

Could Humans Outshine AI in Visual Data Analysis?

RATANOND KOONCHANOK, Indiana University Indianapolis, USA

KHAIRI REDA, Indiana University Indianapolis, USA

People often use visualizations not only to explore a dataset but also to draw generalizable conclusions about underlying models or phenomena. While previous research has viewed deviations from rational analysis as problematic, we hypothesize that human reliance on non-normative heuristics may be advantageous in certain situations. In this study, we investigate scenarios where human intuition might outperform idealized statistical rationality. Our experiment assesses participants' accuracy in characterizing the parameters of known data-generating models from bivariate visualizations. Our findings show that, while participants generally demonstrated lower accuracy than statistical models, they often outperformed Bayesian agents, particularly when dealing with extreme samples. These results suggest that, even when deviating from rationality, human gut reactions to visualizations can provide an advantage. Our findings offer insights into how analyst intuition and statistical models can be integrated to improve inference and decision-making, with important implications for the design of visual analytics tools.

CCS Concepts: • **Human-centered computing** → **empirical studies**.

Additional Key Words and Phrases: Visualization, human-AI collaboration, decision-making

ACM Reference Format:

Ratanond Koonchanok and Khairi Reda. 2024. Could Humans Outshine AI in Visual Data Analysis?. In *TREW@CHI'24: Workshop on Trust and Reliance in Evolving Human-AI Workflows, May 11, 2024, Honolulu, HI*. ACM, New York, NY, USA, 11 pages. <https://doi.org/XXXXXXXX.XXXXXXX>

1 INTRODUCTION

Visualization tools play an increasingly important role in data analysis and decision-making. With recent advances in LLM applications, people can quickly generate visualizations even without having an analytical background. One significant impact of those advanced AI tools such as ChatGPT on data visualization is their ability to enable conversational queries and commands. Users can interact with visualizations using plain language, asking questions, or requesting specific visualizations. There has also been a growing number of proposals for visualization-specific tools that utilize LLMs as their core component [19, 21, 29]. While such advanced tools provide assistance in seeking insights from data [28, 31], making accurate inferences from visualizations can still be challenging, as it requires one to carefully account for potential uncertainties and variabilities in the data. For example, the analyst might overinterpret a visualization displaying an unusual or outlying sample, causing them to infer spurious features that do not exist in the data-generating process (i.e., false discovery). Conversely, they may fail to sufficiently consider data that is in front of them, and instead fall back to their prior belief – a potential manifestation of confirmation bias.

We are motivated by the idea of collaborative exploration and interpretation of visualizations, where human cognition works alongside AI systems to derive insights from complex datasets. Human analysts often bring to bear their own hunches when interpreting visualizations. In coming to an inference, the analyst might choose to adaptively overweigh their prior information and underweigh the evidence implied by a dataset, particularly if the latter is perceived to be

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2024 Copyright held by the owner/author(s).

Manuscript submitted to ACM

noisy or unreliable [20]. Such intuitive thinking is often regarded as untrustworthy and problematic [9, 12]. However, recent research has highlighted the role of heuristics in promoting accurate decisions [1, 7]. Intuitive decision-making can sometimes lead to better outcomes than rational, analytical reasoning [6, 23], particularly in high-risk high-uncertainty situations [11]. Non-statistically rational thinking could, thus, be useful when viewing a visualization with high uncertainty or 'extreme' data. In essence, human inference-making can serve as a valuable complement to a strictly rational agent.

