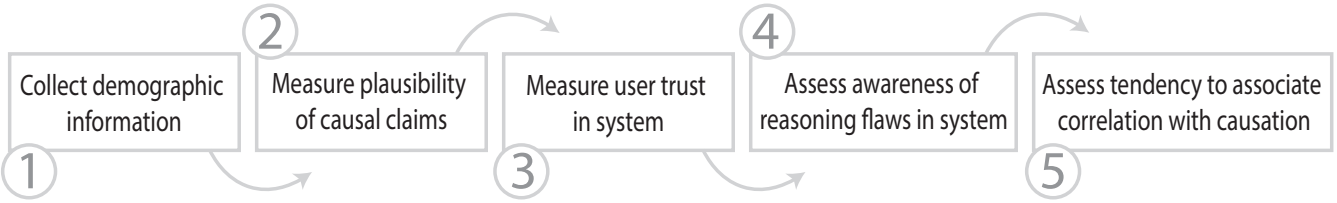

Figure 5: The five stages in study 1 and study 2.

in visualization systems, they are commonly used in other systems (e.g., web browser) to prompt safety-related behaviors (e.g., not to click on phishing websites) [19]. It would be interesting to learn about if a simple statement is enough to raise awareness of the system's potential flaws and reduce users' tendency to confuse correlation and causation.

Based on these three information types, we designed four answer interfaces by adding the information types one by one. Participants were randomly assigned to a condition where the answers adopted one of the four designs:

Claim only (Fig. 4b). The system only shows a claim about cause and effect as an answer to a why question.

Claim + vis (Fig. 4c). Beside the causal claim, the system visualizes the cause and effect using a scatterplot. Prior studies showed that the aspect ratio of point clouds in a scatterplot affects correlation estimation [10, 48]. To support a consistent correlation estimation, we controlled the aspect ratio. For each axis, we set the lowest value to be (min value of the data $-$ 0.15 $\times$ range of the data) and the highest value to be (max + 0.15 $\times$ range).

Claim + vis + description (Fig. 4d). The system additionally states the correlation between the cause and effect variables with a textual description.

Claim + vis + description + warning (Fig. 4e). To encourage users to evaluate the answers carefully, the system warns that the scatterplot only shows correlation, and that correlation is not causation.

4.1.2 Participants. A power analysis indicated that for a significance level of 0.05 and a power of 0.8, detecting a medium effect size of $f = 0.25$ using one-way ANOVA required 180 participants (45 participants per condition). As we planned to conduct non-parametric tests (see Sec. 4.1.4), we targeted a slightly larger sample size (200 participants in total or 50 participants per condition) following guidelines on sample size determination for non-parametric tests [41].

During participant recruitment, we limited the study to workers in the United States, had an acceptance rate of 95% or above, and did not participate in the pre-study. The study took approximately 10-20 minutes, and we compensated participants $2.90. At the end, we recruited 200 unique workers on MTurk.

The survey had two interpretation checks for assessing scatterplot comprehension and three open-ended questions (details in the Procedure section). We excluded participants who did not pass any of the interpretation checks or provided gibberish answers for any of the open-ended questions (but compensated them for participation). Overall, the data quality was poor. For example, many participants provided canned responses for some open-ended questions. We omitted 123 participants and continued recruiting until reaching the target sample size.

Participants aged 20-69 (M=35.4, SD=10.1). 131 were male, 68 were female, and 1 preferred not to say. They reported different educational attainments: high school (29 participants), professional school (22), college (109), graduate school (35), PhD (1), and post-doctoral (4). Concerning data analysis expertise, 44 had none, 64 were beginners, 71 were intermediate, and 21 were advanced. For experience with visualization platforms (e.g., Tableau), 82 had none, 60 were beginners, 36 were intermediate, and 22 were advanced. When asked about the frequency of using question-answering systems, 132 reported never, 30 reported rarely, 23 reported weekly, and 15 reported daily.

4.1.3 Procedure. We first randomly assigned participants to one of the four conditions. For all conditions, the study consisted of five main stages (Fig. 5).

In stage 1 (Fig. 5 ①), participants filled out a demographic survey. After filling out the survey, they completed a practice task to get acquainted with the study interface.

In stage 2 (Fig. 5 ②), participants reviewed a series of nine answers to why questions. In each task, they examined a data observation (e.g., Mississippi has the highest poverty rate among all US states) (Fig. 4a), a why question (e.g., why is poverty rate in Mississippi so high?) (Fig. 4a), and the system's answer to the question (Fig. 4b-d). Depending on the condition, participants saw a different visual design for the answers. Based on the system's answer, participants rated their agreement with a causal claim (e.g., low employment rate in Mississippi is a factor that leads to high poverty rate in Mississippi) on a 7-point Likert scale.

We constructed the nine answers using the top nine causal claims obtained from the pre-study (Fig. 3). Hence, three answers were reasonable, three were unreasonable, and three had plausibility that was difficult to judge. This intended to mirror real-world systems that tend to be unreliable in answering why questions. The order of the answers was randomized to prevent order effects.

After participants reviewed the nine answers, we measured user trust in the system in stage 3 (Fig. 5 ③). Participants rated their trust in the system on a 7-point scale from -3 (I don't trust it at all) to +3 (I fully trust it). They further shared their reasons for trusting or not trusting the system.

Next, we assessed their awareness of the reasoning flaws in the system in stage 4 (Fig. 5 ④). Participants reported whether they

