Left, Right, and Gender: Exploring Interaction Traces to Mitigate Human Biases

Emily Wall*, Arpit Narechania*, Adam Coscia, Jamal Paden, and Alex Endert

Abstract—Human biases impact the way people analyze data and make decisions. Recent work has shown that some visualization designs can better support cognitive processes and mitigate cognitive biases (i.e., errors that occur due to the use of mental "shortcuts"). In this work, we explore how visualizing a user's interaction history (i.e., which data points and attributes a user has interacted with) can be used to mitigate potential biases that drive decision making by promoting conscious reflection of one's analysis process. Given an interactive scatterplot-based visualization tool, we showed interaction history in *real-time* while exploring data (by coloring points in the scatterplot that the user has interacted with), and in a *summative* format after a decision has been made (by comparing the distribution of user interactions to the underlying distribution of the data). We conducted a series of in-lab experiments and a crowd-sourced experiment to evaluate the effectiveness of interaction history interventions toward mitigating bias. We contextualized this work in a political scenario in which participants were instructed to choose a committee of 10 fictitious politicians to review a recent bill passed in the U.S. state of Georgia banning abortion after 6 weeks, where things like gender bias or political party bias may drive one's analysis process. We demonstrate the generalizability of this approach by evaluating a second decision making scenario related to movies. Our results are inconclusive for the effectiveness of interaction history (henceforth referred to as *interaction traces*) toward mitigating biased decision making. However, we find some mixed support that interaction traces, particularly in a summative format, can increase awareness of potential unconscious biases.

Index Terms—Human bias, bias mitigation, decision making, visual data analysis

1 Introduction

As the sheer volume and ubiquity of data increases, data analysis and decision making are increasingly taking place within digital environments, where humans and machines collaborate and coordinate to inform outcomes, facilitated by interactive visual representations of data. These environments provide a new way to measure and characterize cognitive processes: by analyzing users' interactions with data during use. Analyzing user interactions can illuminate many aspects about the user and their process, including identifying personality traits [9], recovering a user's reasoning process [15], and most relevant to the present work, characterizing human biases [46]. In this work, we explore how showing a user prior interaction history might be used to **mitigate potential biases** that may be driving one's data analysis and decision making.

We utilize the technique of **interaction traces**, a form of provenance [35] visualization in which a user's own previous interactions with the data influence the visual representations in the interface. We show interaction traces in two ways: *in-situ* interaction traces alter the color of visited data points in a scatterplot based on the frequency of prior interactions (Figure 1E), and *ex-situ* interaction traces are shown in an additional view of the data that compares the distribution of a user's interactions to the underlying distributions in the data (Figure 1F).

We operationalize **biased behavior** as deviation from a baseline of equally probable interactions with any data point. It can be conceptualized as a model mechanism [48], captured using bias metrics [46], and may correspond to other notions of societal or cognitive bias. Similarly then, a **biased decision** is one which reflects choices that are not proportional to the data. This definition of bias serves as a point of comparison for user behavior and decision making, but, as described in [46], is not inherently negative and requires interpretation in context by the user given their goals. We posit that visualization of interaction traces will lead to reflection on behavior and decision making, increasing awareness of potential biases. Importantly then, our definition of **bias mitigation** is a reduction in *unconscious* biases, which we aim to address by promoting user reflection [40] about factors driving their decision making processes. In particular, we examine the effectiveness of visualizing traces of users' interactions, where effectiveness is measured by (1) *behavioral changes*, (2) *changes in decisions made*, and (3) *increased cognitive awareness*.

To assess the impact of interaction traces toward mitigating potential biases, we designed an interactive scatterplot-based visualization system (Figure 1). We conducted a crowd-sourced experiment in which users performed two decision making tasks in the domains of (1) politics and (2) movies. In the political scenario, we curated a dataset of fictitious politicians in the U.S. state of Georgia and asked participants to select a committee of 10 responsible for reviewing public opinion about the recently passed Georgia House Bill 481 (Georgia HB481), banning abortion in the state after 6 weeks. In this scenario, several types of bias may have impacted analysis, including gender bias (i.e., bias favoring one gender over another), political party bias (i.e., voting along political party lines, regardless of potential ideological alignment from candidates in another party), age bias (i.e., preferential treatment of candidates based on age), and so on. Participants in the experiment also completed a parallel task in the domain of movies: to select 10 representative movies from a dataset of similar size and composition. In this task, we anticipated that participants' decisions would be driven by idiosyncrasies of their individual preferences.

For the given tasks, we assessed four interface variations: CTRL, SUM, RT, and RT+SUM. The CTRL interface served as the control system, which we compared to variations that provided either *real-time* (RT) or *summative* (SUM) views of the user's interaction traces (or both, RT+SUM). Our experiments yielded mixed results, offering support that interaction traces, particularly in a summative format, can lead to behavioral changes or increased awareness, but not substantial changes to final decisions. Interestingly, we find that increased awareness of unconscious biases may lead to amplification of individuals' conscious, intentional biases. We emphasize that regardless of domain, our goal is not to address overt biases (e.g., in the form of discrimination) in this work; rather, we believe visualization *can* have an impact on increasing user awareness of potential unconscious biases that may impact decision making in critical ways.

