Eurographics Conference on Visualization (EuroVis) 2024
W. Aigner, D. Archambault, and R. Bujack
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 3

Guided By AI: Navigating Trust, Bias, and Data Exploration in AI-Guided Visual Analytics

Sunwoo Ha[1], Shayan Monadjemi[2], Alvitta Ottley[1]

[1] Washington University in St. Louis, St. Louis, MO
[2] Oak Ridge National Laboratory, Oak Ridge, TN

Abstract

The increasing integration of artificial intelligence (AI) in visual analytics (VA) tools raises vital questions about the behavior of users, their trust, and the potential of induced biases when provided with guidance during data exploration. We present an experiment where participants engaged in a visual data exploration task while receiving intelligent suggestions supplemented with four different transparency levels. We also modulated the difficulty of the task (easy or hard) to simulate a more tedious scenario for the analyst. Our results indicate that participants were more inclined to accept suggestions when completing a more difficult task despite the AI's lower suggestion accuracy. Moreover, the levels of transparency tested in this study did not significantly affect suggestion usage or subjective trust ratings of the participants. Additionally, we observed that participants who utilized suggestions throughout the task explored a greater quantity and diversity of data points. We discuss these findings and the implications of this research for improving the design and effectiveness of AI-guided VA tools.

CCS Concepts

• ***Human-centered computing*** → ***Visual analytics;*** *Empirical studies in visualization;*

1. Introduction

The visual analytics (VA) community is increasingly interested in using artificial intelligence (AI) algorithms to assist users in data exploration and decision-making. As these mixed-initiative systems become more prevalent, it is critical to understand the users' behaviors when interacting with the algorithm and to what extent they allow the algorithm to influence their data exploration and analytical decisions. Understanding the underlying factors that influence the user's interactions is essential to ensure that the AI guidance is used effectively and will maximize the cost-benefit ratio of creating such tools. However, recent studies show two extreme behaviors of interaction within human-AI decision-making. Some users are unable or unwilling to trust an algorithm's guidance, leading to *underutilization* [MHN*22]. On the other hand, some may be heavily influenced by the AI and blindly accept incorrect suggestions without verifying whether it was correct, commonly referred to as *overreliance* [BMG21].

Developing methods to mitigate these behaviors and allow users to calibrate an appropriate reliance on the AI during decision-making is an ongoing challenge. One option explored by existing work in the XAI community is to provide explanations of the AI suggestions to promote trust calibration. These explanations can be as simple as displaying the AI confidence in a certain prediction or more elaborate explanations that relate information such as feature importance [WY21]. Researchers posit that these explanatory techniques can create more effective, transparent, and trustworthy systems by explaining how the AI teammate arrived at its suggestions [RSG16]. As a result, users can better understand the reasoning behind these suggestions and feel more confident using them during their decision-making process [KDM*19].

Empirical evidence supports these assertions, with prior work showing that providing explanations of the AI suggestions tended to increase people's ability to calibrate trust in the algorithms [KDM*19, ZLB20]. Most of the tasks completed by participants in the prior studies utilized a pre-trained AI for decision-making for scenarios such as medical diagnosis [BSO15, GSR*21], recidivism [ZLB20], and income prediction [MLW*23]. Observing the effects of explanations on the user's trust, bias, and behaviors is underexplored in the context of AI-guided visual data exploration. The exploratory nature of AI-guided VA along with the already utilized visual elements to convey data, adds another layer of complexity to this open challenge. Displaying too much information can overwhelm users and lead to confusion, especially in data-rich settings, which calls the VA community for design guidelines on effectively showing and explaining suggestions whilst exploring.

Motivated by these challenges, we explore the impact of task difficulty and transparency on users' trust and data exploration with AI suggestions within a VA scenario. We conducted a 2×4 between-subject crowdsourced user study with 500 participants tasked with exploring a dataset using a mixed-initiative system. We randomly assigned participants to one of ten groups selected from two task difficulty conditions (easy and hard), four transparency conditions (no transparency, confidence, keyword, keyword + confidence), and two control conditions (no AI) for each task difficulty group.

To mimic an ecologically valid analytic task in the national security domain, participants explored a map visualization containing geo-tagged social media posts to identify people who may be affected by an epidemic. They interacted with the dataset for up to ten minutes by tagging individuals of interest, while also being mindful of the different types of symptoms and affected regions of the city. The VA system for our study utilized an active search algorithm introduced by Monadjemi et al. [MHN*22], which learns the most influential keyword of interest based on the user's data exploration interaction logs and recommends data points that will potentially match their analytic goals.

We find that when completing a more difficult task, the users were more inclined to accept the AI suggestions despite the AI having lower suggestion accuracy. Also, we demonstrate that the level of transparency had no measurable impact on suggestion usage and trust. Moreover, the participants who received no transparency exhibited similar performance measures and subjective trust levels as those who were provided with model transparency. Overall, we found that participants tended to trust AI suggestions regardless of the transparency level. Furthermore, participants overrelied on the suggestions in a more difficult task setting. Lastly, we observed that the participants who utilized the suggestions explored a greater quantity and diversity of data points. We conclude by discussing the implications of our results, highlighting some of the promises and challenges of promoting transparency in difficult task settings along with AI suggestions in a guided visual data discovery framework.

A summary of our contributions is as follows:

- Despite long-standing beliefs about trust and transparency in the AI community, we show that additional transparency to a VA system may not always affect data exploration and suggestion usage. We discuss future directions for transparency methods and the unique consideration for VA systems.
- We demonstrate that users' baseline trust in AI-guided VA may be high, and adding information to promote transparency may have a marginal impact. Still, our findings provide weak confirmation for suggestion usage as a proxy for real-time trust in future AI-assisted VA evaluations as we observe that higher AI reliance was associated with high perceived trust.
- We show that task difficulty was strongly associated with suggestion usage and overreliance. Our findings suggest that designers of VA systems should consider ways to adapt the level of guidance provided by the AI based on the task's difficulty.

2. Related Work

Guidance in VA is a computer-assisted process to address users' knowledge gaps during interactive sessions [CGM*17]. Researchers have developed various systems that utilize AI and ML algorithms [XOW*20] to support user interactions and the discovery of new insights during data exploration [KJO*19, LGG*18, MHN*22]. We contribute to the body of work on guidance in VA by examining the relationship between trust, suggestion usage, data exploration, and transparency levels.

2.1. Evaluating Guidance in VA

Evaluating guidance in VA systems typically involves calculating and comparing metrics such as task accuracy and speed to determine the effectiveness of assisting the user [MHN*22, BLBC12, BCS16]. While these metrics are crucial, observing how users interact with the AI guidance is equally important to improve the design of future AI-guided VA systems.

Some existing work has sought to look beyond speed and accuracy [DC17, LST*21, DLW*17, MHN*22]. For example, work by Dabek et al. [DC17] examined *suggestion usage*. They found that users ultimately performed the action suggested to them 20% of the time in their system's evaluation. Their users said the suggestions were useful but not always necessary to solve the task. Lee et al. [LST*21] also investigated usage with *Frontier*, a system recommending new ways to visualize a given dataset. They observed that users followed suggestions while exploring unfamiliar data attributes or when they did not know what to explore next. Furthermore, they argue that the interpretability of the suggestions positively impacted usage. Dasgupta et al. [DLW*17] examined *trust perception*. They presented a comparative study of domain scientists' trust level in their visual analytics system, Active Data Biology. They argue that domain scientists trust intuitive and transparent systems that allow seamless switching between hypothesis generation and evidence gathering.

2.2. AI Underutilization and Overreliance

Skepticism and low self-reported trust in AI have led users to ignore the AI during the decision-making process and develop algorithmic aversion [MHN*22, DSM15, KYZ23], which is defined as the tendency for users to discount suggestions from AI more heavily than human suggestions. This negative attitude towards AI leads to *underutilization* of intelligent suggestions. Researchers have also observed this behavior when users had high-domain expertise [DC17]. Closely related to domain expertise is task familiarity. Interestingly, another study showed that users with high task familiarity reported more trust in the AI teammate but showed less adherence to its suggestions [SOM*19].

