Trust Your Gut: Comparing Human and Machine Inference from Noisy Visualizations

Ratanond Koonchanok, Michael E. Papka, Khairi Reda

Abstract—People commonly utilize visualizations not only to examine a given dataset, but also to draw generalizable conclusions about the underlying models or phenomena. Prior research has compared human visual inference to that of an optimal Bayesian agent, with deviations from rational analysis viewed as problematic. However, human reliance on non-normative heuristics may prove advantageous in certain circumstances. We investigate scenarios where human intuition might surpass idealized statistical rationality. In two experiments, we examine individuals' accuracy in characterizing the parameters of known data-generating models from bivariate visualizations. Our findings indicate that, although participants generally exhibited lower accuracy compared to statistical models, they frequently outperformed Bayesian agents, particularly when faced with extreme samples. Participants appeared to rely on their internal models to filter out noisy visualizations, thus improving their resilience against spurious data. However, participants displayed overconfidence and struggled with uncertainty estimation. They also exhibited higher variance than statistical machines. Our findings suggest that analyst gut reactions to visualizations may provide an advantage, even when departing from rationality. These results carry implications for designing visual analytics tools, offering new perspectives on how to integrate statistical models and analyst intuition for improved inference and decision-making. The data and materials for this paper are available at `https://osf.io/qmfv6`

Index Terms—Visual inference, statistical rationality, human-machine collaboration.

1 Introduction

Visualizations play an increasingly important role in data analysis and decision-making. Crucially, these tools are not just used for extracting numbers; often, their greater value lies in revealing insights about the processes that *generated* the data in the first place. For example, a business analyst who observes a visualization showing an increase in luxury coat sales might seek to understand the underlying factors, such as unusually cold winters, an increase in disposable incomes, or whether this trend was simply a fluke. This deeper understanding can in turn help analysts perform higher-level tasks, such as attributing causality, validating hypotheses, or predicting future observations. Making accurate inferences from visualizations, however, can be challenging, requiring one to account for potential uncertainties and natural variabilities in the data. This interpretive process could expose the viewer to pitfalls. For example, the viewer might overinterpret a visualization displaying an unusual or extreme sample, causing them to infer spurious features not found in the data-generating process. Conversely, the viewer may fail to consider the data sufficiently and instead fall back to their prior belief — a potential manifestation of confirmation bias. Inferential errors during visual analysis can lead to invalid conclusions and false discoveries [70]. While the likelihood of incorrect machine inference can be quantified analytically or through simulations, our understanding of human inference-making from visualizations remains limited.

Prior research has approached this question by comparing the visual inferences people make to those of an optimal Bayesian agent that observes the same data [40]. This method evaluates whether a viewer's interpretation, such as the level of correlation between two variables, aligns with the conclusions of a Bayesian that accounts for both the viewer's preexisting knowledge and the data observed [37]. The level of divergence between the viewer's inference and the Bayesian model determines how optimal the viewer's interpretation is; visualization techniques eliciting smaller divergence are deemed more effective.

Although useful as a reference point, the suitability of a fully normative agent as a universal benchmark target is not always justified on an ecological level [14]. Unlike statistical machines, humans do not possess the necessary cognitive resources to process information like a perfectly rational agent. Additionally, rational agents do not exhibit the nuances of human cognition. In certain contexts, deviations from rationality can indeed be advantageous, leading to a closer approximation of the underlying reality. For example, a Bayesian agent might perceive a positive correlation between two unrelated variables, such as the number of ice cream cones sold and the number of shark attacks, despite a prior suggesting no connection. In contrast, a human observer could become increasingly skeptical upon observing the same dataset, irrationally reinforcing their belief in no correlation. While the Bayesian approach is normatively superior, the human response could lead to better inference in such scenarios. Specifically, the non-rational response of reinforcing a 'null' belief in the face of contrary evidence can instill healthy skepticism, guarding against false positives.

In this work, we move away from the assumption that a rational agent is always ideal. Instead, we compare the *utility* of both human and Bayesian inference in accurately characterizing data-generating processes under varying uncertainty and sample conditions. Specifically, we investigate situations where human visual inferences might outperform those of an ideal Bayesian (and vice versa). Although we expect a Bayesian agent to be globally optimal, we foresee scenarios where factors such as sample size and outlyingness could give the human analyst an edge. This perspective acknowledges that human heuristics, though seemingly inferior [62], can sometimes give rise to better inference [26]. By characterizing factors that lead analysts to exceed machine-inference performance, we pave the way for designing more intelligent visual analytics systems. Such systems might adaptively encourage analysts to leverage their intuition and 'trust their gut', or, at other times, provide computational assistance when users are more prone to biased inference-making.

We conducted two experiments to study how people make visual inferences in response to bivariate visualizations on a range of topics. Participants externalized both their prior beliefs (i.e., before seeing data) and posteriors (i.e., post-data exposure) using a graphical elicitation device. We manipulated several factors, including the visualization type, sample size, and the extremeness of the sample (i.e., the degree to which it is *inconsistent* with the ground truth). We find that participants were more accurate at inferring the true correlation level than both informed and uninformed Bayesian models when the visualization showed extreme samples. This advantage was particularly pronounced with icon arrays, and in situations where there was a higher consensus around the ground truth. However, participants were less precise in

R. Koonchanok and K. Reda are with Indiana University Indianapolis. Email: {rkoonch, redak}@iu.edu. M. E. Papka is with Argonne National Laboratory and the University of Illinois Chicago. Email: papka@anl.gov

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

estimating the true uncertainty of the data-generating process, often displaying unwarranted confidence. In a second experiment that varied the uncertainty level in the generating models, participants were better at inferring the true correlation at low uncertainty, although they did not necessarily improve their performance relative to machines.

Our findings reveal that individuals frequently diverge from normative inference when interpreting visualizations. Importantly, these deviations proved advantageous, particularly when examining extreme datasets and small samples. In such instances, an analyst's intuition may yield more accurate conclusions on average than those generated by idealized statistical machines, even when the latter are endowed with human priors. Conversely, for larger and more reliable datasets, the precision offered by statistical inference seems to reduce both bias and variability, leading to better estimation of the generating model. These findings highlight the potential for integrating human and machine inference in visualization tools, providing guidance on how to interweave the two capabilities. Moreover, our results suggest an alternative design strategy that seeks to harness human heuristics in visual analytics, as opposed to entirely aligning analyst behaviors with rational frameworks. We discuss the implications and highlight future research directions.

2 Background & Related Work

2.1 Heuristic vs. Rational Decision-Making

Heuristics are commonly associated with the notion of "taking shortcuts" to simplify a cognitively demanding task. As such, heuristics have been viewed as an inferior and unreliable form of cognition compared to rational decision-making [29, 35]. However, recent research challenges this perspective, showing the potential of heuristics in helping people come to good decisions [2, 25]. In certain situations, intuitive decision-making with heuristics can lead to superior outcomes compared to rational, analytical reasoning [16, 58], especially in high-risk, high-uncertainty scenarios [34]. For instance, managers employing heuristics sometimes make more effective decisions than those relying on statistical procedures like logistic regression [46]. In particular, when information is limited, individuals could achieve better judgments by using simple heuristics rather than attempting to process the available information [27]. In other words, by adaptively disregarding certain information, one could enhance their decision quality [28].

Non-rational thinking could also prove beneficial in visual analytics, especially when one is faced with improbable data. For example, analysts could grow skeptical and deviate from normative practices by deliberately underweighting the evidence implied by the data. Such heuristics, which are not always amenable to algorithmic modeling, can enhance inference from visualizations in specific scenarios. Rather than seeking to align analyst performance with normative, statistical inference [37, 39], we seek a better understanding of how human and machine inference-making can complement each other.

2.2 Inference from Visualizations

Statistical inference involves drawing conclusions about a population based on a (limited) sample of data from that population. Although inference is often performed with the aid of formal statistical models, it is possible to make inferences from visualized data [6]. For visual inferences to be reliable, one must be able to distinguish between real effects and spurious patterns [70]. Towards that end, methods have been proposed to safeguard the graphical inference process by controlling the rate of false discovery [59, 60, 71]. The lineup protocol is one of the early influential techniques in this space [68], and is often considered a visual analog to null-hypothesis significance testing [49]. Lineups, however, can be difficult to use in practice due to the need to develop realistic null models that can be compared to real datasets [4].

More recent work aims to allow visual analysts to operationalize their prior beliefs in the interpretation of visualizations [48]. This approach provides affordances for analysts to assess the compatibility of their models with data [10, 11, 36, 42, 56]. By encouraging analysts to explicitly test their models, researchers aim to foster a more normative, Bayesian-grounded inference from visualizations [37, 40]. Generally speaking, Bayesian inference provides a way to tap into expert knowledge [16], encoding the latter as prior models that can help normalize against outlying data, or supply additional information when data is limited. For example, in climate modeling, prior information can improve estimates of rainfall in the face of potentially extreme measurements [12, 13]. Similarly, in wildlife research, where species might not have the same observable abundance, borrowing information from related species can improve inference about rare populations [47]. Koonchanok et al. propose a parallel mechanism in visualizations [43]. They demonstrate that eliciting priors can reduce false discoveries by allowing analysts to remain vigilant against spurious visualizations. Karduni et al. investigated how individuals update their beliefs when making sense of bivariate visualizations [37]. They show that incorporating uncertainty representations enables a more Bayesian-like belief update. Others have explored how graphical model elicitation might engender attitude change through increased cognitive processing of visualizations [33, 50, 57]. Building on prior work in belief elicitation [10, 42, 43, 48], we employ interactions to externalize people's priors and posteriors, and compare human performance to that of Bayesian agents.