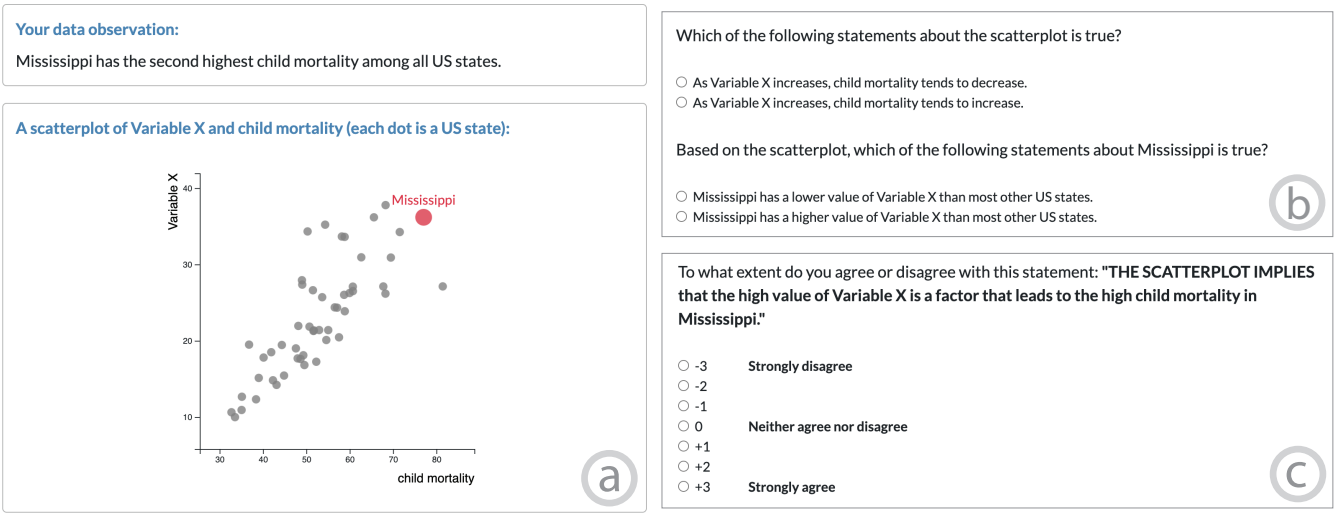

Figure 6: Measuring tendency to associate correlation with causation. Participants saw a data observation and a scatterplot (a), answered interpretation check questions (b), and rated their agreement on a statement suggesting that the scatterplot implied causation (c).

observed any reasoning flaws in the system. If the answer was "yes," we asked them to specify the reasoning flaw(s) they found.

In the final stage (Fig. 5 ⑤), we assessed their tendency to associate correlation with causation. Participants in stage 5 saw a description of a data observation (Mississippi has the second highest child mortality among all US states) and a scatterplot showing a strong correlation between child mortality and an unknown variable X (Fig. 6a). To assess participants' understanding of scatterplots, we first asked participants to answer two interpretation check questions (Fig. 6b). Participants who failed to pass any of the questions were excluded from the data analysis.

Whereas participants rated their agreement with a causal relationship in stage 1, participants rated their agreement with a sentence stating that a scatterplot with a high correlation implied a causal relationship in stage 5. Participants saw a statement: "*The scatterplot implies* that the high value of Variable X is a factor that leads to the high child mortality in Mississippi" (Fig. 6c). They rated the statement on a 7-point Likert scale and explained why they agreed or disagreed.

4.1.4 Quantitative Measures. We derived six measures from participants' response.

Agreement (reasonable). For each participant, we computed the average agreement rating for the three reasonable answers.

Agreement (unreasonable). It is the average rating for the three unreasonable answers.

Agreement (hard to tell). It is the average rating for the three answers that were hard to tell if they were reasonable.

Trust. Some researchers have developed questionnaires to assess user trust in recommender systems [53] and machine learning systems [9]. Since these questionnaires may not be applicable to question-answering systems, we tailored a question to assess trust in question-answering systems. In the post-study survey, we asked, "Overall, how much do you trust or not trust the question-answering system?" and participants rated on a scale from -3 to +3.

Awareness of system's flaws. We computed the number of participants who selected "yes" for the question, "Did you observe any flaw(s) in the reasoning of the question-answering system?" Unlike the other measures that are scales between -3 and +3, this measure is a count between zero and 50. Whereas trust and agreement with answers are more subjective, observations about reasoning flaws in the system are more clear-cut, making a yes/no question more suitable.

Awareness of "correlation is not causation". In the last part, participants rated a statement: "*The scatterplot implies* that the high value of Variable X is a factor that leads to the high child mortality in Mississippi." (Fig. 6c) If participants were cautious about drawing causal conclusions from correlation, they should be inclined to disagree with the statement.

During a pilot study, we observed that when the variable name was shown, participants tended to use their common sense to decide if they agreed with the statement. Yet, we wanted to assess tendency to confuse correlation and causation instead of ability to apply common sense. To reduce the impact of common sense in answering the question, we hid the variable name of X.

Likert-scale data are not continuous and violate the ANOVA assumptions. To study the main effect of answer design, we used a Kruskal-Wallis test, which is a non-parametric equivalence of one-way ANOVA, for the five measures using a 7-point scale (i.e., all measures except awareness of system's flaws). When there is a significant main effect, we conducted post-hoc Wilcoxon rank sum tests with a Holm-Bonferroni correction for pairwise comparisons.

For the awareness of system's flaws, we used a Fisher's exact test to assess if the number of participants who found reasoning flaws in the system was significantly different across conditions.

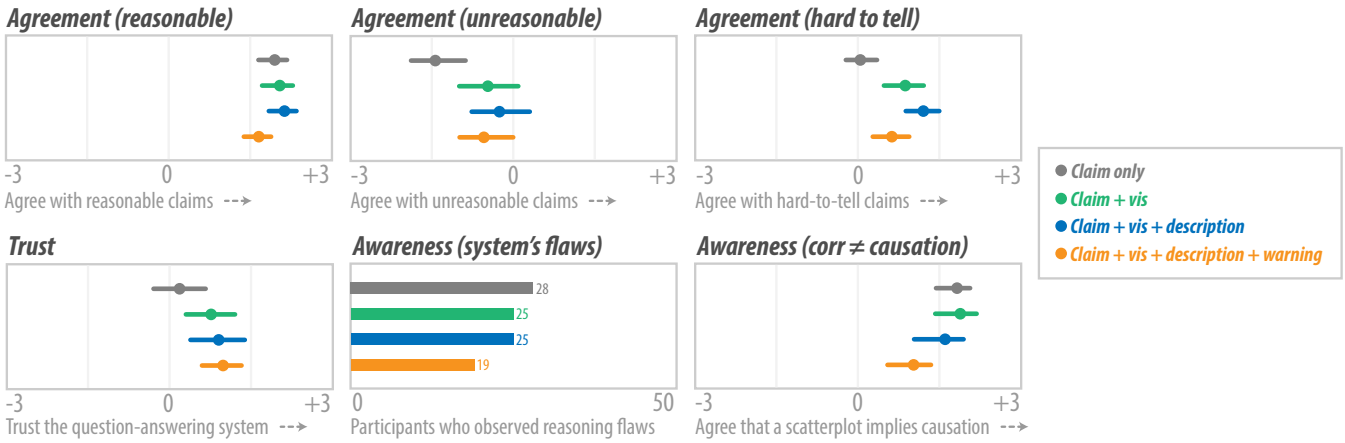

Figure 7: Quantitative results from study 1. All error bars show 95% bootstrapped confidence intervals.

4.1.5 Qualitative Response. There were three open-ended questions in the survey, one for explaining trust or distrust, one for specifying reasoning flaws in the system, and one for explaining why agreed or disagreed that the scatterplot implied causation.

For each question, an author open-coded the responses to identify the emergent categories and develop a codebook. We observed that a response could include multiple categories. Hence, we treated each category as binary: For each response, we labelled whether each category was present or absent. Two coders independently coded all responses. We then discussed inconsistencies, refined code definitions, and independently re-coded the responses based on the new definitions. We iteratively coded the responses until we reached a Cohen's κ above 0.7 for all the categories.