On the opposite extreme of underutilization is *overreliance*. The literature on AI-assisted decision-making has also established that humans can be easily influenced by the AI teammate and often accept incorrect suggestions without verifying whether the AI was actually correct [BMG21]. Jacobs et al. [JPM*21] found that users with low AI literacy were significantly more likely to select medical treatments that were aligned with the AI suggestions. These interaction behaviors depend on a lot of different factors, however, the most popular method used to mitigate these behaviors in existing works is to provide explanations of the AI suggestions [VJGM*22, ZLB20, WY21, NKR20]. The intuition runs that if users see an incorrect explanation, they will more carefully scrutinize the AI suggestion and build an appropriate reliance.

2.3. AI Transparency and Explanations

Explainable AI (XAI) are techniques that enable humans to understand, trust, and manage AI teammates effectively [ADRDS*20]. For example, Cheng et al. [CWZ*19] show that presenting the inner workings of a university admissions algorithm with an interactive interface can enhance users' understanding of the algorithm. However, it is unclear what criteria constitute a sufficient explanation. Researchers have used *examples* [CJH19, YHSA20] and *coun-*

terfactual examples [WY21]. Wang et al. [WY21] explored other explanation methods such as *feature importance*, *feature contribution*, and *nearest neighbors* for recidivism prediction and forest cover prediction tasks. Their study found supportive evidence suggesting that providing information about feature contribution allowed participants to have an awareness of uncertainty within the model and appropriately calibrate their trust.

Transparency and explanations are often interconnected. A common approach to promote transparency is showing *uncertainties* [HFRD13, ZLB20, BSO15] of the AI. For instance, Zhang et al. [ZLB20] studied the effect of showing *confidence score* and local explanation for predictions in an income prediction task. Their findings show that prediction-specific confidence information could support trust calibration. Similarly, Dietvorst et al. [DSM18] communicated a *model's uncertainty* through the outright disclosure of the model's average error rate by a text description (i.e., "the model has an average error rate of x"). Linder et al. [LMY*21] explored how the type and amount of explanations affect users' understanding and performance on a fact-checking task. More detailed explanations such as providing examples of alternative statements with the same classification and information about the influence of the statement's metadata led the users to a better understanding of the AI suggestions. However, this was coupled with lower performance due to the additional time and attention required.

2.4. Manipulating Model Uncertainty and Task Difficulty

High levels of AI prediction uncertainty may indicate an elevated likelihood of poor performance. Sacha et al. [SSK*16] discussed the role of uncertainty, awareness, and trust in VA. They argue that the users' trust in the AI teammate's outcomes is influenced by their awareness of the various kinds of uncertainty that exist or are generated in the system. To this end, prior works have manipulated task difficulty or model uncertainty whilst providing explanations to observe their respective effects on the human's interactions with the AI. However, the results are inconclusive.

For example, Vasconcelos et al. [VJGM*22] manipulated the task's difficulty with a visual search maze-solving task by changing the maze's dimensions such that the harder task involved solving a higher-dimensional maze. They show that explanations of the AI prediction reduced overreliance in the hard task condition. On the other hand, Zhao et al. [ZWM*23] explored the impact of uncertainty visualization on trust and reliance on model predictions using a college admissions forecasting task. Findings show that in low-uncertainty tasks, proper visualization of model uncertainty can enhance an appropriate adoption of model predictions. However, when a decision task had high model uncertainty, the uncertainty visualization did not significantly affect the participants' trust.

The overarching findings from some of the prior work support that local explanations and transparency methods may increase trust and suggestion usage. However, these techniques may not necessarily increase human-AI performance. For example, too much information about the AI could lead to worse performance or decision-making outcomes [LMY*21].

3. Research Goals

Building on prior work, this paper aims to understand how different levels of transparency provided by VA systems impact users' trust and acceptance of those suggestions. We are also interested in how the decision to utilize or not utilize these suggestions affects users' data exploration patterns. Further, we consider how task difficulty, *determined here by the percentage of irrelevant data in the underlying dataset*, might further affect the observed behaviors.

The interplay among trust, suggestion usage, transparency levels, and task difficulty carries significant implications for future VA tools seeking to harness AI technology to assist users in data exploration, making the work in this paper a compelling unresolved research inquiry. It is important to consider the prior work showing that the quality of the explanations provided by the AI plays a pivotal role [LT19, WY21] in the users' interactions with AI suggestions. When the AI provides explanations or displays information about its confidence, users are likelier to trust and rely upon AI suggestions [ZLB20]. Moreover, the intricate relationship between trust and suggestion usage appears to be influenced by task difficulty or model uncertainty [VJGM*22, ZWM*23]. We hypothesize that users may be more inclined to trust and utilize AI suggestions when confronted with difficult tasks where additional guidance in data exploration becomes more essential.

4. Methods

We aimed to create a realistic dataset, scenario, and tasks for our participants. Thus, we adopted the task and dataset from the 2011 Visual Analytics Science and Technology (VAST) Challenge. The VAST Challenge is a yearly competition and workshop supported by IEEE VIS and the Pacific Northwest National Laboratory and publishes datasets and analytic tasks that mimic real-world challenges [CGW*12, CGW14]. In this scenario, the fictional city of Vastopolis is facing a biochemical attack, leading to an epidemic that is spreading throughout the city. The dataset consists of 1,023,077 messages, similar to tweets, that were posted on social media from different parts of the city over 21 days. Additionally, there is a satellite image of the city that includes labeled highways, hospitals, landmarks, and water bodies.

4.1. Task

As part of the study, we informed the participants that hospitals in the city of Vastopolis had a significant rise in reported illnesses. The city authorities have recruited the participants to help identify the affected areas by analyzing social media activity. Their task was to sift through the dataset of social media posts using the interface in Figure 1 and tag posts containing illness-related content for further investigation. Based on their analysis, they then identified the areas of the city that they believed were most impacted by the epidemic. We chose this dataset and task for two main reasons. First, it mimics realistic scenarios where users search through a large space in search for relevant data points (e.g., material discovery, intelligence analysis) [JMMG18, GKX*12]. Secondly, the task does not require any domain expertise, which makes it appropriate for a large-scale crowd-sourced study targeted towards the general public.

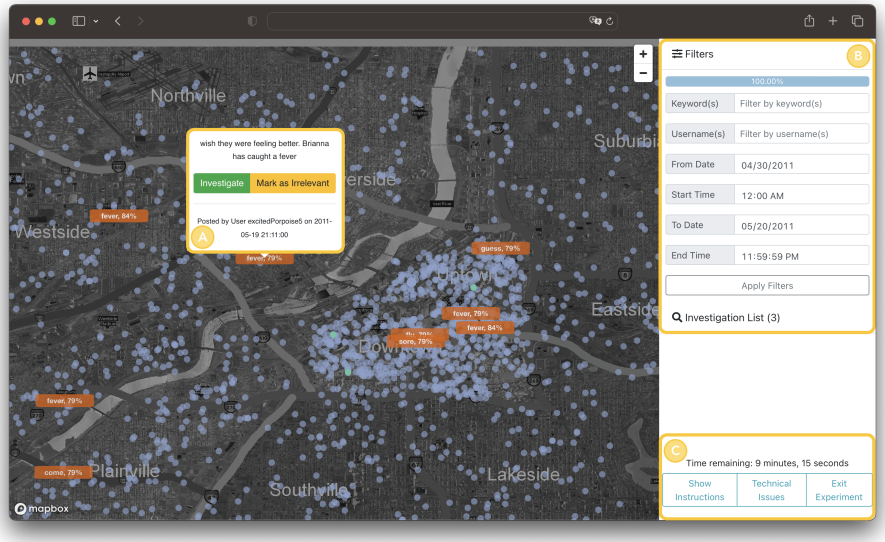

Figure 1: *The interface of the system used for the study. a) Hovering triggers a tooltip with the full text of the social media post. b) The sidebar allows users to search and filter. c) The remaining time for the task, and users could exit at any time or report technical issues.*

4.2. Design

We designed a 2 (task difficulty) × 4 (transparency level) between-subject study, resulting in a total of eight AI conditions. There were two control (no AI) conditions for each level of difficulty.