In this study, we investigate situations where human visual inferences might outperform those of an ideal AI machine in accurately characterizing data-generating processes under varying uncertainty and sample conditions. Rather than assuming that inferences from an AI are always ideal, we compare the benefits and drawbacks of both sides, evaluating conditions in which the strength of one can compensate for the weakness of the other. To achieve this objective, we employ Bayesian models to assess how both humans and AI machines make inferences on datasets. Bayesian models offer a principled framework for incorporating prior beliefs and updating them based on observed evidence, making them suitable for analyzing complex datasets and assessing decision-making processes. Previous research has explored the application of Bayesian cognition approaches to data interpretation through visualization [13, 15] where they evaluated the effectiveness of visualization techniques based on the alignment between human interpretations and the Bayesian model's conclusions. While this method can be helpful in assessing overall effectiveness, relying solely on a model as a reliable reference may not always be ideal. In certain contexts, deviations from AI's rationality can be advantageous. For instance, a model might perceive a minor positive correlation between two unrelated variables, such as the number of ice cream cones sold and the number of shark attacks. In contrast, a human analyst might become more skeptical upon observing the same data, irrationally reinforcing their belief in no correlation, eventually avoiding a false-positive conclusion. This perspective acknowledges that human heuristics, though often simplistic and seemingly inferior [26], can lead to more accurate, ecological inferences [8].

By characterizing factors that lead analysts to exceed machine-inference performance, we lay the foundation for designing human-in-the-loop systems capable of leveraging both the intuition of analysts and the computational power of machines to harness insights from visualizations. These systems could integrate analyst intuition, such as through natural language queries, into their decision-making processes, enhancing their ability to provide meaningful insights and recommendations. To explore the opportunities, we ask the following research questions.

RQ1: How do human visual analysts compare to Bayesian agents that see that same data? We specifically compare the visual inferences of analysts to Bayesians: one informed with the prior knowledge of the human analyst, and another with a uniform prior that provides no preexisting knowledge before exposure to the data.

RQ2: How do characteristics of data samples affect visual inference accuracy relative to Bayesian agents?

We address these questions empirically in two crowdsourced studies. Specifically, we assess how accurately individuals perceive the true bi-variate relationship between pairs of attributes, upon being exposed to (potentially noisy) samples. We collect responses from viewers both before and after exposure to visualizations, allowing us to capture both their prior beliefs and posterior inferences about the data-generating process. We then measure the accuracy of human-vs-Bayesian inferences under different sample configurations, ranging from large samples that closely conform to the true parameters to small and potentially extreme samples. The work presented in this paper is a preliminary version of the research described in Koonchanok et al. [17], where we provide a more detailed analysis and additional experiments.

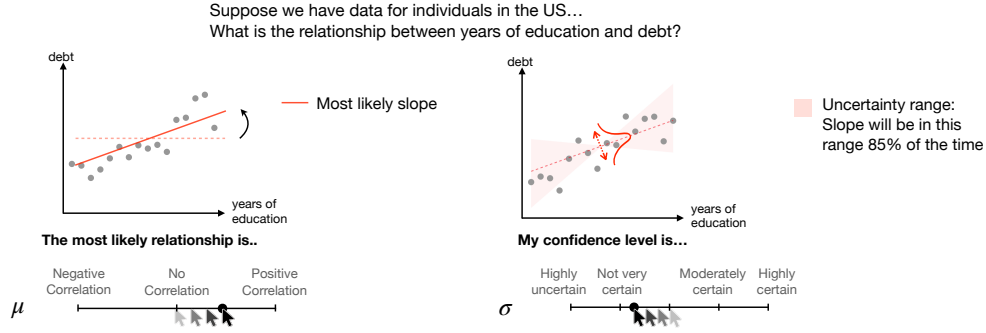


Fig. 1. Elicitation Method. Participants utilize the first slider to adjust the correlation line (left). They then use the second slider to adjust the uncertainty range (right). Data points on the plot are refreshed at the frequency of 5 Hz to allow participants to see the model implication.

2 METHODS

We address our research questions empirically in a crowdsourced study where we assess how accurately individuals perceive the true bi-variate relationship between pairs of attributes when exposed to (potentially noisy) samples. We collect responses from viewers both before and after exposure to visualizations, allowing us to capture both their prior beliefs and post-sample (posterior) inferences about the data-generating process. We then measure the accuracy of human-vs-Bayesian inferences under different sample extremeness, ranging from samples that closely conform to the true parameter to the extreme, potentially misleading samples that arise sporadically.