For each category, we conducted a Fisher's exact test to determine whether its presence was significantly different across conditions.

4.1.6 Hypotheses. We developed hypotheses based on research in visualization's persuasive power, trust in automated systems, and warning science.

Pandey et al. [51] found that when participants did not have a strong attitude towards a topic, visualizations had a strong power to change their attitudes. They also commented on the difficulty to change attitudes for topics of which participants already had a strong prior opinion [51]. We expected that showing a scatterplot would increase the plausibility of hard-to-tell claims because participants likely did not have a strong attitude towards them. We also expected that the scatterplot would not affect the plausibility of reasonable and unreasonable claims.

H1.1: Participants' agreement with the reasonable claims does not differ across conditions.

H1.2: Participants' agreement with the unreasonable claims does not differ across conditions.

H1.3: Participants in the three conditions that show a scatterplot in the answers (i.e., claim + vis, claim + vis + description, and claim + vis + description + warning) agree with the hard-to-tell claims more than participants in the claim-only condition.

Transparency in automated systems can inspire user trust [59]. For example, when a recommender system provides reasons behind its recommendations, users tend to trust the system more [29]. Showing the data can increase the transparency in the question-answering system. We posited that users would trust the system more when it showed the scatterplot.

H1.4: Participants in the three conditions that show a scatterplot in the answers trust the question-answering system more than participants in the claim-only condition.

Some researchers in warning science have compared the effectiveness of passive and active warnings [19]. Whereas active warning forces users to notice it by blocking user tasks, passive warning (e.g., a simple warning message) is less interrupting [19]. In data analysis, passive warning is more suitable because a small latency in interaction can hamper analysis quality [45]. However, Egelman et al. [19] showed that passive warnings were often ineffective because users might ignore them. The ineffectiveness might extend to question-answering systems. Hence, we posited that the warning message would not increase participants' awareness of the system's flaws nor decrease their tendency to associate correlation with causation.

H1.5: Participants' awareness of the system's flaws does not differ across conditions.

H1.6: Participants' awareness of "correlation is not causation" does not differ across conditions.

4.2 Results

Figure 7 summarizes the results for the quantitative measures. We observed that the scatterplot increased the plausibility of unreasonable and hard-to-tell claims but not reasonable claims. The warning message appeared to decrease the plausibility of reasonable claims but not unreasonable and hard-to-tell claims. Trust and awareness of flaws did not seem to differ across conditions. However, the warning message seemed to increase the awareness of "correlation is not causation."

Also, for the claim-only condition, participants tended to give a neutral rating for the hard-to-tell claims (M=0.04), a positive rating

for the reasonable claims (M=1.95), and a negative rating for the unreasonable claims (M=-1.44). This confirmed the validity of the pre-study results.

In the following, we provide the detailed analysis.

4.2.1 Agreement (reasonable). On a scale from -3 (strongly disagree) to +3 (strongly agree), participants in the claim + vis + description condition rated the reasonable claims the highest (M=2.13, SD=0.92), followed by those in the claim + vis condition (M=2.04, SD=0.98), the claim-only condition (M=1.95, SD=0.96), and the claim + vis + description + warning condition (M=1.65, SD=0.89). A Kruskal-Wallis test indicated a significant main effect of answer design on the rating ($\chi^2(3)$=10.8, p=.013). We conducted six post-hoc pairwise comparisons using Wilcoxon rank sum tests with a Holm-Bonferroni correction. Results showed that only the difference between claim + vis + description + warning and claim + vis + description (p=.017) as well as that between claim + vis + description + warning and claim + vis (p=.044) were significant. The results did not support **H1.1**.

4.2.2 Agreement (unreasonable). Participants in the claim + vis + description condition rated the unreasonable claims the highest (M=-0.26, SD=1.92), followed by those in the claim + vis condition (M=-0.47, SD=1.96), the claim + vis + description + warning condition (M=-0.55, SD=1.82), and finally the claim-only condition (M=-1.44, SD=1.74). A Kruskal-Wallis test indicated a significant main effect of answer design on the rating ($\chi^2(3)$=12.2, p=.007), with post-hoc pairwise comparisons showing that all the three conditions with scatterplots in the answers had a significantly higher average rating than the claim-only condition. The results did not support **H1.2**.

4.2.3 Agreement (hard to tell). Participants in the claim + vis + description condition rated the hard-to-tell claims the highest (M=1.2, SD=1.12), followed by those in the claim + vis condition (M=0.87, SD=1.34), the claim + vis + description + warning condition (M=0.62, SD=1.25), and the claim-only condition (M=0.04, SD=1.05). There is a significant main effect of answer design on the rating ($\chi^2(3)$=24.4, p<.001). Pairwise comparisons showed that all the three conditions with scatterplots in the answers had a significantly higher average rating than the claim-only condition. Other pairs were not significantly different. The findings supported **H1.3**.

4.2.4 Trust. On average, the trust ratings across conditions were positive, indicating a tendency to trust the system. Claim + vis + description + warning has the highest rating (M=0.98, SD=1.31), followed by claim + vis + description (M=0.9, SD=1.76), claim + vis (M=0.76, SD=1.67), and claim-only (M=0.18, SD=1.70). However, we did not observe a significant main effect of answer design on trust ($\chi^2(3)$=7.34, p=.062). The results did not support **H1.4**.

4.2.5 Why trust or not trust? We coded participants' reasons for trusting or not trusting the question-answering system. Seven categories of responses emerged from the analysis. We report the core results here and provide the detailed breakdown of the categories across conditions in the supplementary materials.

For each response, we labelled each category as present or absent. We labelled all categories as absent for responses that were too broad or vague (e.g., *"it is nice"*).

The top three reasons for distrusting the system were some answers did not make sense (40.5% of 200), the system confused correlation and causation (9%), and it did not provide enough support for its causal claims (7.5%). A participant felt that some claims lacked support and wrote, *"Some of the answers could be factual but it was hard to determine without further data."*

The top three reasons for trusting the system were that some answers made sense (26%), the system showed the data (8.5%), and the system provided some support for its causal claims (6.5%)

We did not observe a significant difference in the presence of any of the categories across conditions using Fisher's exact tests (details in supplementary materials).

4.2.6 Awareness of system's flaws. Using a Fisher's exact test, we did not find a significant difference in the number of people who found reasoning flaws (these participants selected "yes" for the question asking whether they observed reasoning flaws) across conditions (p=.33). We could not reject **H1.5**.