4.2.1. Task Difficulty

Controlling the difficulty of the task aids in understanding when interactions with the AI change and when guidance may be valuable but more error-prone. To control the level of difficulty of our task, we use the percentage of irrelevant points in the dataset as a proxy. The study had two levels of difficulty:

- **HARD**: The data was a random sample of 2000 points from the entire 21-day period, resulting in ∼9% illness-related posts.
- **EASY**: The data was a random sample of 2000 points from the approximate start date of the epidemic. Hence, adopting the assumption that the starting point of the epidemic is known. This resulted in a dataset with ∼36% illness-related posts.

4.2.2. Transparency Levels

With limited screen real estate in VA and the added stress of solving a task within a limited time, we need to be conservative in how we promote situational awareness of the AI and how much information is shown to the users, especially when the AI provides multiple suggestions. Although the space of explanation techniques is vast, we opted for variations of simple prediction-specific transparency methods for users to understand – the most influential keyword and confidence value. Inspired by prior studies [ZLB20, LMY*21, WY21], we posit that providing the most influential keyword and confidence value may improve trust calibration by giving users an indication to increase their situational awareness of the AI's performance. We define an influential keyword to be one which if eliminated, decreases the probability of a post being relevant by the largest amount. This method of transparency will allow users to quickly understand the context of a certain suggestion, while confidence values can provide users with information about the system's level of certainty regarding the relevance of a particular suggestion. In addition to a **control (CTRL | no AI)** condition, there were four transparency levels in our study:

- **No transparency (NONE)**: Suggestions are presented as dots.
- **Confidence (CONF)**: Suggestions are presented as rectangles containing the AI's confidence percentage that the suggestion is relevant to the user. 87%
- **Keyword (KWD)**: Suggestions are presented as rectangles containing the most influential word for classifying the social media post as relevant to the user. flu
- **Keyword + confidence (KWD+CONF)**: Suggestions are presented as rectangles containing the most influential word and the AI's confidence percentage. flu, 87%

4.3. Visual Analytic System

We expanded upon the prototype presented in [MHN*22] and prioritized the following design features for the VA tool used in this study: (1) a map for identifying regions that are most affected, (2) a search and filter functionality to aid the discovery of illness-related terms (3) the ability to inspect individual social media posts and triage people for contact tracing. See Figure 1 for the tool's interface. Users hovered over data points to trigger a tooltip with the social media post's details. They then could flag the post by clicking the *investigate button*. If the post was previously flagged, there is an option to remove the flag and report an *irrelevant suggestion* from the AI. We utilized three distinct colorblind-safe colors to distinguish among **suggested points**, **points selected for investigation**, and the **remaining points**.

4.4. Modeling and Guidance Engine

Our visual analytic system observed user interactions, inferred their data interest, and made suggestions. To fully specify our guidance engine, we need to describe (1) the classification model which *predicted* the relevance of documents to the task at hand in light of past interactions, and (2) the algorithm which *decided* which documents to suggest given the classification model's belief.

4.4.1. Predicting Document Relevance

We begin with a finite set of n data points displayed on the interface, $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, and as we observe user interactions, we wish to infer if each point in $\mathcal{X}$ is relevant to the task at hand. We consider this dataset to be *unlabeled* initially, meaning that we do not know whether each point is relevant to the task at hand or not. As we observe user interactions with data points, we translate their interactions into labels and collect them in the observation set, $\mathcal{D} = \{(x_1, \hat{y}_1), ..., (x_t, \hat{y}_t)\}$, where $\hat{y}_i = 1$ indicates that x_i is deemed *relevant* and $\hat{y}_i = 0$ indicates that x_i is deemed *irrelevant* based on user interactions. We want to then reason about the relevance of the remaining points to the task at hand, $\Pr(\hat{y}_i = 1 \mid x_i, \mathcal{D})$, where $x_i \in \mathcal{X}$ is an arbitrary unlabeled data point. Specific to our visual analytic tool described above, we consider an *investigate* interaction to result in a $\hat{y}_i = 1$ label (i.e., relevant), whereas *irrelevant suggestion* interactions result in a $\hat{y}_i = 0$ label (i.e., irrelevant).

Since our task and dataset from Section 4.1 involve interacting with social media posts, we wish to express $\Pr(\hat{y}_i \mid x_i, \mathcal{X})$ with support over text data. Therefore, in a pre-processing step, we transform the unstructured text data into a numerical space using a pre-trained *word2vec* auto-encoder [ŘS10]. After this transformation, each social media post is represented as a 300-dimensional numerical vector. We then compute the pairwise cosine distance between vectors to build a k-NN classification model for inferring the relevance of documents to the task at hand in light of past observations where $k = 180$. This model is initially untrained, meaning it does not have any labeled observations. As the user interacts with the data, we re-train this model with the updated observations. One compelling benefit of using a k-NN classifier in this setting is that it is quick to re-train in real-time.

Table 1: *Summary of notations used in Section 4.4*

Notation	Description
$\mathcal{X}$	Set of n data points displayed on the interface.
$\mathcal{D}$	Set of data points and their relevance labels observed through user interactions with the interface.
$\mathcal{S}_t$	Set of data points suggested to the user at time t.
$\hat{y}_i$	Label for data point x_i observed from interactions.
y_i	Ground-truth label for data point x_i given the task.

4.4.2. Making Suggestions for Investigation

We used an active search algorithm to suggest points to the user. Active search is an active learning algorithm that makes queries to maximize the number of relevant data points discovered under a limited querying budget [JMC*17]. In mathematical notation, the utility of $\mathcal{D}$ at the end of the session is defined as:

$$u(\mathcal{D}) = \sum_{j=1}^{|\mathcal{D}|} \hat{y}_j,$$

where the active search algorithm aims to make queries to approximately maximize this utility. Active search has proven effective in accelerating drug discovery [JMMG18] and visual data foraging [MHN*22]. We utilized a greedy active search algorithm, assuming each set of queries is the final batch. We refer the readers to Jiang et al. [JMMG18] for details on search horizon and exploration/exploitation trade-off in active search algorithms.

4.5. Participants

We used G*Power to conduct an a priori power analysis to estimate the sample size for this study. Our effect size of .6 for an ANOVA was based on studies by Monadjemi et al. [MHN*22]. The result from the a priori analysis indicated that we needed a sample size of $N = 450$ to achieve 80% power for detecting a medium effect at a significance level of $\alpha = .05$. Thus, we used a sample size of $N = 500$ to account for exclusions. We provide additional information about the study in the preregistration†.

We recruited our participants through Prolific [PS18] per Washington University's IRB guidelines. Participants were 18 to 65 years old, from the United States, and fluent in English. For more detailed participant demographics, please refer to Table 2. We provided a base pay rate of $15.00 per hour and the participants' median time to complete the study was around 15 minutes (including the tutorial, the task, and the survey).

4.6. Experimental Procedure

Our system randomly assigned each participant to one of the eight conditions. Upon giving consent to participate in our study, participants saw a tutorial demonstrating how to interact with the system. The tutorial also provided explanations as well as examples of the transparency methods tested in this study to all participants.