2.1 Prior and Posterior Inference Elicitation

We capture participants’ prior and posterior beliefs through a graphical inference method, as illustrated in Figure 1. Previous research has shown that eliciting priors can be beneficial in multiple aspects [3, 4, 10, 14, 16, 18]. This interface enables participants to articulate two key parameters regarding their beliefs: the most likely true correlation coefficient between the two variables (μ) and the associated uncertainty (σ). Participants enter their (prior or posterior) beliefs about these two parameters using two sliders. The first prompts them 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 uncertain to ‘highly certain’. These two parameters are then used to populate the following model:

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

Where $\mu \in (-1, 1)$ is the *expected slope* of the relationship as specified by the relationship slider, and $\sigma \in (0, 0.6)$ is the *uncertainty* in the slope, specified in the second confidence slider. β_0 , an intercept for the regression line, centered around 0 with a fixed standard deviation of $\sigma_b = 0.1$. σ_e is a standard deviation of an additional residual term (ϵ_i) and is

Correlation	Question	Ground Truth
Positive	What is the relationship between teachers' experience and the average standardized test score for students?	$\mu = 0.502$ $\sigma = 0.172$
Negative	What is the relationship between time spent on social media and hours of sleep at night?	$\mu = -0.422$ $\sigma = 0.146$
No relationship	What is the relationship between the length of people's first name and last name?	$\mu = 0.030$ $\sigma = 0.236$

Table 1. Examples from the prompt questions used in our experiments. We collected a mean crowdsourced ground truth of the μ to determine both the correlation of and consensus level of each question.

fixed at 0.45. We also display a hypothetical outcome plot (HOP), during which a new set of data points is refreshed every 200 milliseconds.

2.2 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 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 question. We then selected 24 prompt questions, based on the topic variety, to be used in the experiment. These questions comprised six ground truth models with a positive relationship ($\mu > 0$), six questions with a negative relationship ($\mu < 0$), and 12 with no correlation ($\mu \approx 0$).

2.2.1 Social consensus. Individuals often rely on social knowledge when forming their beliefs. Social 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 among the workers. Within each category (positive, negative, or no relationship), we designate half the 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.

2.3 Sample Properties

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

2.3.1 Sample Size. The number of observations in a sample is an important factor for an analyst to consider when making inferences. Larger sample sizes generally supply stronger evidence about the underlying data-generating sample. 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 sample sizes were selected to allow for 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.

2.3.2 Sample Extremeness. A core question that we address in this work is how resilient people are to extreme (or spurious) samples, as compared with idealized statistical machines. Sample extremeness reflects the degree to which it is consistent with the ground truth model. Under a random sampling regime, spurious samples are expected to naturally arise in the study. Consequently, we do not explicitly manipulate sample extremeness. Instead, we simply quantify and record the extremeness of generated stimuli in every trial. Under the law of large numbers, we expect samples generated to conform to the ground truth on the aggregate. In particular, the majority of samples should reflect the expected correlation coefficient of the corresponding ground truth model. However, extreme and spurious would be expected to emerge in the experiment (e.g., a sample showing a positive correlation when the underlying ground truth indicates no relationship).

We use Pearson’s coefficient (R) to characterize the strength of correlation for a sample and, by extension, the degree to which it can be considered *extreme* with respect to its underlying ground truth model. Specifically, we compute a difference (ΔR) between the sample’s correlation and the expected strength of the correlation:

$$\Delta R = R_{Sample} - R_{Expected} \quad (2)$$

Where R_{Sample} is the sample’s correlation coefficient and $R_{Expected}$ is the mean R-value, determined from 10,000 simulated draws from the ground truth model. A sample that is a faithful representation of the data-generating model will have $\Delta R \approx 0$. As ΔR deviates (either 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. Furthermore, we would expect the sample size to affect the likelihood of extreme samples: a sample with a larger number of data points will exhibit a narrower ΔR distribution. Conversely, samples with fewer data points will exhibit a wider extremeness distribution. Data-generating models with high uncertainty (σ) will also exhibit a greater likelihood of extreme samples that deviate from the expected slope.