4.2.7 What are the reasoning flaws? The qualitative coding resulted in four categories. Among the 97 participants who observed reasoning flaws in the system, the majority of participants stated providing nonsensical answers as a reasoning flaw (70.1% of 97). Other observed reasoning flaws were confusing correlation and causation (15.5%), not having enough support for the claims (8.25%), and considering only one factor (3.09%). Fisher's exact tests did not indicate significant differences in the presence of any of the four categories across conditions.

4.2.8 Awareness of "correlation is not causation". We asked participants to rate a sentence stating that a scatterplot implied causation. On a scale from -3 (strongly disagree) to +3 (strongly agree), claim + vis + description + warning had the lowest rating (M=1.02, SD=1.45), followed by claim + vis + description (M=1.6, SD=1.59), claim-only (M=1.82, SD=1.10), and claim + vis (M=1.88, SD=1.33). All conditions got a positive average rating, indicating a tendency to associate correlation with causation. We found a significant main effect of answer design on the rating ($\chi^2(3)$=15.2, p=.002). Post-hoc pairwise comparisons showed that claim + vis + description + warning had a significantly lower average rating than all the other three conditions, indicating that the warning appeared to reduce the tendency to associate correlation with causation. The results did not support **H1.6**.

4.2.9 Why agree or disagree with the statement? The qualitative coding yielded four categories. We again observed that some responses were overly broad (e.g., *"because the graph shows it"*) and coded all categories as absent for such responses.

Among the more specific responses, the majority of participants agreed that the scatterplot implied a causal relationship because the scatterplot showed a correlation (46% of 200). An example response is *"If Variable X did not rise then child mortality would not rise."*

Participants disagreed with the statement because correlation is not causation (8.5%), variable X was unknown and they could not judge (8.5%), and the scatterplot had outliers (4%). A participant who observed outliers said, *"I only slightly agree because other states show otherwise. Texas, for instance, has a much lower Child Mortality rate but Variable X is almost the same."*

We did not find significant differences in the presence of any of the categories across conditions.

5 STUDY 2: PROVIDING ONLY REASONABLE ANSWERS

Several findings from study 1 deviated from our expectations: The simple warning appeared to decrease the plausibility of reasonable claims and increase the awareness of "correlation is not causation"; we did not have enough evidence that user trust was improved by showing the data. A potential explanation lied in the unreliable performance of the system—it made the warning more noticeable and reduced the effectiveness of showing the data in improving user trust (when the system performed poorly, it was untrustworthy no matter whether it showed the data). To investigate whether the observations in study 1 held for a system that had a higher perceived performance, we conducted study 2.

5.1 Methods

Study 2 was the same as the study 1 except that participants reviewed nine reasonable answers to why questions (as opposed to reviewing answers with different levels of plausibility in study 1). We constructed the answers using the top nine claims in the ranked list of 30 reasonable claims obtained from the pre-study.

We similarly recruited 50 participants per condition (200 unique workers in total). Workers who participated in the pre-study and study 1 were excluded from study 2. Participants aged 18-70 (M=36.1, SD=11.0). 121 were male, 77 were female, and 2 preferred not to say. The reported educational attainments were high school (28 participants), professional school (10), college (116), graduate school (37), PhD (8), and postdoctoral (1). Concerning data analysis expertise, 44 had none, 79 were beginners, 54 were intermediate, and 23 were advanced. For experience with visualization platforms (e.g., Tableau), 84 had none, 47 were beginners, 45 were intermediate, and 24 were advanced. When asked about the frequency of using question-answering systems, 123 reported never, 35 reported rarely, 31 reported weekly, and 11 reported daily.

As the system only presented reasonable answers, study 2 only had four measures: agreement (reasonable), trust, awareness of system's flaws, and awareness of "correlation is not causation."

In study 1, participants heeded the warning, causing them to agree less with reasonable claims and be less likely to associate correlation with causation. We expected that both effects would disappear when the system was more trustworthy. Furthermore, in study 1, showing the data using a scatterplot did not seem to improve user trust in the system. We posited that when the system provided only reasonable answers, showing the data would improve user trust. We considered the same set of hypotheses as in study 1:

H2.1: Participants' agreement with the reasonable claims does not differ across conditions.

H2.2: Participants in the three conditions that show a scatterplot in the answers trust the question-answering system more than participants in the claim-only condition.

H2.3: Participants' awareness of the system's flaws does not differ across conditions.

H2.4: Participants' awareness of "correlation is not causation" does not differ across conditions.

5.2 Results

Figure 8 shows the quantitative results. Kruskal-Wallis tests for agreement (reasonable), trust, and awareness of "correlation is not causation" as well as a Fisher's exact test for awareness of system's flaws indicated no significant differences across conditions (details in the supplementary materials). Hence, the results failed to support **H2.2**. However, we could not reject **H2.1**, **H2.3**, and **H2.4**.

We also observed that participants in study 2 appeared to trust the system more than those in study 1. The mean trust rating in study 2 was 1.70 (SD=1.04) while that in study 1 was 0.71 (SD=1.64). Participants in study 2 also found fewer reasoning flaws in the system. The total number of participants who found reasoning flaws in study 2 was 35 (compared with 97 in study 1). We summarize the qualitative results as follows.

5.2.1 Why trust or not trust? Participants provided diverse reasons for trusting or not trusting the system. Seven categories of reasons emerged from the qualitative coding.

The top three reasons for trusting the system were the answers made sense (38.5% of 200), the system provided enough support for its causal claims (19.5%), and it showed the data (8%). The top three reasons for distrusting the system were the system did not provide enough support for its claims (8%), it considered only one factor (7%), and it confused correlation and causation (2.5%).

Fisher's exact tests indicated that the number of participants stating "the answers made sense" as a reason was significantly different across conditions (p=.014). We conducted six post-hoc pairwise comparisons using Fisher's exact tests with a Holm-Bonferroni correction. We only observed that more participants in the claim-only condition stated "the answers made sense" than in the claim + vis + description condition (p=.025). A potential explanation was that providing a claim only led participants to comment mostly on the plausibility of the claim. However, providing other information (e.g., a scatterplot) alongside a claim enabled them to comment on other aspects and less on the plausibility.

In study 2, 56% of the responses contained reasons for trusting the system while 16% contained reasons for not trusting it. The data stood in contrast to those in study 1. In study 1, 39.5% of the responses had reasons for trust while 49.5% had reasons for distrust. This echoed the finding that participants trusted the system more in study 2.

5.2.2 What are the reasoning flaws? Among the 35 participants who answered "yes" for the question asking whether they observed reasoning flaws in the system, we found three categories of responses after omitting those who provided vague answers: The system considered only one factor (28.6% of 35); it confused correlation and causation (20%); it did not provide enough support for the claims (17.1%). Using Fisher's exact tests, we did not observe significant differences in the presence of the categories across conditions.