3 Research Questions & Methods

We investigate the accuracy of human inference-making from visualizations, evaluating how well these inferences align with the data-generating processes under various uncertainty and visualization conditions. Our central hypothesis is that although analysts will deviate from Bayesian rationality (and thus make suboptimal inferences in the aggregate), they can still update their beliefs in ways that more accurately reflect reality. Specifically, non-normative belief updating, guided by intuition and hunches, could be useful in high uncertainty or noisy sample conditions. In effect, human inference-making can complement strictly rational agents. Thus, we pose two questions:

RQ1: How do visual analysts compare to Bayesian agents that see that same data? How well do humans perform under different visualizations? We compare visual inferences of analysts to two Bayesians: one informed with the prior knowledge of the human analyst, and another with a uniform, flat prior that provides no preexisting knowledge, forcing a purely data-driven inference.

RQ2: How do sample and ground-truth characteristics influence visual inferences relative to Bayesian agents? Factors such as the perceived reliability (or extremeness) of the sample and the degree of ground consensus behind the generating process may impact how individuals derive insights about the data-generating mechanism.

To address these questions, we conducted two crowdsourced experiments. We evaluate individuals' accuracy in inferring true bivariate relationships between attribute pairs after exposure to (potentially noisy) samples. We collected responses from viewers before and after exposure to visualizations, recording both their prior beliefs and posterior inferences about the data-generating process. We measure the accuracy of human-vs-Bayesian inferences under different sample configurations, ranging from large samples that reflect the true parameters to small and potentially extreme samples. Additionally, we assess the effectiveness of multiple visualization types for this purpose.

3.1 Model Elicitation

We employ a graphical elicitation device to externalize participants' prior and posterior beliefs (illustrated in Figure 1). This interface enables participants to specify two parameters via two sliders according to their beliefs: the expected correlation coefficient between the two variables (μ) and the associated uncertainty in the correlation (σ). Specifically, the first slider prompts participants to indicate the "most likely relationship" along a continuum from 'negative' to 'positive' correlation. The second slider prompts participants to express their "confidence level" in the relationship, ranging from 'highly uncertain' to 'moderately' and 'highly certain'. These two parameters populate the following model:

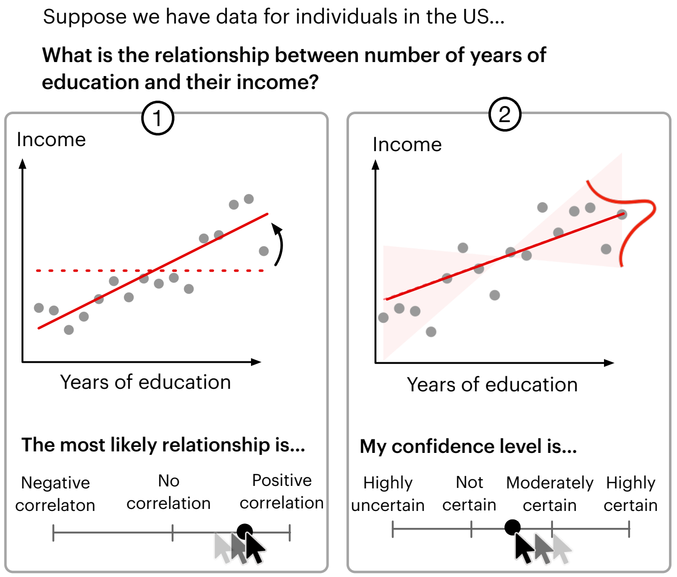
Suppose we have data for individuals in the US...
What is the relationship between number of years of education and their income?
1
Income
Years of education
The most likely relationship is...
Negative correlaton
No correlation
Positive correlation
2
Income
Years of education
My confidence level is...
Highly uncertain
Not certain
Moderately certain
Highly certain

Fig. 1: A graphical belief elicitation device for expressing beliefs about bivariate relationships in response to a prompt (top). Participants externalized their prior and posterior beliefs in two steps: (1) indicating the most likely relationship, and (2) specifying their uncertainty in the relationship. These slider settings update two parameters, μ and σ, in a linear model. During this interaction, participants see hypothetical samples from this model, refreshed at 5Hz, illustrating what the bivariate data might look like if their beliefs were true.

$$
\begin{aligned}
y_i &= \beta_0 + \beta x_i + \varepsilon_i \\
\beta &\sim \mathcal{N}(\mu, \sigma^2) \\
\beta_0 &\sim \mathcal{N}(0, \sigma_b^2) \\
\varepsilon_i &\sim \mathcal{N}(0, \sigma_e^2)
\end{aligned}
\tag{1}
$$

Where μ is the *expected slope* of the relationship as specified by the first slider, and σ is the *uncertainty* in the slope, as specified through the second slider. The intercept β_0 for the regression line is centered around 0, with a fixed standard deviation of $\sigma_b = 0.1$. ε_i is an additional residual term with a fixed standard deviation of $\sigma_e = 0.45$.

During user interaction, the elicitation interface visualizes samples generated from the above model. This allows participants to see hypothetical outcomes, refreshed at 5Hz, showing what the data *might* look like if their belief was true. This type of animated display is meant to help participants intuitively grasp the effects of manipulating the two sliders and the uncertainty implied by their (prior) model. The hypothetical samples are displayed in a scatterplot (Figure 1), a bar chart, or an icon array (Figure 3-A and B), depending on the experimental condition. However, in all conditions, participants utilize the same two sliders to specify their beliefs. The hypothetical samples are only animated during the elicitation step and shown in grey color to help distinguish them from the ground truth data samples (solid black).

3.2 Stimuli and Data-Generating Models

To provide plausible stimuli for the study, we developed a set of 40 prompt questions and corresponding ground truth models. The prompt questions covered a variety of common knowledge topics (see Table 1 for examples), with bivariate relationships ranging from negative correlations to positive correlations. Additionally, we also included attributes with no plausible relationship. The corresponding ground truth models for these prompts followed the same form as Equation 1. To initialize these models with plausible parameters rooted in common wisdom, we employed crowd workers recruited through Amazon Mechanical Turk. Workers ($n = 61$) were prompted to respond to each of the 40 questions, using the model elicitation interface in Figure 1 to provide their belief on the most likely relationship slope and their uncertainty around that relationship. Responses from the workers were then averaged forming the two ground truth parameters for each prompt question. Subsequently, we selected 24 prompt questions to be used in the experiment. These questions comprised six ground truth models with a positive relationship ($\mu_{Truth} > 0$), six questions with a negative relationship ($\mu_{Truth} < 0$), and 12 with no correlation ($\mu_{Truth} \approx 0$).

Table 1: Example prompt questions and associated ground-truth parameters as obtained from crowd workers.

Correlation	Consensus	Prompt question	Ground truth
Positive	Low	What is the relationship between alcohol consumption and an individual's stress level?	$\mu_{Truth} = 0.394$ $\sigma_{Truth} = 0.166$
Negative	High	What is the relationship between time spent on social media and hours of sleep at night?	$\mu_{Truth} = -0.422$ $\sigma_{Truth} = 0.146$
No relation	High	What is the relationship between the length of people's first and last name?	$\mu_{Truth} = 0.03$ $\sigma_{Truth} = 0.236$

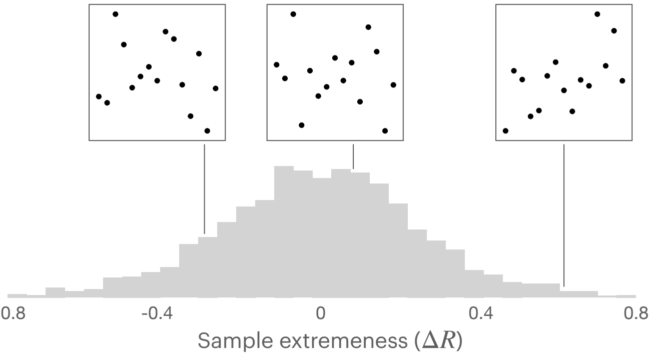

Fig. 2: Distribution of extremeness for 'medium' sample sizes ($N = 15$ points). With $\Delta R \approx 0.1$, the middle sample is a relatively faithful depiction of the ground truth (no correlation in this example). The sample on the left is more extreme with $\Delta R \approx -0.3$. The right-most scatterplot represents an even more extreme occurrence with $\Delta R \approx 0.6$.

Social Consensus: Individuals often rely on social knowledge when forming their beliefs. Consensus (whether perceived or actual) can also serve as a tool to reduce uncertainty. Therefore, we expect the agreement around the ground truth to impact people's inferences from visualizations. To quantify the latter, we measure the consistency of the crowd wisdom: Prompt questions exhibiting smaller variations in the elicited crowd beliefs are considered to reflect a higher degree of social consensus. This was determined based on the standard deviation of the elicited μ responses among workers. Within each category (positive, negative, or no relationship), we designate half the prompt question with the lowest standard deviation as 'high' consensus, with the other half considered 'low' consensus, representing lower agreement between workers on what the data-generating process should be. Including this factor in our analysis helps control for prior beliefs. Participants are expected to have stronger priors when responding to high-consensus questions. Conversely, priors are likely to be much weaker for low-consensus questions, leading participants to become more data-driven in their inference, and potentially more susceptible to extreme samples.

3.3 Sample Configurations

For each stimulus, we display the prompt question and ask the participant to provide their prior belief about the topic (μ and σ). We then expose participants to a random data sample generated from the ground truth model. Subsequently, we prompt them to reflect and provide a posterior inference using the same elicitation device as before. To understand how sample characteristics affect inference accuracy, we varied two properties: sample size and extremeness.