By measuring sample extremeness during the study, we can observe participants’ reactions to spurious samples when those arise. Furthermore, we can gauge the degree to which participants (and Bayesian agents) are resilient to those misleading samples, and whether either can recover the underlying ground truth parameters.

3 EXPERIMENT

The experiment aims to compare the inference accuracy between participants to ideal Bayesian agents. We specifically test the extent to which participants can infer the ground truth parameters as a function of sample size, sample extremeness, and the degree of social consensus (along with interactions of these factors). Participants saw prompt questions described in §2.2, and were asked to predict the parameters of a bi-variate relationship model, namely, the expected slope μ and uncertainty σ . In each stimulus, participants were asked to provide their prior beliefs through the graphical elicitation interface described earlier. Specifically, participants adjusted the two sliders in Figure 1, until the observed HOP sample matched their expectation of the relationship. Following the prior specification, they were then

shown a data sample drawn randomly from the ground truth model side-by-side with their prior belief. During this step, they were asked to indicate how reliable they believed the sample to be, providing their subjective rating via a slider. Upon inspecting the sample, participants were prompted to provide their updated (posterior) belief, which reflect their inference “about the true relationship” via the same elicitation device. We specifically instructed participants to re-adjust the two parameter sliders again after considering the sample. The entire steps are shown in Figure 2.

3.1 Experiment Design

We adopt a mixed design, investigating two independent factors (varied within subjects): Sample size (3 levels) \times Social consensus (2 levels).

Sample characteristics: We varied sample characteristics within-subject. Specifically, participants completed an equal number of 8 trials with each sample size (small, medium, large). Participants also saw an equal number of low and high-consensus prompts. Sample extremeness was included in our analysis as an explanatory variable. Extremeness was left to vary as a consequence of the random sampling process. Hence, while not controlled on a participant level, extremeness followed a predictable distribution and was used as one of the explanatory factors in our modeling and analysis of the results (see §2.3 for a discussion of this).

Inferences: Participants supplied four continuous responses in every trial: the expected slope (μ) and uncertainty (σ) in the relationship, both before (i.e., prior belief) *and* after sample exposure (i.e., posterior inference). For every trial, we generated two comparable posterior 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 same sample observed by the participant. This corresponds to an idealized agent equipped with identical prior knowledge as the participant. The second response, also a Bayesian agent, utilized a flat prior in conjunction with the same sample-implied likelihood, thus representing a ‘blank-slate’ agent that exclusively learns the two parameters (μ_{Flat} and σ_{Flat}) purely from the sample. We measured the distance between the ground truth parameters (i.e., true μ and σ) and the posterior separately for each of the three agents (Participant, Bayesian, and Flat Bayesian).

3.2 Participants

Potential participants were recruited through Prolific. We ended up recruiting 74 participants. We recruited workers who are US residents with a minimum task-approval rate of 98%. Participants received a \$5 compensation upon completing the experiment. The study was approved by Indiana University’s institutional review board.

3.3 Procedure

Before starting the experiment, participants went through a series of tutorials explaining the goal of the study and explained how the graphical elicitation device worked. The instructions emphasized that the samples shown could be noisy, particularly when the sample size is small. Participants then completed a practice trial. After the tutorial, participants completed the main study which consisted of 24 trials, corresponding to the prompt questions developed in §2.2. In addition to the analyzed trials, we included four engagement checks inserted randomly within the trial stream. The engagement questions consisted of questions with the same setup as the rest but with different instructions asking participants to move sliders to a specific location (e.g. extreme right).