Each participant then played the role of a triage investigator and needed to find social media posts that contained information about the hot spot locations of the epidemic and the symptoms being reported. Participants could either use the search feature or browse the data points via the interactive map shown in Figure 1. Hovering over a data point triggered a tooltip containing the post's full text, and clicking on the *investigate* button within the tooltip tagged the post as containing illness-related information. Once the AI observed the participant's first three tagged interactions, it suggested 10 relevant posts to explore. These suggestions were in the form of visual cues that were updated after every investigation. The visual cues depended on the type of transparency condition they were assigned as detailed in Section 4.2.2.

After spending up to 10 minutes tagging posts related to the epidemic, participants answered some of the VAST Challenge questions, which included their beliefs on whether the epidemic was

† This experiment was pre-registered on Open Science Foundation.

Table 2: *Summary statistics and self-reported demographics of participants.*

	Easy Task	Hard Task
Recruited participants	236	236
Excluded participants	44	36
Included participants	192	200
Sex	91 Male, 95 Female, 6 Undisclosed	94 Male, 101 Female, 5 Undisclosed
Age	$\mu = 34.7, \sigma = 12$	$\mu = 32.7, \sigma = 11.1$
Education	68% with at least an associate degree	60% with at least an associate degree

contained, how it was transmitted, and the areas that were most affected. Finally, they completed the exit survey which included demographic and system usability questions, in addition to attention checks and questions about their level of trust towards the AI.

5. Data Collection, Exclusions, and Validation

There were two exclusion criteria listed on our preregistration for this study. First, we initially excluded participants who had technical issues and could not complete the task ($n = 28$), then we excluded the sessions of those who failed to pass the attention checks in the post-experiment survey or had less than 10 hovers on the data points in their interaction log ($n = 64$). In the end, we had a total of 408 participants' sessions for data analysis.

5.1. Manual Coding and Ground Truth Proxy

In addition to measuring trust and suggestion usage, we considered accuracy to enable comparisons with prior work. However, the 2011 VAST Challenge dataset did not include labels indicating posts that were illness-related. We began by compiling a list of all the tagged social media posts from the study participants, producing 1642 unique posts. One author manually coded each post using binary labels of **illness-related** or **non-illness-related**. This process also produced a collection of word stems, allowing us to programmatically label the full dataset used in the study. We evaluated the labels by iteratively selecting random samples of the dataset, correcting labeling errors, and updating the list of word stems. For example, the following post in the dataset was initially not flagged as illness-related but after the manual coding process, the label was corrected as "having the sweats" is a potential symptom of having a fever: *"Nicholas has caught a the **sweats** I hate this."*

These ground truth proxy labels enable us to estimate measures such as the AI and participants' accuracy. This process also uncovered a few participants who did not complete the task as instructed. Specifically, for 16 participants, none of the social media posts they tagged for contact tracing included illness-related information. Thus, we deviated from our preregistration exclusions and excluded these 16 participants from our analysis, resulting in 392 remaining participants. Table 2 summarizes the self-reported demographic information of participants included in the study.

5.2. Data Collection

To measure the impact of task difficulty and transparency level on AI interaction, we calculated the following dependent variables:

- **Suggestion usage** $\in [0...1]$, the ratio of AI suggestion batches that resulted in an interaction (*investigate* or *mark as irrelevant*). USAGE is the quotient of the number of accepted suggestions divided by the number of AI suggestion batches received.

$$\text{USAGE} = \frac{1}{\mid \mathcal{D} \mid -3} \sum_{t=4}^{\mid \mathcal{D} \mid} \mathbb{1}_{\mathcal{S}_{t-1}}(x_t),$$

  where $\mathbb{1}$ denotes the indicator functions [‡].
- **AI accuracy** $\in [0...1]$, the ratio of relevant suggestions presented to the participant.

$$\text{AI ACCURACY} = \frac{1}{\mid \mathcal{D} \mid -3} \sum_{t=4}^{\mid \mathcal{D} \mid} \sum_{i=1}^{\mid \mathcal{S}_t \mid} \frac{y_i}{\mid \mathcal{S}_t \mid}$$

- **No. of positive investigations** is the number of illness-related social media posts tagged for investigation by the participant.

$$\text{INVESTIGATIONS} = \mid \{x_i \in \mathcal{D} \mid (y_i = 1) \wedge (\hat{y}_i = 1)\} \mid$$

- **Overreliance** $\in [0...1]$, the ratio of irrelevant suggestions investigated by the participant.

$$\text{OVERRELIANCE} = \frac{\mid \{x_t \in \mathcal{D} \mid (x_t \in \mathcal{S}_{t-1}) \wedge (\hat{y}_t = 1) \wedge (y_t = 0)\} \mid}{\mid \{x_t \in \mathcal{D} \mid (x_t \in \mathcal{S}_{t-1}) \wedge (\hat{y}_t = 1)\} \mid}$$

- **Symptom diversity** is the number of unique relevant symptoms discovered by the participant through investigations.

The covariate that was measured:

- **TRUST** $\in [1...5]$, self-reported trust in the algorithm collected in the post-experiment survey. Participants responded to the statement "*I trusted AVA throughout the investigation.*" on a 5-point Likert scale with 1 = Strongly disagree to 5 = Strongly agree.

6. Hypotheses

We tested the following hypotheses:

H1 Participants in HARD will use suggestions more frequently than those in EASY, regardless of the level of transparency. Aligning with prior work which suggests that participants have a tendency to overrely on AI with difficult tasks [VJGM*22], we hypothesize that participants in HARD may be more willing to rely on external guidance for direction throughout their analysis.

[‡] The indicator function states that $\mathbb{1}_{\mathcal{S}_{t-1}}(x_t) = 1$ if $x_t \in \mathcal{S}_{t-1}$, and $\mathbb{1}_{\mathcal{S}_{t-1}}(x_t) = 0$ otherwise. In our case, it tells us whether the data point with which the user interacted was suggested to them by the system.

H2 Participants who receive higher transparency (i.e., KWD+CONF) will have higher suggestion usage and subjective trust ratings [LMY*21, KDM*19] compared to those who receive lower transparency (i.e., NONE, CONF, and KWD), regardless of task difficulty.

H3 Participants in HARD may have a higher level of trust in the system than those in EASY. Building on **H1**, we hypothesize that participants in HARD may perceive the VA system as providing valuable guidance in a difficult task. In contrast, participants in EASY may perceive the system as unhelpful since the task is doable without additional guidance [DC17].

H4 Participants in CONF and KWD+CONF will have higher trust in the system than those in NONE and KWD, regardless of task difficulty. We posit that confidence values communicate the level of certainty the AI has and the presence of the most influential keyword further clarifies the reasoning behind the suggestions.

7. Results

Of the 392 participants, 192 were assigned to EASY, and 200 were assigned to HARD. We begin our analysis by establishing a baseline with our control (CTRL) conditions, i.e., participants who completed the task with no AI suggestions.

Is there evidence that the hard task was more difficult? Participants in EASY|CTRL ($n = 40$) had a median value of 70 people tagged for contract tracing (IQR $(40.5, 95.25)$), and those in HARD|CTRL ($n = 45$) had a median value of 50 (IQR $(27.0, 63.0)$) people tagged. An independent sample Mann-Whitney U test comparing the outputs of the two groups found a significant difference in the number of investigations ($U = 1209, p = .007, \eta^2 = .087$). *The data quality in* HARD *made the task more difficult to execute than for those in* EASY*, eliciting lower discovery rates.*

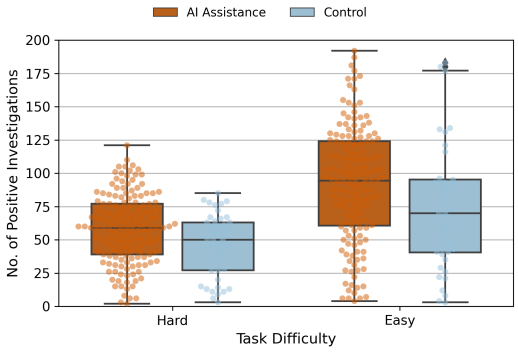

Figure 2: *The spread of positive investigations made for* CTRL *and* AI*-guided groups for both* EASY *and* HARD.