Sample Size: The number of data points in a sample is an important factor for an analyst to consider when making inferences. Larger samples furnish stronger evidence about the underlying data-generating process. We thus varied the size of samples shown to participants, using 7, 15, and 30 data points to represent 'small', 'medium', and 'large' sample sizes, respectively (these labels are only to illustrate the results, and not shown in the experiment). These sizes were selected to provide varying levels of evidence, while still allowing for extreme samples to emerge. Specifically, smaller samples are more prone to noise, giving a potentially misleading picture of the underlying ground truth.

Sample Extremeness: A core question that we address is how resilient people can be to extreme (or spurious) data, as compared with idealized statistical machines. Sample *extremeness* reflects the degree to which it is *inconsistent* with the data-generating process (e.g., a sample suggesting a positive correlation when the underlying model prescribes no relationship). Under a random sampling regime, spurious samples are expected to arise. We take advantage of this phenomenon and quantify the extremeness of the emerging samples in the experiment. We use Pearson's coefficient (R) to measure correlation strength for a sample and, by extension, the degree to which it can be considered *extreme* with respect to its underlying model. Specifically, we compute a difference (ΔR) between the sample's and the *expected* correlation:

$$\Delta R = R_{Sample} - R_{Expected} \tag{2}$$

$R_{Expected}$ is the mean R-value, determined from 10,000 simulated draws from the ground truth. A sample that is a faithful representation of its model will have $\Delta R \approx 0$. As ΔR deviates (negatively or positively), the sample will be considered more extreme. Under the law of large numbers, we would expect the distribution of ΔR to be centered around zero, with the majority of samples exhibiting low extremeness (see Figure 2). However, extreme samples would be expected to emerge. Additionally, we anticipate that sample size will influence the likelihood of extremeness: larger samples are expected to have a narrower ΔR distribution, while smaller samples may yield a wider extremeness distribution. Ground-truth models with higher uncertainty (σ) will also produce more spurious data. By quantifying the extremeness of emerging samples, we can observe how participants react to spurious data. We can then assess the extent to which participants (and Bayesian agents) can resist misleading samples.

4 EXPERIMENT I

This experiment aims to compare the inference of participants to ideal Bayesian agents. We specifically test the extent to which participants can infer the ground truth parameters as a function of visualization type, sample size and extremeness, and the degree of social consensus. Participants saw prompt questions described in §3.2 and were asked to predict the parameters of a bivariate ground truth model, namely, the expected slope μ and uncertainty σ. For each stimulus, participants were asked to provide their prior beliefs through the graphical elicitation interface described earlier. After that, they saw a sample from the model and were asked to reflect and provide their updated (i.e., posterior) inference about the true relationship.

4.1 Experiment Design

We adopted a mixed design, investigating three independent factors: Visualization (3 types) × Sample size (3 levels) × Social consensus (2 levels). Visualization was varied between subjects, whereas sample size and social consensus were varied within-subject.

Visualization: We evaluated scatterplots, bar charts, and icon arrays for representing samples. Scatterplots were chosen for their perceptual effectiveness in illustrating correlations [30]. Bar charts on the other hand aggregate the X axis into ordered groups (e.g., 'low' and 'high income'), averaging the second attribute (e.g., BMI) within these groups (see Figure 3-A). Icon arrays similarly aggregate the X axis into discrete 'low' and 'high' bins but show individual data points with color-coded Y values (Figure 3-B). While scatterplots offer a full view of the sample, both bar charts and icon arrays provide simplified, aggregate representations. Icon arrays were particularly noted for aiding conditional probability assessment of the sort needed for Bayesian reasoning [51]. They have also been used extensively in communicating uncertainty (e.g., medical risks [22, 23]) to non-experts, making them a suitable visualization to test in this experiment. That said, the collapse of continuous variables into discrete ordinals in both icon arrays and bar charts may ultimately prove detrimental.

Sample and Ground Truth Characteristics: Participants completed eight trials with each of the three sample sizes (small, medium, and large). Of these eight trials, four were presented with high-consensus prompts and four were low-consensus. Sample extremeness was included as an explanatory variable but was left to vary as a consequence of the random sampling process. Hence, while not explicitly controlled, extremeness covaried following a predictable distribution (Figure 2).

4.2 Hypotheses

We developed four hypotheses:

H1 – Compared to the two Bayesian agents, participants will be less accurate *overall* in recovering the true slope of the data-generating process. The global suboptimality of human analysts can be attributed to the difficulty in performing a fully normative inference from data on a visual basis. That said, we hypothesize that the informed Bayesian will outperform the uninformed (i.e., flat prior) agent, as the former benefits from participants' knowledge.

H2 – Although participants will make less-optimal inferences in the aggregate, we expect them to handle extreme samples better than Bayesian agents. Specifically, we hypothesize that participants will employ non-statistically normative heuristics to 'adjust' for improbable data, leading to more accurate inference. We expect this tendency to be more evident when there is greater consensus around the ground truth and with smaller sample sizes.

H3 – Participants will perform best with scatterplots, thanks to their effectiveness in displaying bivariate correlations. Icon arrays should also yield good results, though to a lesser degree. Both visualizations allow for individual data point observation. In contrast, participants will struggle with bar charts due to the absence of uncertainty representations, making this visualization somewhat less optimal for inference.

H4 – We anticipate reduced accuracy in participants' judgment of the model's uncertainty. Specifically, we expect human analysts to perform less effectively than both the informed and uninformed Bayesian agents when characterizing uncertainty.

4.3 Participants

We recruited participants from Prolific who are US residents and are at least 18 years old. Prospective participants were initially screened through basic questions on bivariate relationships to ensure sufficient background. We ultimately enrolled 222 participants in the experiment (112 males, 104 females, 6 others) with a mean age of 35.4 years. Education levels for participants were as follows: 56 high school, 24 associate, 89 bachelor's, 39 master's, 13 doctorate, and 1 unspecified. Participants were randomly assigned to one of the three visualization types (74 in each group). They were compensated with $5 for an hourly wage of $16.7. The study was approved by Indiana University's IRB.

4.4 Procedure

Participants first completed a tutorial explaining the goals of the experiment and introducing the graphical elicitation device. The instructions emphasized that the samples shown could be noisy, particularly for small samples. Participants then completed one practice trial, after which they completed the main study which consisted of 24 trials, corresponding to the prompt questions developed in §3.2, presented in random order. In addition to the analyzed trails, we included four engagement checks. The checks instructed participants to set the sliders to specific values (e.g., extreme positive correlation).

In each trial, participants were first presented with a context (e.g., "supposed we have data on US cities"), and were then prompted to predict a linear relationship between two variables (e.g., "what is the relationship between the unemployment rate and the affordability of housing?") Participants were asked to visually provide their prior belief through the elicitation device (§3.1). Next, the participants were presented with a sample drawn from the corresponding ground-truth model. The sample was shown side-by-side with participants' prior.

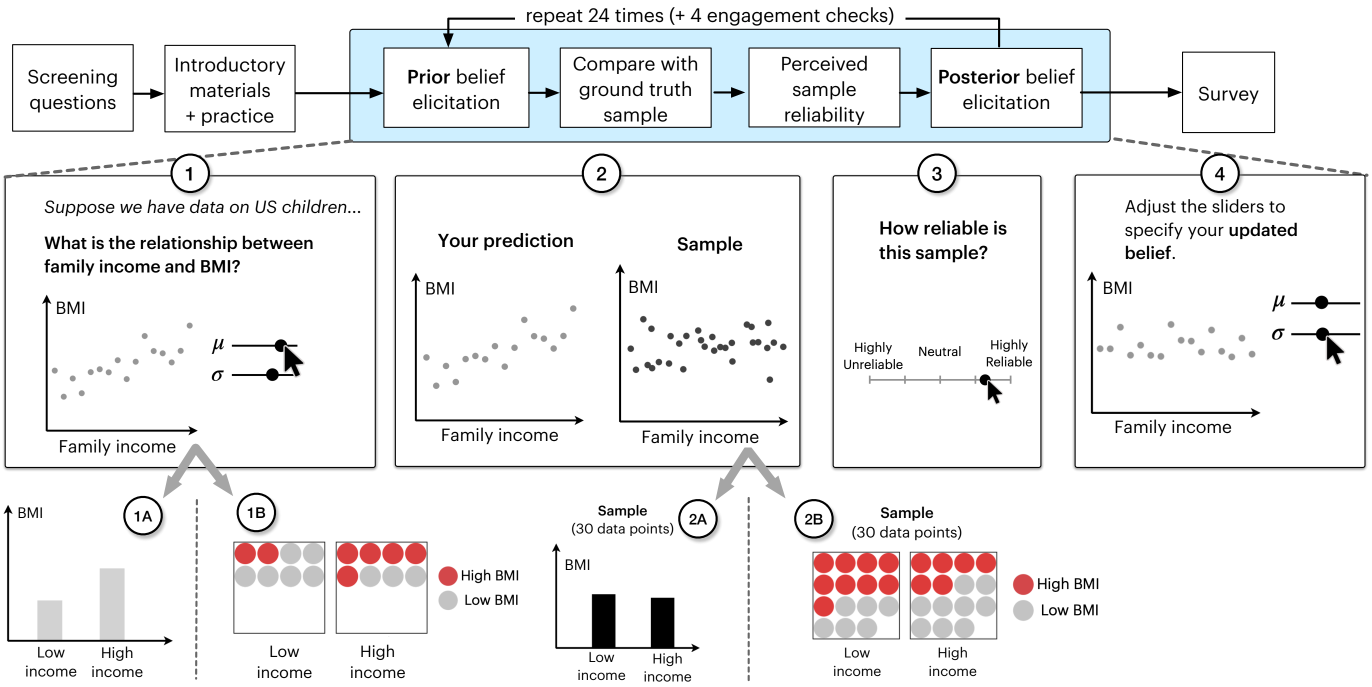

Fig. 3: Top: Flow diagram illustrating the procedure for Exp. 1. The blue-shaded section represents a single trial. In each trial, participants: 1) externalize their prior belief about a prompt question by adjusting two sliders, 2) observe a sample from the ground truth presented alongside their belief, 3) indicate how reliable they believe the sample is, and 4) specify their updated (posterior) belief by re-adjusting the sliders. Depending on the visualization condition, participants are either shown scatterplots, bar charts (A), or icon arrays (B).