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 rate of affordable housing?) The participant was asked to visually provide their prior belief

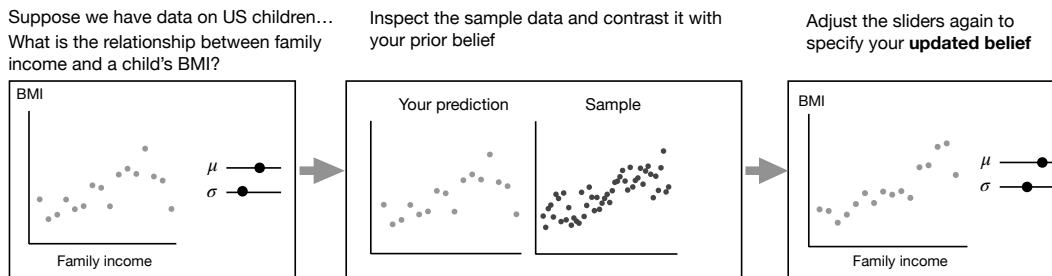


Fig. 2. Steps per trial that each participant. participants start by specifying initial belief of μ and σ using the two sliders. They then observe the data sample and compare it against their prediction. Subsequently, they specify their posterior beliefs using the same interface as the first step.

through the elicitation device described in §2.1. Next, the participants were presented with a sample drawn from the ground-truth model, shown side-by-side with their prior. Lastly, the participant was prompted to update their prior belief by re-adjusting the same set of two sliders. Figure 2 illustrates the sequence of these steps for a trial.

3.4 Accuracy Metrics

To characterize the accuracy of posterior inferences, we use a metric that measures the divergence between the posterior and the true values. For readability, we normalize to obtain relative maximum/minimum deviation based on the range of parameters dictated by the sliders $([-1, 1])$. Specifically:

$$\Delta\mu = \frac{\mu_{Human|Bayes|Flat} - \mu_{Truth}}{2} \tag{3}$$

The *Human* response is the posterior inference of the participant, whereas *Bayes* and *Flat* are the normative Bayesian inferences that are either based on the participant-supplied prior or an assumption of a flat prior, respectively. In the subsequent results, we multiplied the number by 100 to represent percentage values.

4 RESULTS

We used the *brms* package [2] to fit the responses to a Bayesian regression model. The model predicts $\Delta\mu$. We modeled both the *mean* and *variance* of the divergence, although the mean is our primary interest. We started by constructing a simpler model formulation that included the primary experimental factors. We then followed a model-expansion approach, adding interaction effects to improve the model fit (assessed using posterior predictive simulations). We also added random effects to account for individual differences among participants.

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

ΔR is the sample extremeness (as defined in Equation 2), *agent* indicates whether the inference came from a Human (the participant), a Bayesian, or flat-prior agent. *Size* is a categorical variable representing the sample size (Small,

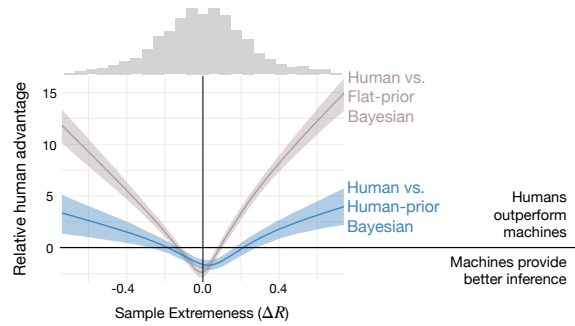


Fig. 3. Comparison between humans and statistical machines in inferring true μ , contingent on sample extremeness. The top histogram illustrates the empirical distribution observed at various ΔR levels.

Medium, or Large), *consensus* is the social consensus around the ground truth, and *questionType* is a categorical variable indicating whether the true model prescribes positive correlation, negative correlation, or no relationship.

Participants completed the experiment in 16.39 minutes on average. They provided 1776 responses in total. We first compare how participants perform compared with the Bayesian machines. We then analyze how each experimental factors affect the inference accuracy.