5.2.3 Why agree or disagree with the statement? The qualitative analysis resulted in five categories of responses. Congruent with

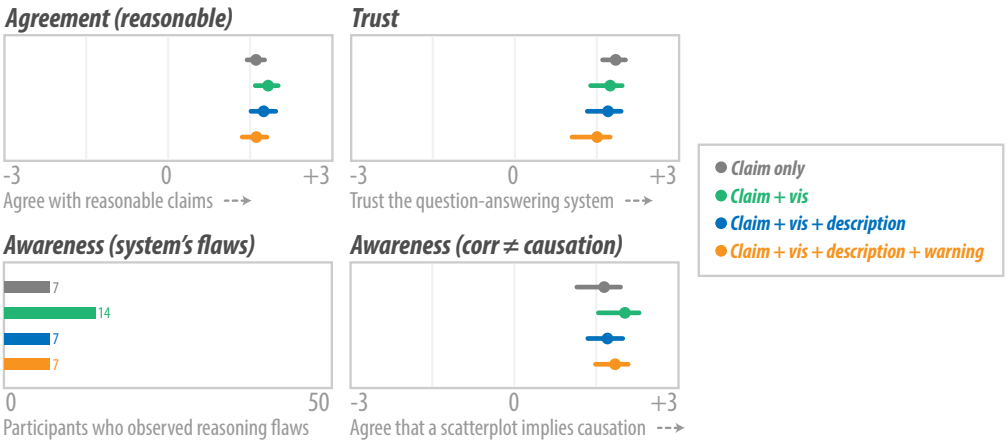

Figure 8: Quantitative results from study 2. All error bars show 95% bootstrapped confidence intervals.

study 1's results, most participants agreed that the scatterplot implied a causal relationship because the scatterplot showed a correlation (43.5% of 200).

Participants who disagreed with the statement commented that correlation is not causation (7.5%), the scatterplot had outliers (7%), variable X was unknown and they could not judge (5.5%), and the dots in the scatterplot looked disperse (1.5%).

Fisher's exact tests did not show a significant difference across conditions for any of the categories.

6 DISCUSSION

Before discussing the implications of our findings, we summarize the results from the two studies and provide potential explanations for the less intuitive observations.

In study 1, participants reviewed answers of different plausibility. We did not observe effects of the textual description about correlation on the perceived plausibility of causal claims, user trust in the system, the awareness of the system's flaws, and the awareness of "correlation is not causation." However, showing a scatterplot caused participants to disagree less with unreasonable claims and agree more with hard-to-tell claims. In contrast, a simple warning message seemed to cause participants to agree less with reasonable claims. The warning also reduced participants' tendency to associate correlation with causation.

Nevertheless, when participants examined only reasonable answers in study 2, the impact of the simple warning message on reducing the plausibility of reasonable claims and on raising the awareness of "correlation is not causation" seemed to disappear. Research in warning science found that arousal strength (i.e., the perceived importance or relevance of a warning) affects the effectiveness of a warning message in motivating safety-related behaviors [28]. Participants in study 2 tended to trust the system more than those in study 1. This likely led participants in study 2 to perceive the warning about the system's reasoning flaws to be less relevant. The warning in study 2 became less effective possibly because participants tended to ignore the warning.

In both studies, we did not observe significant differences in user trust and the awareness of the system's flaws across conditions. The qualitative results provided an explanation. In study 1, when asked about why they did not trust the system or what were the reasoning flaws in the system, most participants simply stated that the answers did not make sense. In study 2, when asked about why they trusted the system, the majority commented that the answers made sense to them. The results appeared to indicate that system performance in answering why questions had a dominating effect on user trust and the awareness of reasoning flaws in the system. In other words, when users can assess system performance, showing other information (e.g., a scatterplot or a warning) may play a small role in shaping user trust and the awareness of flaws.

We observed a tendency for participants to conclude causation from correlation. In both studies, we found that the ratings for the awareness of "correlation is not causation" were positive (i.e., agreeing that the scatterplot implied causation) even when participants were warned that correlation is not causation. How do we reduce the illusion of causality when using question-answering systems? Here, we devise design considerations based on the study results.

6.1 Encouraging Skepticism When Using Question-Answering Systems

A core implication of our results is that question-answering systems could utilize visualizations of correlation to create an illusion of causality: By showing a scatterplot, these systems could increase the likelihood for users to accept causal claims that are unfounded. Mitigating the illusion of causality entails a deliberate design effort. In this section, we argue that encouraging users to be skeptical about automated answers could promote an appropriate interpretation of automatically generated causal claims and propose design ideas to inspire skepticism.

Why should users be skeptical when considering causal claims that are automatically generated? In a perfect world, user trust in a system's answer should match the ground truth—users should trust causal claims only when they are true and distrust false claims (Fig. 9a). In reality, however, belief in causal claims depends on their perceived plausibility despite the ground truth (Fig. 9b). Very often, users can only determine the plausibility of a causal claim but not whether it is true. Hence, we advocate that users should

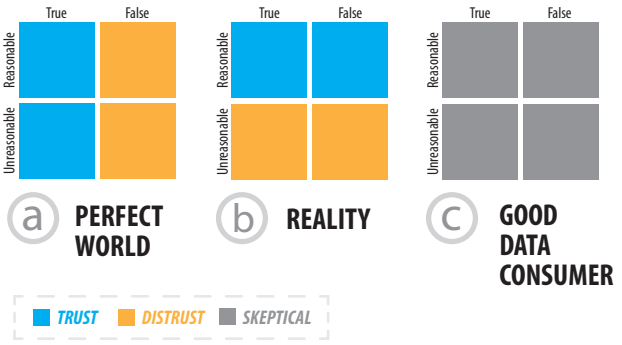

Figure 9: The relationship among trust, a claim's plausibility and the ground truth in different scenarios. In each bigger square, the y-axis is a claim's plausibility and the x-axis is the ground truth. In a perfect world (a), users should trust a claim only if it is true. In reality (b), users tend to trust a reasonable claim and distrust an unreasonable claim. When the truth is unknown, a good data consumer (c) should be skeptical despite a claims' plausibility.

be skeptical whenever they cannot assess the veracity of a causal claim (Fig. 9c): A good data consumer should question the validity of a reasonable claim because the causal relationship could be fake; she should not refute the possibility of an unreasonable claim since the claim could hold true.

6.1.1 Encouraging Skepticism for Reasonable Claims. How do we encourage users to be skeptical about reasonable causal claims through interface design? Warning could be a potential solution. In study 1, we observed that participants tended to be more cautious in agreeing with a reasonable claim given a simple warning message. However, a simple warning could be unreliable: When the system only provided reasonable causal claims in study 2, the warning did not seem to promote such caution. To improve the effectiveness of warning in inspiring skepticism, its design could be improved based on research in warning science. For example, Wogalter [67] proposed the Communication-Human Information Processing (C-HIP) Model to describe the perceptual and cognitive processes after people see a warning. The model suggests asking a series of questions to assess the effectiveness of warning messages. For instance, do people notice the warning? Is the message in the warning being effectively communicated?