They were prompted to provide their impression of sample reliability via a slider. Lastly, participants were prompted to update their beliefs by re-adjusting. Participants, however, were instructed that it is up to them to decide "how much (or little) to adjust [their] initial beliefs." Figure 3 illustrates this sequence. The visualized samples along with the belief elicitation device varied depending on the visualization condition (scatterplots, bar charts, or icon arrays). However, participants entered their responses using an identical set of sliders, with the difference being in how the samples, priors, and posteriors were visualized.

4.5 Response and Accuracy Metrics

Participants supplied five continuous responses in every trial: the most-likely slope (μ_{Human}) and uncertainty (σ_{Human}) in the relationship, both before (i.e., prior belief) and after sample exposure (i.e., posterior inference). Additionally, they provided their perception of sample reliability (reported in the supplemental materials). For every trial, we generated two comparable statistical inferences. The first (μ_{Bayes} and σ_{Bayes}) was derived through Bayesian inference using the participant-provided prior and the likelihood function implied by the sample observed by the participant. This corresponds to an idealized statistical agent equipped with identical prior knowledge as the participant. The second response, also a Bayesian, utilized a flat prior in conjunction with the same likelihood function, representing an uninformed agent that exclusively learns the two parameters (μ_{Flat} and σ_{Flat}) from the sample. We quantify inference accuracy for each of the three agents by measuring the divergence of the inferred parameters from the ground truth:

$$\begin{aligned} \Delta\mu &= \frac{\mu_{Human|Bayes|Flat} - \mu_{Truth}}{2} \times 100 \\ \Delta\sigma &= \frac{\sigma_{Human|Bayes|Flat} - \sigma_{Truth}}{0.6} \times 100 \end{aligned} \tag{3}$$

We normalize relative to the most likely parameter range of $\mu \in [-1,1], \sigma \in [0,0.6]$ and report 100× values for readability.

4.6 Analysis and Modeling

Participants completed the experiment in 18 minutes on average (*sd* = 8.7). They provided 5,328 responses in total. We used the brms package [7] to fit the responses to a Bayesian regression model. The model predicts $\Delta\mu$ or $\Delta\sigma$ (i.e., the divergence from true slope or uncertainty). We modeled both the *mean* and *spread (sd)* of the divergence. We followed a Bayesian analysis workflow in creating the model [24], evaluating with posterior predictive checks and incorporating additional parameters to improve the posterior fit. Details on the model construction process can be found in the supplementary material. The model for $\Delta\mu$ is given by (with a nearly identical formula to predict $\Delta\sigma$):

$$\begin{aligned} \Delta\mu &\sim Normal(\text{mean}, \text{sd}^2) \\ \text{mean} &= \Delta R \times agent \times vis \times consensus \times size \\ &\quad + \Delta R \times agent \times questionType \\ &\quad + (1 + \Delta R \times agent \mid participant) + (1 \mid question) \\ log(\text{sd}) &= agent \times vis \times consensus \times size \\ &\quad + (1 + size \times agent \mid participant) + (1 \mid question) \end{aligned}$$

ΔR is the sample extremeness (Equation 2). *Agent* indicates whether the inference came from a human (the participant), an informed Bayesian, or an uninformed flat-prior agent. *Size* is a categorical variable representing the sample size (small, medium, or large). *Vis* represents the visualization type (scatterplot, bar chart, or icon array). *Consensus* is the agreement around the ground truth. The first interaction term allows us to model our primary effects of interest. After conducting posterior predictive checks, we introduced an effect for *questionType*; this categorical variable signifies whether the ground truth prescribes a positive correlation, negative correlation, or no relationship – the latter played a role in modulating participants' responses. Lastly, we included random slopes and intercepts to model individual differences among participants and effects of topic variation (*question*). We used flat priors for all parameters, except for the spread which followed a half-student's T distribution ($df = 3$, $bias = 0$, $scale = 2.5$).

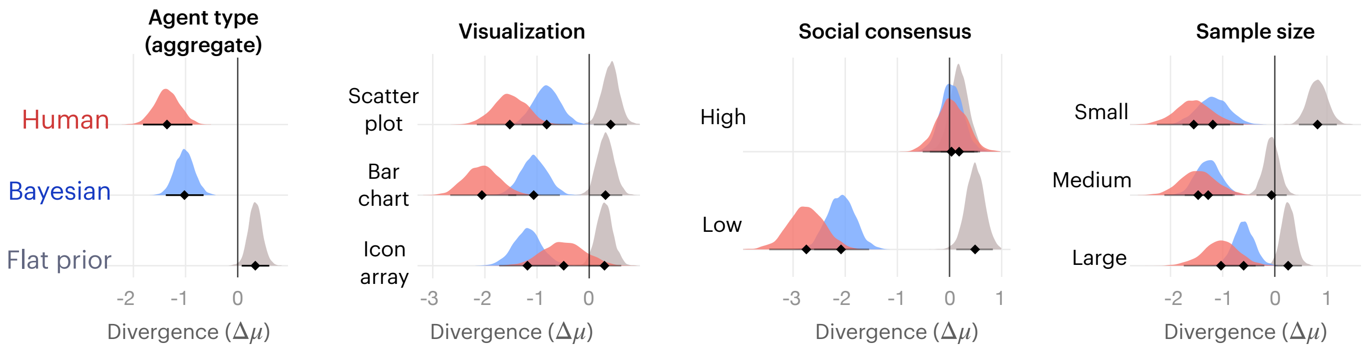

Fig. 4: Estimated mean divergence (± 95% credible intervals) for participants vs. informed and uninformed (flat prior) agents. Smaller divergence from zero indicates better accuracy at inferring the true slope. Right: marginalized effects of Visualization, Consensus, and Sample size.

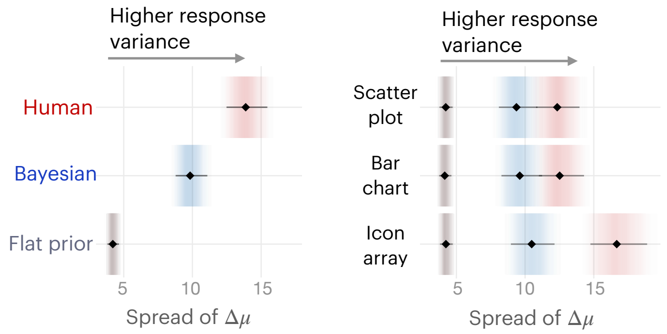

Fig. 5: Estimated spread (standard deviation) of $\Delta\mu$ (± 95% CIs) for participants (red) vs. informed and uninformed Bayesians. A larger spread implies higher variability in inference accuracy. The plot on the right shows the marginalized effects of visualization type.

4.7 Results

4.7.1 How accurate are participants in inferring the true slope compared to Bayesian agents?

Figure 4 illustrates the mean accuracy of three agents overall and across all manipulated factors. Zero divergence is ideal. Overall, the flat-prior agent exhibits the smallest distance to the real slope (est. divergence: 0.34, 95% credible intervals: $[0.08, 0.61]$). This was followed by the informed Bayesian agent (-1.02, 95% CI: $[-1.38, -.65]$), and lastly, participants who provided the least accurate inferences (-1.36, CI: $[-1.81, -.87]$). Humans and informed Bayesians overlapped substantially in their accuracy, although they both consistently underestimated the true relationship. By contrast, an uninformed Bayesian appears to provide near-optimal inference on the aggregate. We analyze how accuracy varies across visualization types and social consensus levels.

Visualization: Participants appear to perform best when viewing icon arrays. On average, their divergence from the true slope was only $-.49$ with a 95% credible interval that overlaps zero ($[-1.27, 0.3]$), indicating a good chance of inferring the true relationship. Notably, human performance with icon arrays slightly surpassed though remained comparable to that of an informed Bayesian (-1.18, CI: $[-1.73, -0.65]$). In contrast, participants tended to make less accurate inferences with scatterplots (-1.52, CI: $[-2.15, -0.9]$) and bar charts (-2.06, CI: $[-2.66, -1.4]$), with both visualizations leading to an underestimation of the true relationship.

Consensus: All agents performed equally well when the consensus around the ground truth was high, with divergence remarkably centered around zero. By comparison, low-consensus topics elicited less accurate inferences from participants (divergence: -2.75, CI: $[-3.46, -2.07]$). This reduction in accuracy was also observed to a lesser degree in the informed Bayesian (-2.08, CI: $[-2.6, -1.54]$). As would be expected, a flat-prior agent is virtually unaffected by whether there is ground consensus (0.18, CI: $[-0.17, 0.54]$ for high consensus vs. 0.49, CI: $[0.13, 0.84]$ for low).

Sample Size: The data-driven agent consistently demonstrated good inference across medium and large sample sizes, with credible intervals that intersected zero. As expected, its performance was reduced in small sample sizes, resulting in a non-zero deviation (0.82, CI: $[0.46, 1.19]$). This degree of deviation in smaller datasets was somewhat comparable to that of an informed Bayesian (-1.19, CI: $[-1.77, -0.6]$), and to human performance (-1.55, CI: $[-2.26, -0.86]$).