4.1 Human vs machine inferences

Figure 3 illustrates the relative resilience of humans compared to statistical inference. Specifically, we depict the estimated difference in divergence between participants on one hand and informed Bayesian agents (blue) and uninformed Bayesian agents (grey) on the other. An estimate above 0 suggests superior inferential performance for humans compared to statistical machines, while a value below zero indicates the opposite. When $\Delta R = 0$ (indicating a sample fully consistent with the ground truth), the model predicts that humans will underperform, showing a deficit of -1.62 (CI: $[-2.06, -1.14]$) against informed Bayesian agents and -2.37 (CI: $[-2.90, -1.84]$) relative to an uninformed, data-driven agent. At $\Delta R = 0.2$, however, the model predicts nearly equal performance for humans and informed Bayesian agents and a substantial advantage for humans over an uninformed agent. The performance advantage for humans can be expected to further widen at $\Delta R = .4$, relative to informed Bayesian.

4.2 To what extent are humans resilient to misleading samples compared to Bayesian agents?

Figure 4-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 13.53 (95% CI: $[11.43, 15.82]$) increase in divergence from the true relationship. By contrast, a Bayesian agent is influenced more strongly by misleading samples (18.61, CI: $[17.1, 20.16]$), leading to higher divergence at more extreme samples. As would be expected, an agent with a flat prior learning purely from the data is influenced the most (31.73, CI: $[30.96, 32.52]$), leading to a strong association between sample quality and inference accuracy. When there is more consensus around the ground truth, participants are more resilient to spurious samples (9.65, CI: $[7.32, 12.11]$ for high consensus vs., 17.46 CI: $[14.82, 20.29]$ for low consensus). Lastly, sample size also plays a role, with participants most skeptical of small samples (8.57, CI: $[6.36,$

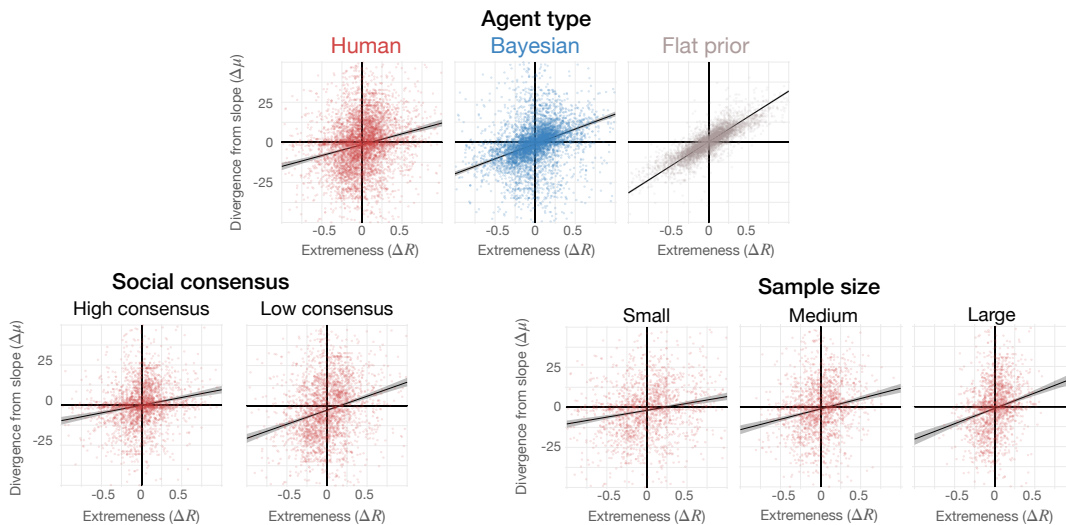


Fig. 4. The sensitivity of the three agents to sample extremeness. The regression line shows model estimates. Points depict the observed, empirical responses. A weaker correlation implies more resilience to spurious samples.

10.91]), followed by medium (12.90, CI: [10.16, 15.74]) and large (19.21, CI: [15.79, 22.77]). For both Bayesian agents, their sensitivities to extremeness by social consensus and sample size closely align with the overall estimate.