Do we replicate prior work showing that AI improves discovery rate? Based on prior findings [MHN*22], this work assumes that AI guidance would improve data discovery. We validate this assumption by comparing the INVESTIGATIONS averaged across all AI transparency conditions and the CTRL(no AI) condition. Figure 2 shows the spread of INVESTIGATIONS for the CTRL ($n = 85$) and AI-guided groups ($n = 307$) for both EASY and HARD. An independent samples Mann-Whitney U test comparing the outputs of the two groups found a significant difference between the number of people tagged for contact tracing ($U = 16790.5, p < .001, \eta^2 = .042$). *This finding replicates prior work [MHN*22] and reaffirms that* AI *guidance is associated with more efficient data exploration.*

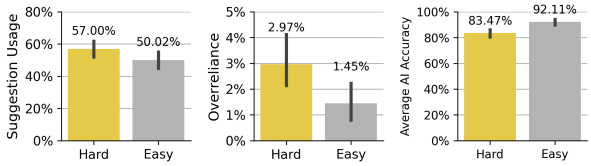

Figure 3: *Comparison of suggestion usage, overreliance, and average* AI *accuracy between task difficulty conditions. Participants in* HARD *utilized more suggestions despite having less accurate* AI *suggestions, which may have led participants to overrely.*

7.1. Is it more likely to use AI suggestions when completing a more difficult task?

With **H1**, we hypothesized that participants in HARD might feel more uncertain about their ability to perform the task and may be more willing to rely on external guidance. The first chart in Figure 3 compares the average USAGE between EASY (50.02%) and HARD (57%). An independent sample Mann-Whitney U test revealed a significant difference in the USAGE between the two groups ($U = 10178.5, p = .039, \eta^2 = .014$). Supporting **H1**, we reject the null hypothesis that there is no difference in the proportion of suggestions used between HARD and EASY. *Participants in* HARD*, were more likely to utilize* AI *suggestions during exploration. However, the effect is small.*

Is there evidence of overreliance? The finding that participants in HARD were more likely to use AI suggestion than EASY is particularly interesting when we consider the AI accuracy. As shown in Figure 3, we observed a significantly lower AI accuracy for HARD compared to EASY ($U = 16790.5, p < .001, \eta^2 = .135$). Overreliance on AI can lead to blindly accepting incorrect suggestions and potentially harming outcomes. Thus, we examine participants' reliance and whether the difficulty of the task may influence it. Overall, the rate of overreliance was low, with evidence suggesting that, on average, 2.97% and 1.45% of participants' accepted AI suggestions in the HARD and EASY groups were irrelevant to the task. An independent sample Mann-Whitney U test revealed a significant difference in the OVERRELIANCE distributions between EASY and HARD ($U = 9315.5, p < 0.0001, \eta^2 = .033$). This finding aligns with prior works that observed an *increase in overreliance on the* AI *when completing more difficult tasks [VJGM*22]*. The second chart in Figure 3 compares the average overreliance rates between EASY and HARD.

7.2. Is it more probable to use AI suggestions when there is more transparency?

Figure 4 shows the spread of USAGE among the participants in all eight AI-guided condition groups. Despite the median USAGE for KWD+CONF being higher than all the other conditions within HARD, we fail to reject our null hypothesis that there is no significant difference among the transparency levels for both EASY

($H(3) = 1.378, p = .711$) and HARD ($H(3) = .483, p = .923$). Thus, *contrary to prior work in the* AI *community [BWZ*21, BLGG20, BMG21, LLT20], the amount of displayed information did not significantly affect the use of* AI *suggestions*, opposing **H2**.

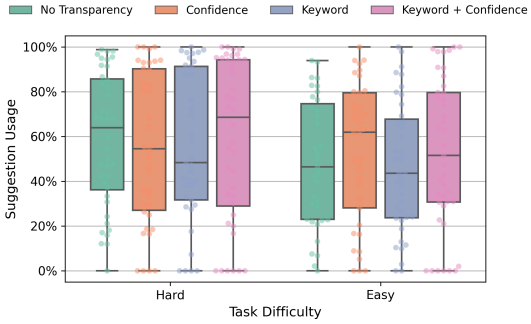

Figure 4: *The spread of* USAGE *across conditions. We found no significant difference among these groups for both* EASY *and* HARD.

Does transparency reduce overreliance? We ran separate Kruskal-Wallis tests to compare the rate of overreliance for participants within the transparency levels for EASY and HARD. Although the spread of overreliance within CONF was more dispersed than the rest of the transparency conditions in EASY, we found no significant difference among the transparency levels within both EASY ($H(3) = 2.648, p = .449$) and HARD ($H(3) = 913, p = .822$). Again, contrary to prior work in the AI community, *promoting transparency did not increase the participants' overreliance on the* AI *[BWZ*21] nor did it reduce overreliance [VJGM*22].* Figure 5 shows the spread of overreliance rates among the participants in all eight condition groups.

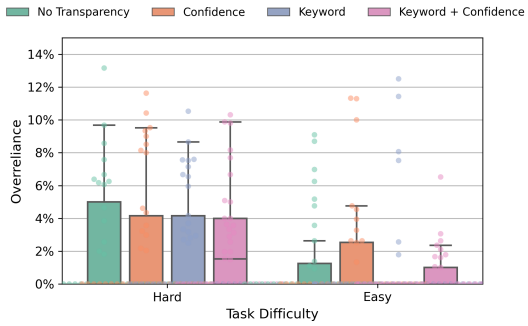

Figure 5: *The spread of overreliance across all conditions. Regardless of the transparency condition, participants in* HARD *overrelied on the suggestions at similar rates throughout the experiment.*

7.3. Does task difficulty affect subjective trust?

Overall, we observed high levels of subjective trust among our participants regardless of the transparency level and task difficulty. Out of all the recruited participants, 71% of them either agreed or strongly agreed that they trusted the system's suggestions, see Figure 6 for a breakdown of self-reported trust among all conditions.

With **H3**, we hypothesized that participants in HARD may have a higher level of TRUST than those in EASY. To test this hypothesis, we compared the distribution of TRUST between EASY and HARD. An independent sample one-tailed Mann-Whitney U test revealed that the distribution of TRUST levels in the AI for EASY was stochastically greater than the distribution of self-reported TRUST levels of those in HARD ($U = 13000.0, p = 0.0417, d = .008$), which is opposite to what we hypothesized in **H3**. We find suggestive evidence that *completing a difficult task with* AI *guidance elicited beliefs that the system was less trustworthy.*

However, lower AI accuracy in HARD conditions, as we saw in Section 7.1, is a confounding factor. We, therefore, considered the correlation between AI accuracy and self-reported trust. A Spearman's rank correlation found no measurable association overall ($r(305) = .0491, p < 0.391$)) nor for the HARD ($r(153) = -.0705, p < 0.383$)).

Is there a relationship between suggestion usage and trust? The literature on AI-assisted decision-making often utilizes suggestion usage as a proxy for trust [BWZ*21, ZLB20, WY21] as behavioral trust (measured in this study via suggestion usage) can be defined by an action based on cognitive (or subjective) trust [KS21]. A Spearman's rank correlation was computed to determine whether suggestion usage is a valid proxy for real-time trust in a VA system. For both EASY ($r(150) = .312, p < .0001$) and HARD ($r(153) = .245, p < .002$), there is a weak positive correlation between USAGE and TRUST, *supporting that suggestion usage has potential to be a valid indicator of subjective trust of users in real-time.*

7.4. Does transparency increase subjective trust?

In **H4**, we hypothesized that higher transparency would induce higher subjective trust, regardless of task difficulty. We ran separate Kruskal-Wallis tests to compare TRUST for participants within the transparency levels for both EASY and HARD. Our analysis found no significant difference in TRUST among the transparency levels for both EASY ($H(3) = 0.702, p = .873$) and HARD ($H(3) = 2.564, p = .464$). Thus, we conclude that *promoting* AI *transparency did not influence subjective trust.*

7.5. Does AI guidance induce bias?

A vital aspect of the VAST challenge was to generate hypotheses for how and where the epidemic was being transmitted. To answer this question, the participants need to understand the scope of symptoms and identify hotspots of the epidemic. We analyzed how AI guidance impacted the number of relevant symptoms discovered and the distribution of locations explored.