Spread: Figure 5 illustrates the estimated spread in $\Delta\mu$. Increased spread corresponds to higher variability in responses. Overall, the flat prior agent displays the least spread (est. SD: 4.21, CI $[3.81, 4.65]$), indicating a consistent response. By contrast, the informed Bayesian (9.83, CI: $[8.78, 11.1]$) exhibits higher variability. Participants, at the upper end, demonstrate the greatest spread (13.9, CI: $[12.5, 15.5]$), indicating higher variance in their ability to infer the true slope. Figure 5-right further dissects this by visualization type. Among participants, icon arrays led to the widest spread (16.7, CI: $[14.8, 19]$), consistently surpassing the other two visualizations (12.3, CI: $[10.8, 14]$ for scatterplot and 12.5, CI: $[11, 14.3]$ for bar charts).

4.7.2 How resilient are humans and Bayesian agents to spurious samples?

Figure 6-top illustrates the impact of sample extremeness on the divergence from the ground truth for the three agent types. A weaker correlation (i.e., smaller slope) indicates lower sensitivity and, hence, better resilience to spurious samples. Participants appear to be less influenced by extreme samples than both informed and uninformed (flat prior) Bayesian agents. Specifically, a unit increase in sample extremeness leads to a 14.3 (CI: $[12.2, 16.47]$) increase in participant divergence from the true relationship. By contrast, a Bayesian is influenced more strongly (18.80, CI: $[17.25, 20.27]$), giving rise to higher error with spurious samples. As would be expected, an agent with a flat prior learning purely from the data is influenced the most (31.41, CI: $[30.68, 32.1]$), leading to a strong association between sample quality and inference accuracy.

Visualization, Consensus, and Sample Size: Icon arrays afford the highest resiliency (est. slope: 12.98, CI: $[9.78, 16.55]$), followed by scatterplots (13.79, CI: $[10.94, 16.64]$) and bar charts (16.10, CI: $[13.41, 18.88]$). These slopes, however, overlap indicating largely similar effects for the different visualizations. When there is more consensus around the ground truth, resiliency is consistently better (10.62, CI: $[8.19, 13.04]$ vs. 17.95, CI: $[15.26, 20.79]$). Lastly, sample size also played a role, with participants least influenced by small samples (9.43, CI: $[7.24, 11.84]$), followed by medium (13.33, CI: $[10.59, 16.02]$) and large (20.18, CI: $[16.59, 23.74]$).

Comparing Human vs. Machine Resilience: Figure 7 illustrates the relative robustness of humans against statistical inference. Specifically, we estimate the log-divergence ratio for Bayesian agents over participants: $log(|\Delta\mu_{Bayes|Flat}/\Delta\mu_{Human}|)$. An estimate above 0 suggests superior inferential performance for humans compared to statistical machines, while a value below zero indicates the opposite. When $\Delta R = 0$ (a sample fully consistent with the ground truth), uninformed

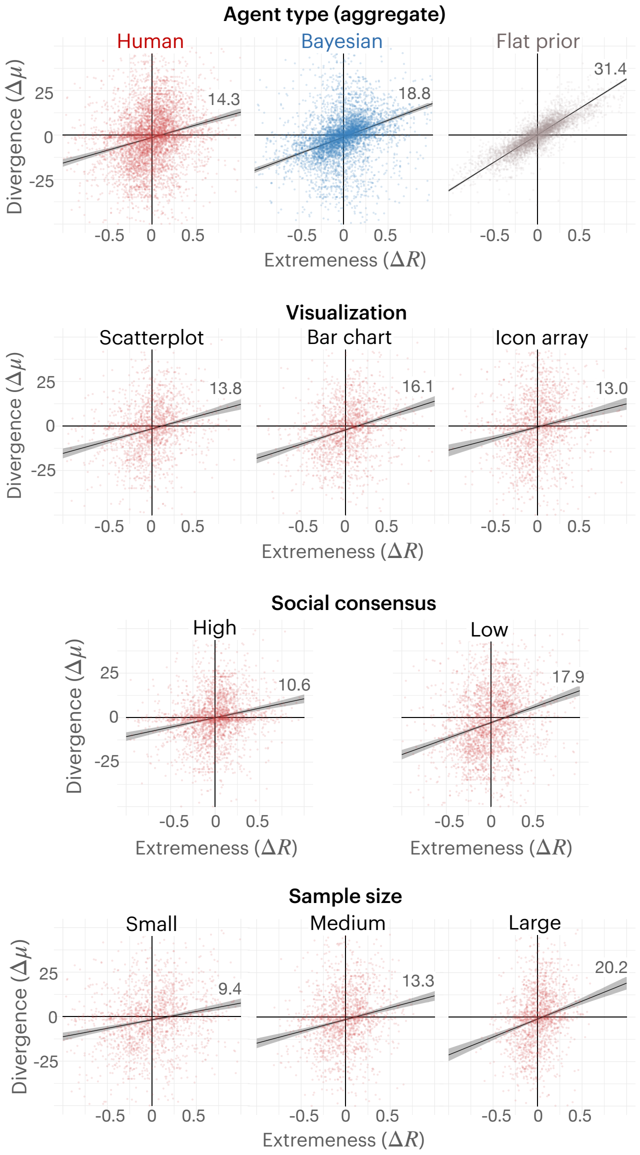

Fig. 6: Sensitivity to extreme samples for the three agents (top) and for participants alone (bottom rows). Lines depict mean estimate of $\Delta\mu/\Delta R$ slopes (± 95% CIs). A lower slope indicates better resilience to spurious samples. The point clouds depict the observed responses.

Bayesians perform 3.82 times better than humans (CI: $[2.57, 5.81]$). Similarly, informed Bayesian inference is expected to be 2.13 times closer to the real data-generating process than the average participant (CI: $[1.59, 2.97]$).

However, as data becomes more extreme, the model predicts an inversion of this relationship. For instance, at $\Delta R = 0.2$ (where 40.7% of the data has $|\Delta R| \geq 0.2$), humans exhibit similar performance to an informed Bayesian (1.12× advantage, CI: $[0.81, 1.52]$) while showing a substantial advantage over a flat-prior agent (2.97×, CI: $[2.31, 3.9]$). This advantage widens when $\Delta R = 0.4$ (with 14% of samples having $|\Delta R| \geq 0.4$). Here, an average human will reliably outperform both informed and uninformed Bayesians by factors of 1.73 (CI: $[1.33, 2.16]$) and 4.04 (CI: $[3.24, 5.06]$), respectively.

4.7.3 How good are the agents at characterizing uncertainty?

Figure 8-left displays the mean accuracy of three agents in characterizing σ_{Truth} (the uncertainty in the ground truth model). The flat prior agent demonstrates a more conservative and accurate inference of true uncertainty (-9.15, CI: $[-11.3, -7.12]$). In contrast, both humans and informed Bayesian agents seemed overconfident in their uncertainty estimation (33.1, CI: $[30.7, 35.5]$ for humans vs. 24.7, CI: $[22.4, 26.8]$ for a Bayesian). The mean accuracy for each visualization type (Figure 8-right) aligns with the aggregate estimate.

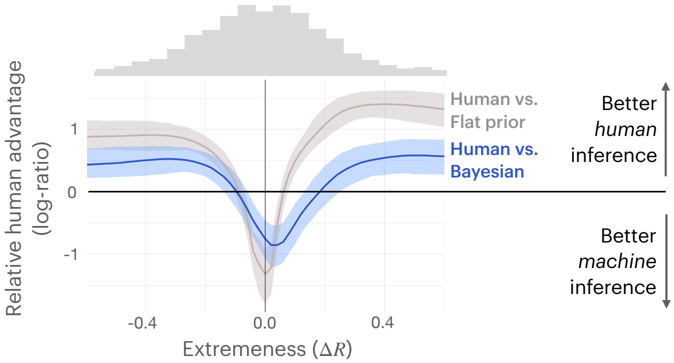

Fig. 7: Mean estimated advantage for participants over statistical machines (for inferring the true μ) subject to sample extremeness. Intervals are 95% CIs, with values above zero indicating a human advantage. The histogram (top) shows the empirically observed ΔR distribution.

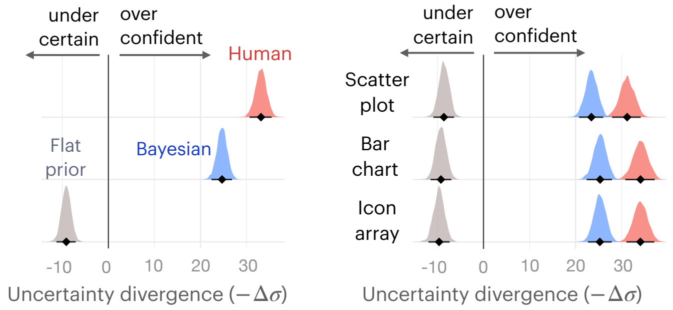

Fig. 8: Estimated mean accuracy (±95% CIs) in characterizing uncertainty at the aggregate level (left) and by visualization type.

4.8 Discussion

Overall, participants were worse at inferring the true slope than at least one of the two Bayesian agents. These results are partially consistent with H1. Surprisingly, however, the flat prior achieved the highest accuracy. This suggests that subject-provided priors were less reliable than the likelihood function implied by the sample. Despite being globally less optimal, participants exhibited better resiliency to spurious samples, especially in high-consensus topics and with smaller sample sizes. Interestingly, a cluster of zero-divergence human responses (Figure 6, particularly at high consensus) suggests that many participants 'correct' for extreme samples, leading to better inference. The model estimates that a typical participant will begin to outperform an informed Bayesian when $|\Delta R| \geq 0.2$ (40.7% of the data). The flat-prior agent, though best performing in the aggregate, was by comparison highly susceptible to spurious samples. These results support H2.