5 DISCUSSION

Overall, both informative and uninformative Bayesian machines are more accurate than participants when data samples closely represent the data-generating models. However, as samples become more extreme, participants become more accurate relative to the Bayesian machines. Participants eventually outperform the machines at certain points and continue to widen the gap as extremeness increases. Even though participants outperform the Bayesian machines for a wide range of extremeness, the machines are still more accurate overall because extreme samples are less likely to be generated.

Our results reveal that individuals often deviate from normative inference while analyzing visualizations. Interestingly, these deviations can be beneficial, especially when viewing visualizations that potentially include extreme data from small sample sizes. In such cases, an analyst’s intuition may afford more accurate inferences than those produced by an ideal inference machine, even if the machine is calibrated with the participants’ pre-existing knowledge. Conversely, for larger and more reliable datasets, the precision offered by machine-based inferences can reduce bias, leading to more accurate inference.

These findings suggest that humans and machines excel in different scenarios, highlighting the potential for collaboration between humans and AI-based agents in inference-making. This collaboration allows each to complement the limitations of the other [25, 27]. For instance, a system could offer feedback and recommendations on visual inference, taking into account the sample’s characteristics and the user’s prior knowledge. If the current sample seems reliable, factoring in aspects like sample size, then one might lean toward relying on a machine. Conversely, in cases where the data appears noisy, relying on human heuristics might prove more suitable. While including humans as part of a system has demonstrated benefits in several machine learning studies [5, 22, 24, 25, 30], the challenge in building such visual

analytics systems is to determine how much trust to allocate to the AI and when to be skeptical. One reason is that it is not straightforward to pre-determine factors such as sample extremeness when encountering new data. Furthermore, it might be a challenge for humans to accurately elicit prior knowledge that reflects statistical parameters. With the assistance of LLMs, we believe that the latter challenge could be addressed. Through natural language input, we have the potential to enhance precise knowledge elicitation, thereby bolstering the effectiveness of visual data analysis.

6 CONCLUSION

In this study, we investigate whether human inference can outperform AI models when making inferences from graphical patterns. We conducted a crowdsourced study to compare the inference accuracy of humans against Bayesian agents, incorporating multiple factors such as sample size and sample extremeness. Participants externalized both their prior beliefs (i.e., before seeing data) and posterior beliefs (i.e., after data exposure) using graphical elicitation. Our results demonstrate that humans were more accurate at inferring the true correlation level when the visualization depicted extreme data samples. These results underscore the potential effectiveness of integrating human-AI workflows to enhance the quality of decision-making.

ACKNOWLEDGEMENT

This paper is based upon research supported by the National Science Foundation under award 1942429.