Does AI guidance encourage symptom diversity? Figure 7 shows the spread of relevant symptoms discovered between participants who did and did not receive AI guidance. We conducted an independent Mann-Whitney U test, which is robust to unequal sample size, to test whether there was a difference in SYMPTOM DIVERSITY between the CTRL (*median* = 14.0, IQR (10.0, 18.0)) and AI-guided (*median* = 16.0, IQR (12.0, 21.0)) groups. We found that the CTRL group *interacted with a significantly less diverse set of relevant symptoms* than the AI guided group ($U = 10477.5, p = .0054, \eta^2 = .02$).

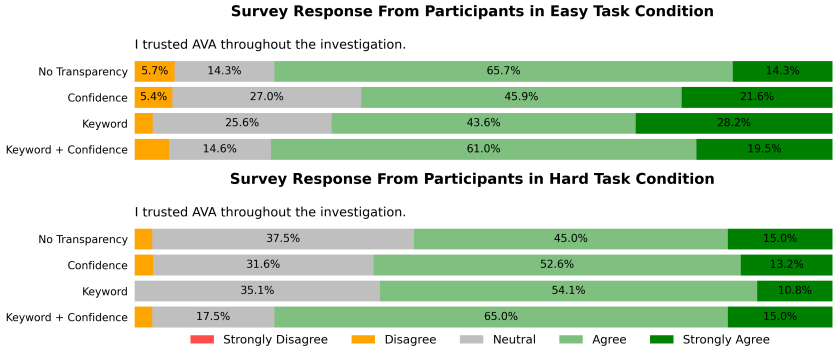

Figure 6: *Post-experiment survey responses on trust toward the* AI *separated by task difficulty and transparency level. We observed high ratings of trust across all conditions with* KWD+CONF *inducing the highest rating in both* EASY *and* HARD.

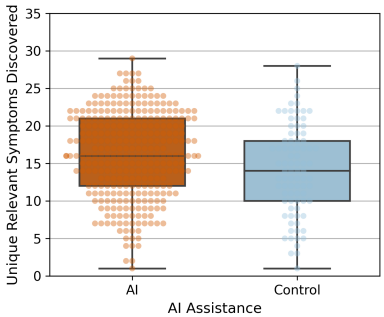

Figure 7: *The spread of relevant, unique symptoms discovered between participants who did and did not receive* AI *guidance. We observed that participants who received* AI *guidance found more unique symptoms related to the epidemic than the* CTRL.

Does the presence of AI guidance bias exploration? To gain a deeper understanding of whether AI guidance biased data exploration, we analyzed how the participants who did and did not receive AI guidance covered the dataset during their investigation by comparing the distribution of locations explored. See supplementary materials for visual distributions of locations investigated for the control and AI assistance groups that completed the hard task. *We observed similar exploration patterns between the two groups, with the* CTRL *exploration being moderately more biased than the* AI*-guided participants towards the lower right corner of the city.*

Is there a relationship between suggestion usage and symptom diversity? In the previous analysis, we observed that the AI-guided participants interacted with a significantly more diverse set of symptoms. Based on this finding, we then explored whether there was a relationship between USAGE and symptom diversity. A Spearman's rank correlation was computed, and we observed a weak positive correlation ($r(307) = .312, p < 0.001$)), *supporting that as the participants utilized more suggestions, the participants were able to uncover more relevant symptoms*.

8. Discussion

This user study aimed to observe how task difficulty and the promotion of transparency impact users' trust, AI suggestion usage, and data exploration whilst performing an analytical data foraging task with AI guidance. We discuss the implications of our findings.

8.1. Users were more likely to use suggestions for hard tasks.

Our analysis of how users interact with AI suggestions considered task difficulty as a variable. In particular, we examined a scenario where guidance might be more valuable but prone to errors. We observed that users tend to rely on AI suggestions more when finding relevant data is more challenging. While this is not surprising, our findings reveal that, at times, users may depend too much on suggestions. In particular, our study found that participants in more challenging conditions were more likely to accept suggestions from AI that were not entirely accurate or appropriate for the task. Participants who received no transparency similarly overrelied on AI suggestions compared to those who received some level of transparency. Moreover, we corroborate the prior findings in the AI community [VJGM*22], which showed an overreliance on AI suggestions in more difficult tasks and demonstrate that this overreliance extends to data exploration with VA systems.

Designers should consider these findings that indicate a higher dependence on AI for challenging tasks when developing VA tools with AI guidance. One solution to account for higher dependence is to create tools that can automatically detect the difficulty level of a task and adjust the level of AI support accordingly. For instance, if a task is deemed easy, the tool can reduce the frequency or intrusiveness of AI suggestions but provide more guidance for challenging tasks. One way to determine task difficulty is by analyzing the user's interaction with the system. The AI can learn from the duration of completing sub-tasks, the number of repeated or redundant interactions, and the frequency of guidance requests. These methods could rely on the existing body of work on analytic provenance [XOW*20]. Additionally, designers could consider incorporating user feedback, such as the user's belief in their domain expertise or the difficulty of the task.

8.2. Trust remained unaffected by transparency.

Our study aimed to investigate the lack of trust and suggestion usage in some participants, which was observed in Monadjemi et al. [MHN*22] with their VA scenario. We incorporated various levels of transparency to help establish an appropriate level of trust.

Our findings indicate that transparency did not significantly impact the user's interactions with the AI suggestions or subjective trust. Moreover, we contradict previous research that suggested transparency can enhance users' trust, confidence, and understanding of AI systems [BWZ*21, BLGG20].

VA may have a unique advantage in terms of the level of trust that users have in the system. If the data source is reliable, analysts may tend to approach problems with a belief in transparency and agency, which are common antecedents of trust. Therefore, the broader findings of AI may not necessarily apply to VA. Additionally, we observed null results which could be due to a ceiling effect. Since our participants' baseline trust was already high, the impact of transparency may not be significant.

It is also plausible that the lack of significant results was due to the simplicity of the transparency techniques employed. To provide more context for the user's final decision, it may be helpful to supplement the confidence value with additional explanations. However, this requires further investigation to determine the appropriate balance of awareness, to avoid information overload, and unwanted interaction behaviors such as aversion or overreliance.

8.3. A possible link between usage and exploration diversity.

A potential risk of providing AI guidance within VA systems is introducing or reinforcing the analyst's bias [WBP*18]. Our study found no evidence of bias, as suggestion usage led participants to discover a more diverse set of symptoms. However, this may highlight a limitation of this study as this phenomenon could be explained by the "wisdom of the crowd effect." The *wisdom of the crowd* refers to the collective intelligence of a group of individuals, which can sometimes lead to better outcomes than those made by individual experts. In contrast, many real-world VA scenarios do not benefit from such group dynamics. Future work (e.g., utilizing a case study approach or priming to elicit cognitive biases) could shed more light on the risks associated with designing and implementing AI-guided VA systems.