Surprisingly, participants demonstrated higher inference accuracy with icon arrays than scatterplots, despite the latter's perceptual superiority for bivariate data. This unexpected result (inconsistent with H3) may be attributed to the advantages of icon arrays in conditional probabilistic reasoning [54]. Such cognitive benefits might outweigh the perceptual advantages of scatterplots. However, it is noteworthy that the spread in performance was most pronounced in icon arrays, indicating a wider variability in response quality, even when this visualization tended to lead to more accurate inference on average. Lastly, and consistent with H4, participants were less accurate in characterizing model uncertainty compared to both Bayesian agents.

5 Experiment II

Exp. 1 shows that consensus, a proxy for the level of social agreement, strongly impacts participants' inferences. This factor could hypothetically influence *when* human analysts might resort to corrective,

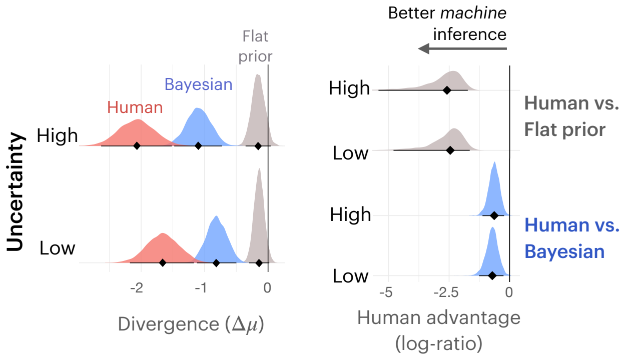

Fig. 9: Left: Mean estimated divergence ($\Delta\mu \pm$ 95% CI). Right: comparison of human and machine accuracy at each uncertainty level.

non-statistically normative heuristics, by drawing upon their socially rooted knowledge. In this experiment, we explore if this social information can compensate for uncertainty in the data-generating process. Specifically, in the presence of heightened statistical uncertainty, we investigate whether participants can leverage their domain intuition to attain a relatively better inference as compared to statistical agents.

5.1 Hypotheses

We expect participants to handle increased uncertainty better than Bayesian agents, particularly when they can compensate with social information:

H5 – The gap between participants and statistical agents will decrease under higher uncertainty. Participants, benefiting from an increased ratio of intuitive, social to statistical information, are expected to reduce the overall performance advantage for Bayesian agents seen in Exp. 1.

H6 – Participants will strategically employ corrective heuristics at higher rates when faced with both high consensus and high uncertainty, making them even more resilient to spurious samples. This behavior will manifest as an interaction between consensus and uncertainty.

5.2 Participants, Experimental Design, and Procedures

We recruited 148 participants from Prolific (95 males 50 females 3 others), with a mean age of 37.4 for a $5 compensation ($16.6/hour on average). Self-reported education levels were: 42 high school, 21 associate, 60 bachelor's, 17 master's, 5 doctorate, and 3 unspecified. The experiment design and procedures were similar to Exp. 1. However, this experiment only employs scatterplots. Participants completed 24 trials. Half the trials were sampled from a model with decreased uncertainty ($\frac{1}{2}\sigma_{Truth}$) and the other half were sampled from a higher-uncertainty model ($2\sigma_{Truth}$). Uncertainty was manipulated as a within-subject factor. We slightly reduced σ_e from 0.45 to 0.3 (see Equation 1) to allow for more uncertainty control via σ_{Truth}.

5.3 Results

Participants completed the experiment in 18.1 minutes on average (*sd* = 10.4). We analyzed the results by fitting a Bayesian regression model similar to Exp. 1 (§4.6), removing the visualization effect and replacing it with *uncertainty* as a factor (two levels: low and high). We first investigate how increasing (or decreasing) uncertainty affects inference accuracy ($\Delta\mu$) for the three agents. We then analyze the interaction between social consensus and uncertainty.

5.3.1 How does uncertainty in the data-generating process impact inference?

Figure 9-left shows the mean accuracy of slope inference for the three agents by uncertainty. Informed Bayesian agents and participants do slightly better when uncertainty is low. However, as depicted in Figure 9-right, the gap (represented by the log-ratio) between participants and the two statistical machines remains relatively consistent regardless of the uncertainty level in the data-generating process. Specifically, on average, the flat-prior agent is 11.7 times (CI: $[5.21, 121.51]$ better than humans when the uncertainty is low, and 13.33 times (CI: $[5.64, 225.88]$ better at high uncertainty. The informed Bayesian also exhibits better performance over humans, though at a lesser degree (2.04× advantage, CI: $[1.29, 3.53]$ at low uncertainty vs. 1.89×, CI: $[1.26, 3.06]$ at high uncertainty. Overall, the observed human-machine gap appears unaffected by changes in uncertainty.

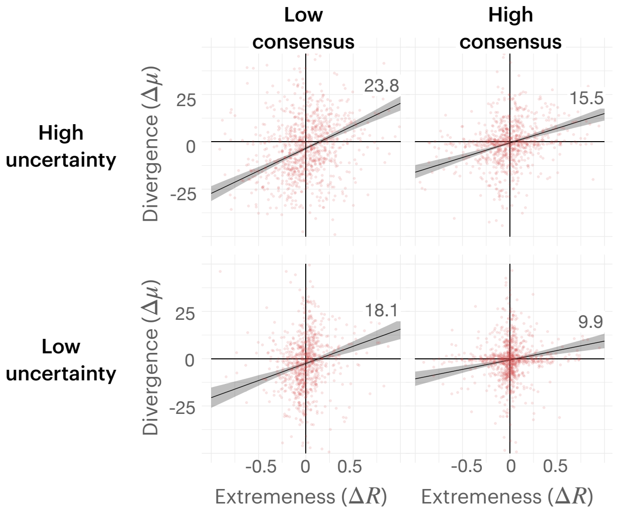

Fig. 10: Participants' sensitivity to sample extremeness, showing the interplay between social consensus and uncertainty. Lines show model estimates of $\Delta\mu/\Delta R$ (± 95% CIs). Points correspond to empirical responses. A weaker correlation implies more resilience to spurious data.

5.3.2 Can social information compensate for uncertainty?

Figure 10 illustrates how sample extremeness affects participants' inference across different levels of uncertainty and social consensus. In general, heightened uncertainty increases susceptibility to spurious samples, but this sensitivity seems to be moderated by social consensus. For instance, when uncertainty is high and the consensus is low, the $\Delta\mu/\Delta R$ rate reaches its peak at 23.8 (CI: $[19.96, 27.67]$). However, higher consensus reduces this rate to 15.48 (CI: $[11.76, 18.82]$). Likewise, lower uncertainty decreases the divergence-extremeness rate from 23.8 to 18.07 (CI: $[12.80, 23.33]$). The rate is at its minimum when uncertainty is low, but consensus is high (9.95, CI: $[6.17, 13.80]$).

We investigated the interplay between uncertainty and social consensus using the WAIC criterion [63]. A reduced model, which excludes an interaction effect, explained over 99% of the weights compared to the full model, indicating the absence of an interaction. Accordingly, the observed results can be attributed to a linear combination of social consensus and low uncertainty, both of which appear to enhance participants' ability to disregard spurious data.

5.4 Discussion

The results do not support the notion that participants use social information to compensate for statistical uncertainty. Instead, the results suggest a simpler explanation: participants benefit, separately, from increased social consensus and reduced uncertainty in the data-generating model. Consequently, we find no evidence to support an interaction as put forth by H6. Likewise, there is no indication that participants can narrow the advantage for statistical machines at higher uncertainty levels. On average, human inference performance remains less accurate than that of Bayesians, and this performance gap persists consistently across the two uncertainty levels (contrary to H5). Therefore, while social information may offer inferential benefits, it seems unlikely to aid visualization observers in reducing inherent statistical uncertainties.

6 General Discussion

6.1 Human versus Statistical Inference

Our findings suggest that human analysts and Bayesian machines excel in different scenarios. While Bayesian agents generally provide more optimal inference, humans showed less deviation from the ground truth with spurious data. For instance, at the 60th percentile of extreme data, humans matched informed Bayesians and outperformed flat prior agents by a factor of 3. This advantage increased at the 86th percentile, with factors of 1.7 and 4, respectively. Participants appear to rely on their internal models to disregard extreme samples, reducing bias. Even when the Bayesian agent is provided with a human prior model, the human inference remains less affected by extreme data.

These findings support our central hypothesis that analysts' intuitive and non-rational heuristics, such as ignoring or underweighting seemingly implausible data samples, can be advantageous. While overlooking improbable evidence can generally be problematic [8,41], this tendency may prove useful in exploratory data analysis, by making analysts more skeptical of patterns they encounter in visualizations, effectively reducing the rate of false discoveries from visualizations. For instance, when viewing a visualization suggesting a positive increase in sales that the analyst believes to be counterintuitive, the analyst may be justified in underweighting this evidence and concluding that there was no substantial increase in this case. Although such a conclusion may be non-statistically rational given the data at hand, it could prevent the analyst from committing a false discovery.

Prior work by Zgraggen et al. suggests that visual analysts frequently make Type I errors, such as concluding that there is an effect, like a positive bivariate correlation or a non-zero difference between two groups, when in reality there are none [70]. In contrast, our findings suggest that, on average, individuals may be more adept at discerning true effects from noise by using their internal models and heuristics to resist spurious patterns. This ability, however, appears contingent on having good intuition about the underlying phenomena. Moreover, our results indicate that formalizing this intuitive process, for example, by enforcing or encouraging universally rational (e.g., Bayesian) inference, could be counterproductive. Visual analytics systems should retain a role for analyst intuition, even when their conclusions may seem biased.