REFERENCES

- [1] Justin S Albrechtsen, Christian A Meissner, and Kyle J Susa. 2009. Can intuition improve deception detection performance? *Journal of Experimental Social Psychology* 45, 4 (2009), 1052–1055.
- [2] Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80 (2017), 1–28.
- [3] In Kwon Choi, Taylor Childers, Nirmal Kumar Raveendranath, Swati Mishra, Kyle Harris, and Khairi Reda. 2019. 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*. 1–14.
- [4] In Kwon Choi, Nirmal Kumar Raveendranath, Jared Westerfield, and Khairi Reda. 2019. Visual (dis) confirmation: Validating models and hypotheses with visualizations. In *2019 23rd International Conference on Information Visualization-Part II*. IEEE, 116–121.
- [5] Pedram Daei, Tomi Peltola, Aki Vehtari, and Samuel Kaski. 2018. User modelling for avoiding overfitting in interactive knowledge elicitation for prediction. In *23rd International Conference on Intelligent User Interfaces*. 305–310.
- [6] Erik Dane, Kevin W Rockmann, and Michael G Pratt. 2012. When should I trust my gut? Linking domain expertise to intuitive decision-making effectiveness. *Organizational behavior and human decision processes* 119, 2 (2012), 187–194.
- [7] Gerd Gigerenzer. 2008. Why heuristics work. *Perspectives on psychological science* 3, 1 (2008), 20–29.
- [8] Gerd Gigerenzer and Henry Brighton. 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science* 1, 1 (2009), 107–143.
- [9] Thomas Gilovich, Dale Griffin, and Daniel Kahneman. 2002. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.
- [10] Jeremy Heyer, Nirmal Kumar Raveendranath, and Khairi Reda. 2020. 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*. 1–14.
- [11] Laura Huang. 2018. The role of investor gut feel in managing complexity and extreme risk. *Academy of Management Journal* 61, 5 (2018), 1821–1847.
- [12] Daniel Kahneman, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.
- [13] Yea-Seul Kim, Paula Kayongo, Madeleine Grunde-McLaughlin, and Jessica Hullman. 2020. Bayesian-assisted inference from visualized data. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 989–999.
- [14] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the gap: Visualizing one’s predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1375–1386.
- [15] Yea-Seul Kim, Logan A Walls, Peter Krafft, and Jessica Hullman. 2019. A bayesian cognition approach to improve data visualization. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
- [16] Ratanond Koonchanok, Parul Baser, Abhinav Sikharam, Nirmal Kumar Raveendranath, and Khairi Reda. 2021. Data prophecy: Exploring the effects of belief elicitation in visual analytics. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [17] Ratanond Koonchanok, Michael E Papka, and Khairi Reda. 2024. Trust Your Gut: Comparing Human and Machine Inference from Noisy Visualizations. *IEEE Transactions on Visualization and Computer Graphics* (2024).

- [18] Ratanond Koonchanok, Gauri Yatindra Tawde, Gokul Ragunandhan Narayanasamy, Shalmali Walimbe, and Khairi Reda. 2023. Visual Belief Elicitation Reduces the Incidence of False Discovery. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [19] Shuaimin Li, Xuanang Chen, Yuanfeng Song, Yunze Song, and Chen Zhang. 2024. Prompt4Vis: Prompting Large Language Models with Example Mining and Schema Filtering for Tabular Data Visualization. *arXiv preprint arXiv:2402.07909* (2024).
- [20] Haihan Lin, Derya Akbaba, Miriah Meyer, and Alexander Lex. 2022. Data hunches: Incorporating personal knowledge into visualizations. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 504–514.
- [21] Paula Maddigan and Teo Susnjak. 2023. Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models. *IEEE Access* (2023).
- [22] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. 2022. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5323–5331.
- [23] Eugene Sadler-Smith and Erella Shefy. 2004. The intuitive executive: Understanding and applying ‘gut feel’ in decision-making. *Academy of Management Perspectives* 18, 4 (2004), 76–91.
- [24] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 215–227.
- [25] Mark Steyvers, Heliodoro Tejada, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.
- [26] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185, 4157 (1974), 1124–1131.
- [27] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. 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*. 1–6.
- [28] Jinge Wang, Qing Ye, Li Liu, Nancy Lan Guo, and Gangqing Hu. 2024. Scientific figures interpreted by ChatGPT: strengths in plot recognition and limits in color perception. *NPJ Precision Oncology* 8, 1 (2024), 84.
- [29] Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim, and Yong Wang. 2023. LLM4Vis: Explainable visualization recommendation using ChatGPT. *arXiv preprint arXiv:2310.07652* (2023).
- [30] Zijie J Wang, Alex Kale, Harsha Nori, Peter Stella, Mark E Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. 2022. Interpretability, then what? editing machine learning models to reflect human knowledge and values. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4132–4142.
- [31] Yuheng Zhao, Yixing Zhang, Yu Zhang, Xinyi Zhao, Junjie Wang, Zekai Shao, Cagatay Turkay, and Siming Chen. 2024. LEVA: Using Large Language Models to Enhance Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* (2024).