9. Limitations and Future Work

To balance control and generalizability, we designed this study using a VAST Challenge task that mimics real-world analytic tasks. In this task, our participants explored a dataset related to an epidemic using a map-based visualization, which means that our results may not apply to other scenarios beyond the specific task used. However, we observed that some of our findings confirm prior research. For example, we found that the use of AI can improve data discovery [MHN*22], and participants are more likely to accept AI suggestions when the task is challenging or the model is uncertain [WY21, JPM*21]. Considering that map-based visualizations are common on the web [BDM*18], and the study's exploratory nature allowed participants to adopt their desired exploration strategies, we believe that our findings can be widely applicable.

Our system leveraged the user's analytical provenance to continually retrain the model based on updated observations to suggest relevant data points. Since the AI did not have access to the ground truth beyond labels provided by the user, the suggestions were contingent upon the data tagged for investigation. Given the relatively high average AI accuracy in both the HARD and EASY task conditions as seen in Figure 3, we can see the potential of utilizing AI suggestions for collaborations in exploration and decision-making settings. Especially, there is an opportunity for future work to explore how transparency can affect subjective trust and suggestion usage in more complex and nuanced AI-guided VA scenarios.

Unlike previous work [ZLB20, LMY*21, BWZ*21, BMG21], we observed no effect of transparency on trust and suggestion usage. The transparency levels tested in this study do not represent all the established XAI methods in AI-assisted decision-making. We chose these transparency techniques based on what would be best suited for our specific AI algorithm and VA scenario. Therefore, further research is necessary to explore the observed high baseline trust phenomenon with the VA system, the design of explanations, and transparency-promoting techniques for visualization interfaces.

Although our crowdsourced approach captured a large and diverse study population, we acknowledge that our controlled user study has some limitations and confounds. One limitation is that the participants recruited for the study were neither asked nor screened about their experiences with visual data exploration tasks or their familiarity with AI guidance before participating. Therefore, our findings may not apply to expert user scenarios. Moreover, our participants could have altered their behavior or responses due to their awareness of being studied. This issue may have impacted our observations, limiting our ability to draw accurate conclusions about users' natural behavior. Also, our between-subjects comparisons prevented us from controlling for individual differences, potentially limiting our ability to capture the full extent of how the different transparency levels impact users' trust and interactions with the AI.

Lastly, consistent with prior work [MHN*22], we presented 10 relevant suggestions at a time from the AI to the participants who received guidance. Future work is needed to explore how varying the number of suggestions displayed at once impacts suggestion usage, data exploration, and cognitive load.

10. Conclusion

This paper explored the impact of AI suggestions on user behavior during data exploration, specifically with an AI-guided VA tool. We aimed to understand how task difficulty and transparency of the AI can affect users' trust, interactions, and data exploration. Our results suggest that participants tended to trust the AI regardless of the amount of transparency provided. We observed that the more difficult the task, the more likely users were to rely on suggestions provided, despite the AI providing suggestions at a lower rate of accuracy. Furthermore, we demonstrate that the level of detail provided to promote transparency had no measurable impact on data exploration, suggestion usage, and trust. Finally, we discuss the implications of our results, including some of the promises and challenges of promoting transparency of the AI in guided data discovery tools. Our findings underscore the importance of transparency in such AI-guided data discovery tools, prompting further inquiry into its role in fostering appropriate reliance on AI within VA systems.

Acknowledgements

This work is supported in part by the National Science Foundation under Grant No. OAC-2118201 and IIS-2142977.

References

[ADRDS*20] ARRIETA A. B., DÍAZ-RODRÍGUEZ N., DEL SER J., BENNETOT A., TABIK S., BARBADO A., GARCÍA S., GIL-LÓPEZ S., MOLINA D., BENJAMINS R., ET AL.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion 58* (2020), 82–115. 2

[BCS16] BATTLE L., CHANG R., STONEBRAKER M.: Dynamic Prefetching of Data Tiles for Interactive Visualization. In *Proceedings of the 2016 International Conference on Management of Data* (San Francisco California USA, June 2016), ACM, pp. 1363–1375. URL: https://dl.acm.org/doi/10.1145/2882903.2882919, doi:10.1145/2882903.2882919. 2

[BDM*18] BATTLE L., DUAN P., MIRANDA Z., MUKUSHEVA D., CHANG R., STONEBRAKER M.: Beagle: Automated extraction and interpretation of visualizations from the web. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (2018), pp. 1–8. 10

[BLBC12] BROWN E. T., LIU J., BRODLEY C. E., CHANG R.: Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Seattle, WA, USA, Oct. 2012), IEEE, pp. 83–92. URL: http://ieeexplore.ieee.org/document/6400486/, doi:10.1109/VAST.2012.6400486. 2

[BLGG20] BUÇINCA Z., LIN P., GAJOS K. Z., GLASSMAN E. L.: Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari Italy, Mar. 2020), ACM, pp. 454–464. URL: https://dl.acm.org/doi/10.1145/3377325.3377498, doi:10.1145/3377325.3377498. 8, 10

[BMG21] BUÇINCA Z., MALAYA M. B., GAJOS K. Z.: To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW1 (Apr. 2021), 1–21. URL: https://dl.acm.org/doi/10.1145/3449287, doi:10.1145/3449287. 1, 2, 8, 10

[BSO15] BUSSONE A., STUMPF S., O'SULLIVAN D.: The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics* (Dallas, TX, USA, Oct. 2015), IEEE, pp. 160–169. URL: http://ieeexplore.ieee.org/document/7349687/, doi:10.1109/ICHI.2015.26. 1, 3

[BWZ*21] BANSAL G., WU T., ZHOU J., FOK R., NUSHI B., KAMAR E., RIBEIRO M. T., WELD D.: Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), ACM. 8, 10

[CGM*17] CENEDA D., GSCHWANDTNER T., MAY T., MIKSCH S., SCHULZ H.-J., STREIT M., TOMINSKI C.: Characterizing Guidance in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 111–120. URL: https://ieeexplore.ieee.org/document/7534883/, doi:10.1109/TVCG.2016.2598468. 2

[CGW*12] COOK K., GRINSTEIN G., WHITING M., COOPER M., HAVIG P., LIGGETT K., NEBESH B., PAUL C. L.: Vast challenge 2012: Visual analytics for big data. In *2012 IEEE conference on visual analytics science and technology (VAST)* (2012), IEEE, pp. 251–255. 3

[CGW14] COOK K., GRINSTEIN G., WHITING M.: The vast challenge: History, scope, and outcomes: An introduction to the special issue, 2014. 3

[CJH19] CAI C. J., JONGEJAN J., HOLBROOK J.: The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray California, Mar. 2019), ACM, pp. 258–262. 2

[CWZ*19] CHENG H.-F., WANG R., ZHANG Z., O'CONNELL F., GRAY T., HARPER F. M., ZHU H.: 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* (Glasgow Scotland Uk, May 2019), ACM, pp. 1–12. 2

[DC17] DABEK F., CABAN J. J.: A Grammar-based Approach for Modeling User Interactions and Generating Suggestions During the Data Exploration Process. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 41–50. URL: http://ieeexplore.ieee.org/document/7534759/, doi:10.1109/TVCG.2016.2598471. 2, 7

[DLW*17] DASGUPTA A., LEE J.-Y., WILSON R., LAFRANCE R. A., CRAMER N., COOK K., PAYNE S.: Familiarity Vs Trust: A Comparative Study of Domain Scientists' Trust in Visual Analytics and Conventional Analysis Methods. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 271–280. URL: http://ieeexplore.ieee.org/document/7536106/, doi:10.1109/TVCG.2016.2598544. 2

[DSM15] DIETVORST B. J., SIMMONS J. P., MASSEY C.: Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General 144*, 1 (2015), 114–126. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000033, doi:10.1037/xge0000033. 2

[DSM18] DIETVORST B. J., SIMMONS J. P., MASSEY C.: Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science 64*, 3 (Mar. 2018), 1155–1170. 3

[GKX*12] GARNETT R., KRISHNAMURTHY Y., XIONG X., SCHNEIDER J., MANN R.: Bayesian optimal active search and surveying. In *Proceedings of the 29th International Conference on Machine Learning* (2012). 3