Despite the advantage at interpreting extreme data samples, human inference seems to exhibit higher variability. Individual analyst responses may thus be further from the truth, even though collectively, humans may be less prone to overinterpreting extreme samples. This variability may be attributed to noise from two sources: the perception of visualizations, and the setting of priors and posteriors through graphical elicitation devices. These channels allow for visual-perceptual, cognitive, and manual errors to seep in, thereby affecting response quality. Computational statistical agents, on the other hand, are unaffected by these noise sources, and may thus provide a less varied response. Similarly, in terms of inferring uncertainty, participants consistently exhibited overconfidence. The Bayesian agent's response, while also expressing higher certainty about the data-generating process than warranted, was closer to the ground truth. Conversely, the uninformed agent provided a more conservative assessment that was almost always closer to the true uncertainty in the generating model.

6.2 People-Machine Collaboration for Inferential Analysis

Our findings suggest the potential for collaboration between humans and statistical (or AI-based) agents for inference-making, enabling each to complement the other's limitations [61,66]. This collaboration could take the form of delegation [5,32,44,52], where each agent (human or statistical) takes responsibility for certain inferential tasks depending on the sample size and the quality of the human intuition. For example, systems can delegate the responsibility of making inferences from large datasets to a rational statistical agent. Conversely, humans could manually handle interpretive scenarios involving smaller and less ideal datasets, where analyst intuition are likely to provide an edge. One challenge to enabling this sort of collaborative inference-making is to predict the potential extremeness of the data at hand so that effective delegation can take place [21,55]. A second challenge is the need to reduce variability in individual human responses without compromising their advantages. We suspect that improving knowledge elicitation in visualizations (e.g., by redesigning and validating graphical model representation techniques [37]) could help reduce response variability by reducing noise in externalizing human priors and posteriors. The idea of using AI for certain aspects of the data analysis pipeline, such as automating the creation of machine learning models [18,65], has indeed been suggested. However, data professionals continue to express reservations about this model [67]. By contrast, a collaborative workflow that genuinely involves analysts could garner more positive reception over 'Auto Data Science' approaches [1,38].

6.3 The Utility of a Biased Mind

In visual analytics, the concept of mixed-initiative collaboration between humans on one hand and algorithms and statistical models on the other is well established [15, 19]. Recent efforts in this space aim to algorithmically identify deviations from statistical rationality [37,40], and introduce interventions to mitigate biases in human decision-making [3,9,17,64]. This view often holds that any deviation from idealized, analytic normativeness is problematic and should be reduced to a minimum. Our findings, however, suggest that analyst deviations in graphical inference could indeed be useful. Thus, rather than solely focusing on 'debiasing' human reasoning (e.g., by seeking to align human responses with rational models [69]), our work suggests the need for a more balanced perspective that acknowledges the utility of both rational and intuitive thinking [16,26,31,58], even if the latter is subject to biases. Instead of attempting to eliminate these biases in visual analytics, a more fruitful approach involves considering how to productively leverage analyst intuition. However, it is important to acknowledge that not all biases are useful, and some may lead to systematic errors and bad judgments, especially in response to perverse incentives (e.g., the pressure to infer a positive (non-null) conclusion from data for publication purposes [20,45,53]). With this understanding, systems can learn to trust human judgment and, conversely, to intervene with computational support when necessary, for example, by furnishing a (more) rational solution for human consideration.

7 Limitations and Future Work

There are limitations to consider when interpreting our results. First, our experiments utilized varying sample sizes, ranging from 'small' to 'large'. Yet, even the largest size ($N = 30$) presents a relatively small amount of data. Larger datasets, in particular, could lead to a greater advantage for statistical machines over humans than we observed in our study. Future studies should therefore include a wider range of sample sizes, including more realistically large datasets. Second, the ground-truth models we used were primarily based on common-sense semantics. These models do not necessarily reflect more contentious or expertise-dependent topics. Additionally, our study tested a limited range of visualizations (scatterplots, bar charts, and icon arrays), hence the results may not generalize to other representations. The involvement of crowdsourced participants, likely non-experts, is another limitation. Lastly, the workflow in our experiments was highly controlled. In reality, analysts are more likely to engage in fluid, open-ended visual exploration. Unlike participants, analysts also have the autonomy to choose which visualizations to view based on their priors. Future studies should therefore attempt to replicate our findings with practicing data scientists under more realistic tool usage conditions.

8 Conclusion

We investigated human inference-making from (at times noisy) visualizations, contrasting the accuracy of human judgments with those of Bayesian agents. Our central hypothesis was that, while humans exhibit non-rational tendencies, they may possess advantages in certain scenarios. The experiment findings support this notion: Although participants were generally less optimal than Bayesian benchmarks, they surpassed machines in specific instances and were better at discerning (and disregarding) spurious datasets. However, our results also indicate that humans exhibit higher variability and consistently exhibit overconfidence. Our findings suggest future design avenues where intuitive thinking on the part of analysts and statistical models can complement each other. Furthermore, our work challenges the assumption that normative rationality is always ideal and shows that analyst gut reactions could provide an advantage in certain visual analytic contexts.

Acknowledgments

This research was supported by the National Science Foundation under award 1942429, and by the Office of Science, US Department of Energy under contract DE-AC02-06CH11357. KR was also supported by the Argonne faculty sabbatical program.

References

[1] C. Aggarwal, D. Bouneffouf, H. Samulowitz, B. Buesser, T. Hoang, U. Khurana, S. Liu, T. Pedapati, P. Ram, A. Rawat, et al. How can ai automate end-to-end data science? *arXiv preprint arXiv:1910.14436*, 2019. doi: 10.48550/arXiv.1910.14436 9

[2] J. S. Albrechtsen, C. A. Meissner, and K. J. Susa. Can intuition improve deception detection performance? *Journal of Experimental Social Psychology*, 45(4):1052–1055, 2009. doi: 10.1016/j.jesp.2009.05.017 2

[3] M. A. Bedek, A. Nussbaumer, L. Huszar, and D. Albert. Methods for discovering cognitive biases in a visual analytics environment. *Cognitive Biases in Visualizations*, pp. 61–73, 2018. doi: 10.1007/978-3-319-95831-6_5 9

[4] R. Beecham, J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood. Map lineups: Effects of spatial structure on graphical inference. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):391–400, 2016. doi: 10.1109/TVCG.2016.2598862 2

[5] E. Bondi, R. Koster, H. Sheahan, M. Chadwick, Y. Bachrach, T. Cemgil, U. Paquet, and K. Dvijotham. Role of human-ai interaction in selective prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 5286–5294, 2022. doi: 10.1609/aaai.v36i5.20465 9

[6] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383, 2009. doi: 10.1098/rsta.2009.0120 2

[7] P.-C. Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28, 2017. doi: 10.18637/jss.v080.i01 5

[8] C. F. Camerer and H. Kunreuther. Decision processes for low probability events: Policy implications. *Journal of Policy Analysis and Management*, 8(4):565–592, 1989. doi: 10.2307/3325045 9

[9] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 116–126. IEEE, 2017. doi: 10.1109/VAST.2017.8585665 9

[10] I. K. Choi, T. Childers, N. K. Raveendranath, S. Mishra, K. Harris, and K. Reda. Concept-driven visual analytics: an exploratory study of model- and hypothesis-based reasoning with visualizations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019. doi: 10.1145/3290605.3300298 2

[11] I. K. Choi, N. K. Raveendranath, J. Westerfield, and K. Reda. Visual (dis) confirmation: Validating models and hypotheses with visualizations. In *2019 23rd International Conference in Information Visualization–Part II*, pp. 116–121. IEEE, 2019. doi: 10.1109/IV-2.2019.00032 2

[12] S. G. Coles and E. A. Powell. Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review/Revue Internationale de Statistique*, pp. 119–136, 1996. doi: 10.2307/1403426 2

[13] S. G. Coles and J. A. Tawn. A bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 45(4):463–478, 1996. doi: 10.2307/2986068 2

[14] V. Crupi and F. Calzavarini. Critique of pure bayesian cognitive science: A view from the philosophy of science. *European Journal for Philosophy of Science*, 13(3):28, 2023. doi: 10.1007/s13194-023-00533-w 1

[15] W. Cui. Visual analytics: A comprehensive overview. *IEEE Access*, 7:81555–81573, 2019. doi: 10.1109/ACCESS.2019.2923736 9

[16] E. Dane, K. W. Rockmann, and M. G. Pratt. When should i trust my gut? linking domain expertise to intuitive decision-making effectiveness. *Organizational behavior and human decision processes*, 119(2):187–194, 2012. doi: 10.1016/j.obhdp.2012.07.009 2, 9

[17] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic. A task-based taxonomy of cognitive biases for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(2):1413–1432, 2018. doi: 10.1109/TVCG.2018.2872577 9

[18] J. Drozdal, J. Weisz, D. Wang, G. Dass, B. Yao, C. Zhao, M. Muller, L. Ju, and H. Su. Trust in automl: exploring information needs for establishing trust in automated machine learning systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pp. 297–307, 2020. doi: 10.1145/3377325.3377501 9

[19] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews. The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems*, 43:411–435, 2014. doi: 10.1007/s10844-014-0304-9 9

[20] W. Forstmeier, E.-J. Wagenmakers, and T. H. Parker. Detecting and avoiding likely false-positive findings–a practical guide. *Biological Reviews*, 92(4):1941–1968, 2017. doi: 10.1111/brv.12315 9