What are other interface design ideas to help encourage skepticism? In study 1, we found that participants tended to disagree less with unreasonable claims and agree more with hard-to-tell claims given a scatterplot. This indicates that correlation depicted in a scatterplot could induce an illusion of causality. To mitigate this illusion, it seems plausible to hide scatterplots from causal claims. Nevertheless, some participants felt that they trusted the system more because the scatterplots enabled them to see the data. Ideally, designers should keep the benefits of scatterplots while mitigating their side effects. Ritchie et al. [54] found that transitioning from a non-deceptive view to a deceptive one could reduce the deception caused by the second view while enabling users to access the benefits of first view. Following this idea, a system could hide scatterplots by default while providing an option for users to view them. It would be interesting to investigate whether this design could reduce users' tendency to confuse correlation and causation.

How do we improve the general awareness of "correlation is not causation"? Besides inspiring skepticism about a reasonable claim, warnings also appeared to raise awareness of "correlation is not causation" in study 1. Again, a simple warning alone could be ineffective: Even when participants were warned that correlation is not causation, they tended to agree that the scatterplot implied causation. This suggests that system developers might need to look beyond interface design to help users acquire correct statistical knowledge. Alternatives include pedagogical approaches such as tutorials. For example, when Tableau introduced Explain Data, they emphasized that users were the data experts, and they should judge the veracity of the causal claims based on their knowledge [62, 63]. Future work will study the effectiveness of such tutorials in reducing users' tendency to associate correlation with causation.

6.1.2 Encouraging Open-Mindedness for Unreasonable Claims. While this work sheds light on ways to inspire skepticism for reasonable claims, designs to keep users open-minded when they see unreasonable claims are yet to be explored. Open-mindedness is a different form of skepticism: Instead of being skeptical about the automatically generated causal claims, users are skeptical about their beliefs and expectations about the data. Different models (e.g., Bayesian statistics [5] and the data-frame model [33, 34]) have been developed to explain the process through which people update their beliefs. Prior research in misinformation showed that existing beliefs are rigid, and people are inclined to resist changes to their beliefs [42].

Although encouraging open-mindedness could be challenging, what are some potential ideas to keep people open-minded when they see unreasonable causal claims? An idea is to enable users to tell the system if an answer makes sense. If users consider an answer questionable, the system could explain why a causal relationship might exist to prevent users from prematurely rejecting a causal claim that seems unreasonable. However, further evidence would be required to demonstrate the effectiveness of this approach.

6.2 Study Limitations and Future Work

Our results hint at the potential for scatterplots to create an illusion of causality and the potential for a simple warning to reduce this illusion. We note that these are observations under controlled experiments, and we are prudent in drawing conclusions about the practical significance of the findings. First, to collect data from hundreds of people on MTurk, we needed to sacrifice realism to adapt the study for an online setting. For example, participants examined a series of answers provided by the system rather than really interact with a working system. Second, collaboration could protect users from being misled in practice: While an analyst might draw causal conclusions from correlational evidence, colleagues might remind the analyst of the flaw. Learning about the practical implications of our findings will require observing how people employ systems such as Explain Data [64] in their workflow and studying how people collaborate during data analysis.

We also note that the effectiveness of a simple warning in reducing causal illusion warrants further studies. Our work only

compared four experimental conditions (Fig. 4). Evidence from further comparisons (e.g., a comparison between an additional claim + vis + warning condition and the original claim + vis condition) could support our findings about the effectiveness of the warning. An ideal experiment is to consider each information type (claim, vis, textual description, and warning) an independent variable with two levels (with and without). This experiment will enable comparisons among all possible experimental conditions. Nevertheless, adding more conditions greatly reduces power given Bonferroni correction, and interesting findings might be missed. In future studies, experimenters would likely want to preserve power by honing in on a smaller set of comparisons. Our findings could provide guidance on what focused comparisons to make.

Our target population was potential end users of question-answering systems. These users include both people who are less proficient in data analysis and those who are more proficient. Our participants ranged from beginner users to more advanced analysts and appeared to be a reasonable proxy for our target. Yet, the focus on these users also implies that some findings (e.g., the tendency among participants to confuse correlation and causation) may not generalize if we conduct the studies with professional analysts only. Future work will replicate our study with these experts.

In both studies, we used a single question to measure trust in the question-answering system. In future, a questionnaire with multiple questions could be developed for assessing user trust in these systems. Such a questionnaire will measure sub-dimensions of trust (e.g., understanding) and enable researchers to learn about more fine-grained reasons for trusting a system (e.g., the system is trustworthy because users can easily understand the answers).

We have investigated whether showing correlational evidence could induce causal illusion. One form of correlational evidence we studied was textual description about correlation. We note that creating a description that completely eliminates casual perception could be challenging because people might easily mistake correlation for causation. Future research will investigate how different phrasing of correlation descriptions will affect casual perception.

Our study focused on numerical variables. Visualizing correlation between numerical variables using scatterplots is common. For other data types (e.g., categorical variables), other charts are often used. A natural extension to our work is to investigate the generalizability of our findings to other data types and charts.

Finally, although we are advocates for encouraging skepticism when using question-answering systems, we note that inspiring the right level of skepticism could be challenging. Ideally, users' skepticism about a causal claim should match the evidence they have about the claim: For causal claims with good support (e.g., carbon dioxide emission leads to global warming), users could be less skeptical; for claims that lack supportive evidence, users could evaluate them more critically. However, it is difficult for a system to infer the amount of evidence users have about a claim and encourage skepticism accordingly. Moving forward, researchers could investigate whether telling users to be skeptical (e.g., through warnings) promotes an appropriate level of skepticism or engenders excessive and unhealthy skepticism.

7 CONCLUSION

Our work is situated in the discourse about the the ethical implications of data visualization [12]. We highlighted another scenario where visualizations might mislead users—question-answering systems could visualize correlation to create an illusion of causality. In particular, we found that in a system that occasionally provided unreasonable answers, showing a scatterplot next to a causal claim increased the plausibility of unreasonable and hard-to-tell claims. However, providing a simple warning about "correlation is not causation" seemed to lead participants to accept reasonable claims more cautiously. We further observed that our participants had a tendency to associate correlation with causation, but the warning appeared to reduce the tendency. We did not observe these effects of warning in a system that only provided reasonable answers. Based on the findings, we advocate that system developers could encourage users to be skeptical about answers generated by question-answering systems and have proposed ideas for doing so.