[GSR*21] GAUBE S., SURESH H., RAUE M., MERRITT A., BERKOWITZ S. J., LERMER E., COUGHLIN J. F., GUTTAG J. V., COLAK E., GHASSEMI M.: Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine 4*, 1 (Feb. 2021), 31. URL: https://www.nature.com/articles/s41746-021-00385-9, doi:10.1038/s41746-021-00385-9. 1

[HFRD13] HELLDIN T., FALKMAN G., RIVEIRO M., DAVIDSSON S.: Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Eindhoven Netherlands, Oct. 2013), ACM, pp. 210–217. 3

[JMC*17] JIANG S., MALKOMES G., CONVERSE G., SHOFNER A., MOSELEY B., GARNETT R.: Efficient Nonmyopic Active Search. In *Proceedings of the 34th International Conference on Machine Learning* (06–11 Aug 2017), Precup D., Teh Y. W., (Eds.), vol. 70 of *Proceedings of Machine Learning Research*, PMLR, pp. 1714–1723. 5

[JMMG18] JIANG S., MALKOMES G., MOSELEY B., GARNETT R.: Efficient nonmyopic active search with applications in drug and materials discovery. *Machine Learning for Molecules and Materials Workshop at NeurIPS* (2018). 3, 5

[JPM*21] JACOBS M., PRADIER M. F., MCCOY T. H., PERLIS R. H., DOSHI-VELEZ F., GAJOS K. Z.: How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry 11*, 1 (Feb. 2021), 108. URL: https://www.nature.com/articles/s41398-021-01224-x, doi:10.1038/s41398-021-01224-x. 2, 10

[KDM*19] KUNKEL J., DONKERS T., MICHAEL L., BARBU C.-M., ZIEGLER J.: Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland Uk, May 2019), ACM, pp. 1–12. URL: https://dl.acm.org/doi/10.1145/3290605.3300717, doi:10.1145/3290605.3300717. 1, 7

[KJO*19] KERY M. B., JOHN B. E., O'FLAHERTY P., HORVATH A., MYERS B. A.: Towards Effective Foraging by Data Scientists to Find Past Analysis Choices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow Scotland Uk, May 2019), ACM, pp. 1–13. URL: https://dl.acm.org/doi/10.1145/3290605.3300322, doi:10.1145/3290605.3300322. 2

[KS21] KIM T., SONG H.: How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics 61* (2021). URL: https://linkinghub.elsevier.com/retrieve/pii/S0736585321000344, doi:10.1016/j.tele.2021.101595. 8

[KYZ23] KIM A., YANG M., ZHANG J.: *When Algorithms Err* : Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms. *ACM Transactions on Computer-Human Interaction 30*, 1 (Feb. 2023), 1–36. URL: https://dl.acm.org/doi/10.1145/3557889, doi:10.1145/3557889. 2

[LGG*18] LIN H., GAO S., GOTZ D., DU F., HE J., CAO N.: RCLens: Interactive Rare Category Exploration and Identification. *IEEE Transactions on Visualization and Computer Graphics 24*, 7 (July 2018), 2223–2237. URL: https://ieeexplore.ieee.org/document/7939996/, doi:10.1109/TVCG.2017.2711030. 2

[LLT20] LAI V., LIU H., TAN C.: "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA, Apr. 2020), ACM, pp. 1–13. URL: https://dl.acm.org/doi/10.1145/3313831.3376873, doi:10.1145/3313831.3376873. 8

[LMY*21] LINDER R., MOHSENI S., YANG F., PENTYALA S. K., RAGAN E. D., HU X. B.: How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters 2*, 4 (Dec. 2021). 3, 4, 7, 10

[LST*21] LEE D. J.-L., SETLUR V., TORY M., KARAHALIOS K. G., PARAMESWARAN A.: Deconstructing Categorization in Visualization Recommendation: A Taxonomy and Comparative Study. *IEEE Transactions on Visualization and Computer Graphics* (2021), 1–1. URL: https://ieeexplore.ieee.org/document/9444894/, doi:10.1109/TVCG.2021.3085751. 2

[LT19] LAI V., TAN C.: On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta GA USA, 2019), ACM, pp. 29–38. 3

[MHN*22] MONADJEMI S., HA S., NGUYEN Q., CHAI H., GARNETT R., OTTLEY A.: Guided data discovery in interactive visualizations via active search. In *2022 IEEE Visualization and Visual Analytics (VIS)* (2022), pp. 70–74. 1, 2, 4, 5, 7, 9, 10

[MLW*23] MA S., LEI Y., WANG X., ZHENG C., SHI C., YIN M., MA X.: Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg Germany, Apr. 2023), ACM, pp. 1–19. URL: https://dl.acm.org/doi/10.1145/3544548.3581058, doi:10.1145/3544548.3581058. 1

[NKR20] NOURANI M., KING J., RAGAN E.: The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8* (Oct. 2020), 112–121. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7469, doi:10.1609/hcomp.v8i1.7469. 2

[PS18] PALAN S., SCHITTER C.: Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance 17* (2018), 22–27. 5

[ŘS10] ŘEHŮŘEK R., SOJKA P.: Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta, Malta, May 2010), ELRA, pp. 45–50. http://is.muni.cz/publication/884893/en. 5

[RSG16] RIBEIRO M. T., SINGH S., GUESTRIN C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco California USA, Aug. 2016), ACM, pp. 1135–1144. URL: https://dl.acm.org/doi/10.1145/2939672.2939778, doi:10.1145/2939672.2939778. 1

[SOM*19] SCHAFFER J., O'DONOVAN J., MICHAELIS J., RAGLIN A., HÖLLERER T.: I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray California, Mar. 2019), ACM, pp. 240–251. URL: https://dl.acm.org/doi/10.1145/3301275.3302308, doi:10.1145/3301275.3302308. 2

[SSK*16] SACHA D., SENARATNE H., KWON B. C., ELLIS G., KEIM D. A.: The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 240–249. 3

[VJGM*22] VASCONCELOS H., JÖRKE M., GRUNDE-MCLAUGHLIN M., GERSTENBERG T., BERNSTEIN M., KRISHNA R.: Explanations can reduce overreliance on ai systems during decision-making. *arXiv preprint arXiv:2212.06823* (2022). 2, 3, 6, 7, 8, 9

[WBP*18] WALL E., BLAHA L. M., PAUL C. L., COOK K., ENDERT A.: Four Perspectives on Human Bias in Visual Analytics. In *Cognitive Biases in Visualizations*. Springer International Publishing, 2018, pp. 29–42. URL: http://link.springer.com/10.1007/978-3-319-95831-6_3, doi:10.1007/978-3-319-95831-6_3. 10

[WY21] WANG X., YIN M.: Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (2021), ACM, pp. 318–328. 1, 2, 3, 4, 8, 10

[XOW*20] XU K., OTTLEY A., WALCHSHOFER C., STREIT M., CHANG R., WENSKOVITCH J.: Survey on the analysis of user interactions and visualization provenance. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 757–783. 2, 9

[YHSA20] YANG F., HUANG Z., SCHOLTZ J., ARENDT D. L.: How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari Italy, Mar. 2020), ACM, pp. 189–201. URL: https://dl.acm.org/doi/10.1145/3377325.3377480. 2

[ZLB20] ZHANG Y., LIAO Q. V., BELLAMY R. K. E.: Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona Spain, Jan. 2020), ACM, pp. 295–305. 1, 2, 3, 4, 8, 10

[ZWM*23] ZHAO J., WANG Y., MANCENIDO M. V., CHIOU E. K., MACIEJEWSKI R.: Evaluating the Impact of Uncertainty Visualization on Model Reliance. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–15. URL: https://ieeexplore.ieee.org/document/10058545/, doi:10.1109/TVCG.2023.3251950. 3