[21] A. Fügener, J. Grahl, A. Gupta, and W. Ketter. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696, 2022. doi: 10.1287/isre.2021.1079 9

[22] M. Galesic, R. Garcia-Retamero, and G. Gigerenzer. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health psychology*, 28(2):210, 2009. doi: 10.1037/a0014474 4

[23] R. Garcia-Retamero, M. Galesic, and G. Gigerenzer. Do icon arrays help reduce denominator neglect? *Medical Decision Making*, 30(6):672–684, 2010. doi: 10.1177/0272989X10369000 4

[24] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020. doi: 10.48550/arXiv.2011.01808 5

[25] G. Gigerenzer. Why heuristics work. *Perspectives on psychological science*, 3(1):20–29, 2008. doi: 10.1111/j.1745-6916.2008.00058.x 2

[26] G. Gigerenzer and H. Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143, 2009. doi: 10.1111/j.1756-8765.2008.01006.x 1, 9

[27] G. Gigerenzer and W. Gaissmaier. Heuristic decision making. *Annual review of psychology*, 62:451–482, 2011. doi: 10.1146/annurev-psych-120709-145346 2

[28] G. Gigerenzer and P. M. Todd. Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart*, pp. 3–34. Oxford University Press, 1999. 2

[29] T. Gilovich, D. Griffin, and D. Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002. doi: 10.1017/CBO9780511808098 2

[30] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014. doi: 10.1109/TVCG.2014.2346979 4

[31] M. G. Haselton, G. A. Bryant, A. Wilke, D. A. Frederick, A. Galperin, W. E. Frankenhuis, and T. Moore. Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition*, 27(5):733–763, 2009. doi: 10.1521/soco.2009.27.5.733 9

[32] P. Hemmer, M. Westphal, M. Schemmer, S. Vetter, M. Vössing, and G. Satzger. Human-ai collaboration: The effect of ai delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 453–463, 2023. doi: 10.1145/3581641.3584052 9

[33] J. Heyer, N. K. Raveendranath, and K. Reda. Pushing the (visual) narrative: the effects of prior knowledge elicitation in provocative topics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020. doi: 10.1145/3313831.3376887 2

[34] L. Huang. The role of investor gut feel in managing complexity and extreme risk. *Academy of Management Journal*, 61(5):1821–1847, 2018. doi: 10.5465/amj.2016.1009 2

[35] D. Kahneman, P. Slovic, and A. Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982. 2

[36] A. Kale, Z. Guo, X. L. Qiao, J. Heer, and J. Hullman. Evm: Incorporating model checking into exploratory visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2023. doi: 10.1109/TVCG.2023.3326516 2

[37] A. Karduni, D. Markant, R. Wesslen, and W. Dou. A bayesian cognition approach for belief updating of correlation judgement through uncertainty visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):978–988, 2020. doi: 10.1109/TVCG.2020.3029412 1, 2, 9

[38] S. K. Karmaker, M. M. Hassan, M. J. Smith, L. Xu, C. Zhai, and K. Veeramachaneni. Automl to date and beyond: Challenges and opportunities. *ACM Computing Surveys (CSUR)*, 54(8):1–36, 2021. doi: 10.1145/3470918 9

[39] Y.-S. Kim, P. Kayongo, M. Grunde-McLaughlin, and J. Hullman. Bayesian-assisted inference from visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):989–999, 2020. doi: 10.1109/TVCG.2020.3028984 2

[40] Y.-S. Kim, L. A. Walls, P. Krafft, and J. Hullman. A bayesian cognition approach to improve data visualization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019. doi: 10.1145/3290605.3300912 1, 2, 9

[41] J. J. Koehler and L. Macchi. Thinking about low-probability events: An exemplar-cuing theory. *Psychological Science*, 15(8):540–546, 2004. doi: 10.1111/j.0956-7976.2004.00716.x 9

[42] R. Koonchanok, P. Baser, A. Sikharam, N. K. Raveendranath, and K. Reda. Data prophecy: Exploring the effects of belief elicitation in visual analytics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2021. doi: 10.1145/3411764.3445798 2

[43] R. Koonchanok, G. Y. Tawde, G. R. Narayanasamy, S. Walimbe, and K. Reda. Visual belief elicitation reduces the incidence of false discovery. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2023. doi: 10.1145/3544548.3580808 2

[44] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, and C. Tan. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2022. doi: 10.1145/3491102.3501999 9

[45] M. D. Lindner, K. D. Torralba, and N. A. Khan. Scientific productivity: An exploratory study of metrics and incentives. *PloS one*, 13(4):e0195321, 2018. doi: 10.1371/journal.pone.0195321 9

[46] S. Luan, J. Reb, and G. Gigerenzer. Ecological rationality: Fast-and-frugal heuristics for managerial decision making under uncertainty. *Academy of Management Journal*, 62(6):1735–1759, 2019. doi: 10.5465/amj.2018.0172 2

[47] D. I. MacKenzie, J. D. Nichols, N. Sutton, K. Kawanishi, and L. L. Bailey. Improving inferences in population studies of rare species that are detected imperfectly. *Ecology*, 86(5):1101–1113, 2005. doi: 10.1890/04-1060 2

[48] S. Mahajan, B. Chen, A. Karduni, Y.-S. Kim, and E. Wall. Vibe: A design space for visual belief elicitation in data journalism. In *Computer Graphics Forum*, vol. 41, pp. 477–488. Wiley Online Library, 2022. doi: 10.1111/cgf.14556 2

[49] M. Majumder, H. Hofmann, and D. Cook. Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, 108(503):942–956, 2013. doi: 10.1080/01621459.2013.808157 2

[50] D. Markant, M. Rogha, A. Karduni, R. Wesslen, and W. Dou. When do data visualizations persuade? the impact of prior attitudes on learning about correlations from scatterplot visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2023. doi: 10.1145/3544548.3581330 2

[51] L. Micallef, P. Dragicevic, and J.-D. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2536–2545, 2012. doi: 10.1109/TVCG.2012.199 4

[52] H. Mozannar, A. Satyanarayan, and D. Sontag. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 5323–5331, 2022. doi: 10.48550/arXiv.2111.11297 9

[53] B. A. Nosek, J. R. Spies, and M. Motyl. Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, 2012. doi: 10.1177/1745691612459058 9

[54] A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. Han, and R. Chang. Improving bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE transactions on visualization and computer graphics*, 22(1):529–538, 2015. doi: 10.1109/TVCG.2015.2467758 7

[55] M. Pinski, M. Adam, and A. Benlian. Ai knowledge: Improving ai delegation through human enablement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2023. doi: 10.1145/3544548.3580794 9

[56] K. Reda, A. E. Johnson, M. E. Papka, and J. Leigh. Modeling and evaluating user behavior in exploratory visual analysis. *Information Visualization*, 15(4):325–339, 2016. doi: 10.1177/1473871616638546 2

[57] M. Rogha, S. Sah, A. Karduni, D. Markant, and W. Dou. The impact of elicitation and contrasting narratives on engagement, recall and attitude change with news articles containing data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2024. doi: 10.1109/TVCG.2024.3355884 2

[58] E. Sadler-Smith and E. Shefy. The intuitive executive: Understanding and applying 'gut feel'in decision-making. *Academy of Management Perspectives*, 18(4):76–91, 2004. doi: 10.5465/ame.2004.15268692 2, 9

[59] R. Savvides, A. Henelius, E. Oikarinen, and K. Puolamäki. Significance of patterns in data visualisations. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1509–1517, 2019. doi: 10.1145/3292500.3330994 2

[60] R. Savvides, A. Henelius, E. Oikarinen, and K. Puolamäki. Visual data exploration as a statistical testing procedure: Within-view and between-view multiple comparisons. *IEEE Transactions on Visualization and Computer Graphics*, 2022. doi: 10.1109/TVCG.2022.3175532 2

[61] M. Steyvers, H. Tejeda, G. Kerrigan, and P. Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022. doi: 10.1073/pnas.2111547119 9

[62] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124 1

[63] A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432, 2017. doi: 10.1007/s11222-016-9696-4 8

[64] E. Wall, L. M. Blaha, L. Franklin, and A. Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 ieee conference on visual analytics science and technology (vast)*, pp. 104–115. IEEE, 2017. doi: 10.1109/VAST.2017.8585669 9

[65] D. Wang, J. Andres, J. D. Weisz, E. Oduor, and C. Dugan. Autods: Towards human-centered automation of data science. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2021. doi: 10.1145/3411764.3445526 9

[66] D. Wang, E. Churchill, P. Maes, X. Fan, B. Shneiderman, Y. Shi, and Q. Wang. From human-human collaboration to human-ai collaboration: Designing ai systems that can work together with people. In *Extended abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2020. doi: 10.1145/3334480.3381069 9

[67] D. Wang, J. D. Weisz, M. Muller, P. Ram, W. Geyer, C. Dugan, Y. Tausczik, H. Samulowitz, and A. Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–24, 2019. doi: 10.1145/3359313 9

[68] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):973–979, 2010. doi: 10.1109/TVCG.2010.161 2

[69] Y. Wu, Z. Guo, M. Mamakos, J. Hartline, and J. Hullman. The rational agent benchmark for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2023. doi: 10.1109/TVCG.2023.3326513 9

[70] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2018. doi: 10.1145/3173574.3174053 1, 2, 9

[71] Z. Zhao, L. De Stefani, E. Zgraggen, C. Binnig, E. Upfal, and T. Kraska. Controlling false discoveries during interactive data exploration. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pp. 527–540, 2017. doi: 10.1145/3035918.3064019 2