REFERENCES

[1] Outlier AI. 2017. *How to Conduct a Proper Root Cause Analysis*. https://towardsdatascience.com/how-to-conduct-a-proper-root-cause-analysis-789b9847f84b

[2] Robert Amar, James Eagan, and John Stasko. 2005. Low-Level Components of Analytic Activity in Information Visualization. In *IEEE Symposium on Information Visualization*. IEEE, 111–117. https://doi.org/10.1109/INFVIS.2005.1532136

[3] Apple. 2020. Siri - Apple. https://www.apple.com/siri/

[4] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3290605.3300484

[5] William M Bolstad and James M Curran. 2016. *Introduction to Bayesian Statistics*. John Wiley & Sons.

[6] Michelle A Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2015. Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2015), 519–528. https://doi.org/10.1109/TVCG.2015.2467732

[7] Jeremy Boy, Anshul Vikram Pandey, John Emerson, Margaret Satterthwaite, Oded Nov, and Enrico Bertini. 2017. Showing People Behind Data: Does Anthropomorphizing Visualizations Elicit More Empathy for Human Rights Data?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5462–5474. https://doi.org/10.1145/3025453.3025512

[8] United States Census Bureau. 2020. Census Bureau. https://www.census.gov

[9] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms Through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3290605.3300789

[10] William S Cleveland, Persi Diaconis, and Robert McGill. 1982. Variables on Scatterplots Look More Highly Correlated When the Scales Are Increased. *Science* 216, 4550 (1982), 1138–1141. https://doi.org/10.1126/science.216.4550.1138

[11] Wharton County Junior College. 2020. Root Cause Analysis Training. https://www.wcjc.edu/Programs/continuing-education/root-cause.aspx

[12] Michael Correll. 2019. Ethical Dimensions of Visualization Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. https://doi.org/10.1145/3290605.3300418

[13] Michael Correll, Enrico Bertini, and Steven Franconeri. 2020. Truncating the Y-Axis: Threat or Menace?. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3313831.3376222

[14] Michael Correll and Jeffrey Heer. 2017. Black Hat Visualization. In *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe)*. https://decisive-workshop.dbvis.de/wp-content/uploads/2017/09/0115-paper.pdf

[15] Coursera. 2020. Root Cause Analysis - Root Cause Analysis | Coursera. https://www.coursera.org/lecture/six-sigma-analyze/root-cause-analysis-w01Qj

[16] Kedar Dhamdhere, Kevin S McCurley, Ralfi Nahmias, Mukund Sundararajan, and Qiqi Yan. 2017. Analyza: Exploring Data With Conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 493–504. https://doi.org/10.1145/3025171.3025227

[17] Evanthia Dimara, Gilles Bailly, Anastasia Bezerianos, and Steven Franconeri. 2018. Mitigating the Attraction Effect With Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 850–860. https://doi.org/10.1109/TVCG.2018.2865233

[18] Evanthia Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic. 2020. A task-based taxonomy of cognitive biases for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 26, 2 (2020), 1413–1432. https://doi.org/10.1109/TVCG.2018.2872577

[19] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of the 2008 CHI Conference on Human Factors in Computing Systems*. ACM, 1065–1074. https://doi.org/10.1145/1357054.1357219

[20] Ethan Fast, Binbin Chen, Julia Mendelsohn, Jonathan Bassen, and Michael S Bernstein. 2018. Iris: A Conversational Agent for Complex Tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. https://doi.org/10.1145/3173574.3174047

[21] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the New World of HCI. *Interactions* 24, 4 (2017), 38–42. https://doi.org/10.1145/3085558

[22] National Center for Education Statistics. 2020. National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education. https://nces.ed.gov

[23] Kaiser Family Foundation. 2020. KFF - Health Policy Analysis, Polling and Journalism. https://www.kff.org

[24] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G Karahalios. 2015. Datatone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 489–500. https://doi.org/10.1145/2807442.2807478

[25] Marti Hearst and Melanie Tory. 2019. Would You Like a Chart With That? Incorporating Visualizations Into Conversational Interfaces. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 1–5. https://doi.org/10.1109/VISUAL.2019.8933668

[26] Marti Hearst, Melanie Tory, and Vidya Setlur. 2019. Toward Interface Defaults for Vague Modifiers in Natural Language Interfaces for Visual Analysis. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 21–25. https://doi.org/10.1109/VISUAL.2019.8933569

[27] Jeffrey Heer and Maneesh Agrawala. 2006. Multi-Scale Banking to 45 Degrees. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (2006), 701–708. https://doi.org/10.1109/TVCG.2006.163

[28] Elizabeth Hellier, Daniel B Wright, Judy Edworthy, and Stephen Newstead. 2000. On the Stability of the Arousal Strength of Warning Signal Words. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 14, 6 (2000), 577–592. https://doi.org/10.1002/1099-0720(200011/12)14:6<577::AID-ACP682>3.0.CO;2-A

[29] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*. ACM, 241–250. https://doi.org/10.1145/358916.358995

[30] Enamul Hoque, Vidya Setlur, Melanie Tory, and Isaac Dykeman. 2017. Applying Pragmatics Principles for Interaction With Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 309–318. https://doi.org/10.1109/TVCG.2017.2744684

[31] Jessica Hullman and Nick Diakopoulos. 2011. Visualization Rhetoric: Framing Effects in Narrative Visualization. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2231–2240. https://doi.org/10.1109/TVCG.2011.255

[32] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Data Through Others' Eyes: The Impact of Visualizing Others' Expectations on Visualization Interpretation. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 760–769. https://doi.org/10.1109/TVCG.2017.2745240

[33] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems* 21, 4 (2006), 70–73. https://doi.org/10.1109/MIS.2006.75

[34] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems* 21, 5 (2006), 88–92. https://doi.org/10.1109/MIS.2006.100

[35] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2018. Frames and Slants in Titles of Visualizations on Controversial Topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3173574.3174012

[36] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2019. Trust and Recall of Information Across Varying Degrees of Title-Visualization Misalignment. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. https://doi.org/10.1145/3290605.3300576

[37] Po-Ming Law, Rahul C Basole, and Yanhong Wu. 2018. Duet: Helping Data Analysis Novices Conduct Pairwise Comparisons by Minimal Specification. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 427–437. https://doi.org/10.1109/TVCG.2018.2864526

[38] Po-Ming Law, Subhajit Das, and Rahul C Basole. 2019. Comparing Apples and Oranges: Taxonomy and Design of Pairwise Comparisons within Tabular Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3290605.3300409

[39] Po-Ming Law, Alex Endert, and John Stasko. 2020. Characterizing Automated Data Insights. *arXiv preprint arXiv:2008.13060* (2020). https://arxiv.org/abs/2008.13060

[40] Victoria Lee. 2020. *Why We Quarantine: A Data Driven Love Letter to You and the Loves of Your Life*. https://medium.com/swlh/why-we-quarantine-a-data-driven-love-letter-to-you-and-the-loves-of-your-life-c19de2bca87f

[41] Erich L Lehmann. 2006. *Nonparametrics: Statistical Methods Based on Ranks*. Springer-Verlag, New York, NY, USA.

[42] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest* 13, 3 (2012), 106–131. https://doi.org/10.1177/1529100612451018

[43] Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. 2016. What Can You Do? Studying Social-Agent Orientation and Agent Proactive Interactions With an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. ACM, 264–275. https://doi.org/10.1145/2901790.2901842

[44] Q Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N Sadat Shami, and Werner Geyer. 2018. All Work and No Play? Conversations With a Question-And-Answer Chatbot in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. https://doi.org/10.1145/3173574.3173577

[45] Zhicheng Liu and Jeffrey Heer. 2014. The Effects of Interactive Latency on Exploratory Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2122–2131. https://doi.org/10.1109/TVCG.2014.2346452

[46] m00nlight Wang. 2018. *Income Inequality Analysis and Visualization*. https://medium.com/@m00nlight/income-inequality-analysis-and-visualization-f688a4fc6609

[47] Helena Matute, Fernando Blanco, Ion Yarritu, Marcos Díaz-Lago, Miguel A Vadillo, and Itxaso Barberia. 2015. Illusions of Causality: How They Bias Our Everyday Thinking and How They Could Be Reduced. *Frontiers in Psychology* 6 (2015), 888. https://doi.org/10.3389/fpsyg.2015.00888

[48] Luana Micallef, Gregorio Palmas, Antti Oulasvirta, and Tino Weinkauf. 2017. Towards Perceptual Optimization of the Visual Design of Scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 23, 6 (2017), 1588–1599. https://doi.org/10.1109/TVCG.2017.2674978

[49] Andrew Vande Moere, Martin Tomitsch, Christoph Wimmer, Boesch Christoph, and Thomas Grechenig. 2012. Evaluating the Effect of Style in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2739–2748. https://doi.org/10.1109/TVCG.2012.221

[50] Makeover Monday. 2020. Makeover Monday | a Weekly Social Data Project. https://www.makeovermonday.co.uk

[51] Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. 2014. The Persuasive Power of Data Visualization. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2211–2220. https://doi.org/10.1109/TVCG.2014.2346419

[52] Anshul Vikram Pandey, Katharina Rall, Margaret L Satterthwaite, Oded Nov, and Enrico Bertini. 2015. How Deceptive Are Deceptive Visualizations? an Empirical Analysis of Common Distortion Techniques. In *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*. ACM, 1469–1478. https://doi.org/10.1145/2702123.2702608

[53] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender System. In *Proceedings of the Fifth ACM Conference on Recommender Systems*. ACM, 157–164. https://doi.org/10.1145/2043932.2043962

[54] Jacob Ritchie, Daniel Wigdor, and Fann Chevalier. 2019. A Lie Reveals the Truth: Quasimodes for Task-Aligned Data Presentation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. https://doi.org/10.1145/3290605.3300423

[55] Julia M Rohrer. 2018. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science* 1, 1 (2018), 27–42. https://doi.org/10.1177/2515245917745629

[56] Dominik Sacha, Andreas Stoffel, Florian Stoffel, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. 2014. Knowledge Generation Model for Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1604–1613. https://doi.org/10.1109/TVCG.2014.2346481

[57] Alper Sarikaya and Michael Gleicher. 2017. Scatterplots: Tasks, Data, and Designs. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 402–412. https://doi.org/10.1109/TVCG.2017.2744184

[58] Vidya Setlur, Sarah E Battersby, Melanie Tory, Rich Gossweiler, and Angel X Chang. 2016. Eviza: A Natural Language Interface for Visual Analysis. In *Proceedings of the 29th Annual ACM Symposium on User Interface Software & Technology*. ACM, 365–377. https://doi.org/10.1145/2984511.2984588

[59] Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *Extended Abstracts of the 2002 CHI Conference on Human Factors in Computing Systems*. ACM, 830–831. https://doi.org/10.1145/506443.506619

[60] Arjun Srinivasan, Steven M Drucker, Alex Endert, and John Stasko. 2018. Augmenting Visualizations With Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2018), 672–681. https://doi.org/10.1109/TVCG.2018.2865145

[61] Arjun Srinivasan and John Stasko. 2017. Orko: Facilitating Multimodal Interaction for Visual Exploration and Analysis of Networks. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 511–521. https://doi.org/10.1109/TVCG.2017.2745219

[62] Tableau. 2019. Explain Data Internals: Automated Bayesian Modeling | Tableau Conference 2019. https://tc19.tableau.com/learn/sessions/explain-data-internals-automated-bayesian-modeling?_ga=2.242994050.1845292459.1583776901-580893601.1583776901&_fsi=H59ZIxRV

[63] Tableau. 2019. Inspect a View using Explain Data – Tableau. https://help.tableau.com/current/pro/desktop/en-us/explain_data.htm

[64] Tableau. 2020. Explain Data | Tableau Software. https://www.tableau.com/products/new-features/explain-data

[65] Andre Calero Valdez, Martina Ziefle, and Michael Sedlmair. 2017. Priming and Anchoring Effects in Visualization. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2017), 584–594. https://doi.org/10.1109/TVCG.2017.2744138

[66] Jun Wang and Klaus Mueller. 2015. The Visual Causality Analyst: An Interactive Interface for Causal Reasoning. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2015), 230–239. https://doi.org/10.1109/TVCG.2015.2467931

[67] Michael S Wogalter. 2006. Communication-Human Information Processing (C-HIP) Model. *Handbook of warnings* (2006), 51–61.

[68] Cindy Xiong, Joel Shapiro, Jessica Hullman, and Steven Franconeri. 2019. Illusion of Causality in Visualized Data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 853–862. https://doi.org/10.1109/TVCG.2019.2934399

[69] Cindy Xiong, Lisanne van Weelden, and Steven Franconeri. 2019. The Curse of Knowledge in Visual Data Communication. *IEEE Transactions on Visualization and Computer Graphics* (2019). https://doi.org/10.1109/TVCG.2019.2917689

[70] Bowen Yu and Cláudio T Silva. 2019. Flowsense: A Natural Language Interface for Visual Data Exploration Within a Dataflow System. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 1–11. https://doi.org/10.1109/TVCG.2019.2934668

[71] Mingqian Zhao, Huamin Qu, and Michael Sedlmair. 2019. Neighborhood Perception in Bar Charts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. https://doi.org/10.1145/3290605.3300462

[72] Qiyu Zhi and Ronald Metoyer. 2020. GameBot: A Visualization-Augmented Chatbot for Sports Game. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–7. https://doi.org/10.1145/3334480.3382794