In this work, we highlight the following contributions:

- *Emily Wall is with Emory University. E-mail: emily.wall@emory.edu.*
- *Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert are with Georgia Tech. E-mail: {arpitnarechania, acoscia6, jpaden, endert}@gatech.edu.*
- **Authors contributed equally.*

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

1. We utilize a technique for showing interaction history, (referred to as *interaction traces*, Section 4.2),
2. We present results of three formative in-lab studies that describe exploratory and qualitative findings (Section 5), and
3. We present results of a crowd-sourced study that describes quantitative effects of interaction traces (Section 6).

In the following sections, we present a description of the datasets and interface used in the studies, findings from the in-lab and crowd-sourced experiments, and a discussion of how these results can inform the design of future systems that can mitigate potentially biased analyses.

2 Related Work

Wall et al. introduced four definitions of the term "bias" in data visualization, relating to human cognitive, perceptual, and societal biases, and a fourth usage as a *model mechanism* [48]. We adopt the fourth perspective. Namely, we utilize computational metrics to characterize how a person's interactive behavior deviates from a baseline model of expected behavior [46]. Specifically, we model and visualize how a user's interaction sequences deviate from uniform behavior. This model serves as a benchmark against which a user can compare, interpret, and reflect on their behavior. We intend this usage to have a neutral connotation – deviation from a baseline is neither good nor bad, but relies on a user's interpretation of the metrics in context. In a political scenario (one task in our experiment), these metrics can be used to indicate when a user's attention is skewed toward e.g., a particular political party, politicians' genders or ages, etc.

These metrics capture deviations which may correspond to systematic biases, e.g., cognitive or societal, which inherently impact the lens through which a person analyzes and makes decisions from data. In Cognitive Science, bias can describe an irrational error that results from heuristic decision making [29, 30, 44]. Alternatively, it can refer to a rational decision made under certain constraints (e.g., limited time or high cognitive load) [21–23]. Cognitive biases can thus influence how people make decisions when "fast and frugal" heuristics [21] are employed in place of concerted, deliberative thinking [17].

In Social Sciences, bias often refers to prejudices or stereotypes that are relevant in society (e.g., racial bias or gender bias). In this work, we refer to such biases as *social biases*. These biases can have far-reaching impacts, such as propagating racial or gender bias to machine learning [20, 32].

Social biases may be influenced by cultural norms, individual experiences or personality variations, and they can shape our decision making in a conscious or an unconscious manner [25]. These biases can have severe implications in a variety of decision making domains. For example, consider the impact of racial bias in hiring. Researchers have found discrimination, either conscious or unconscious, based on racial name trends [8], showing that equivalent resumes with traditionally White names receive 50% more callbacks from job applications than resumes with traditionally African American names. As a result, companies may lack a diverse workforce, which can have implications on employee turnover, group isolation or cohesion, workplace stress, and so on [37].

In the visualization community, bias has garnered increasing attention. Researchers have cataloged relevant biases [14] and proposed methods for detecting the presence of a particular type of bias [11, 13, 24, 45–47]. Other recent works proposed or categorized methods for mitigating bias [12, 31, 41, 49]. Within Wall et al.'s design space of bias mitigation techniques for visualizations [49], our proposed system manipulates the visual representation to show metrics about a user's analysis in a minimally intrusive, orienting [10] fashion, to ultimately facilitate more balanced decision making. Distinct from prior work on bias mitigation in visualization, we focus on increasing awareness of *unconscious* biases which could correspond to cognitive or social biases, including gender bias and political bias (e.g., bias towards one political party), among others.

To mitigate potential biases driving decision making, we are motivated by literature in Cognitive Science on nudging [42] and boosting [27], that can influence people's *behavior* and *decision making* by altering the choice architecture (i.e., the way that choices are presented) or improving individuals' decision making competences. We apply this analogy in the context of visualization with the goal of "nudging" users toward a less biased analysis process. In visualization research, prior work has shown some ability to impact user behavior, resulting in more broad exploration of the data (e.g., by coloring visited data points differently [18] or by adding widgets that encode prior interactions [50]). Furthermore, we are inspired by work on reflective design [40], wherein our purpose is not to prescribe an optimal decision to users, but rather to encourage thoughtful reflection on motivating factors of those decisions while users maintain full agency. We describe the visualization system and interaction traces in Section 4.2.

3 Bias Metric Review

While several metrics have been proposed to quantify aspects of a user's analysis process (e.g., [19, 28, 33, 46]), here we focus on bias metrics introduced by Wall et al. [46] which are theoretically applicable to various types of bias and have been used for initial characterization of anchoring bias [47]. We quantify bias using the data point distribution (DPD) and attribute distribution (AD) metrics [46]. These metrics characterize, along a scale from 0 (no bias) to 1 (high bias), how a user's interactive behavior deviates from expected behavior. In this case, expected behavior is defined by equal probability of interaction with any given data point in the dataset.

Consider a dataset of politicians. Data point distribution (DPD) describes how the user's interactions are distributed over the points (politicians) in the dataset. Uniform interactions over all politicians will result in a low metric value (less biased), while repeated interaction with a subset of the data (e.g., only Republicans) will result in a higher metric value (more biased).

Attribute distribution (AD) considers how the users' interactions across the data map to the underlying distributions of each attribute. That is, if the dataset has politicians with an average Political EXPERIENCE of 9 years, but the user focuses almost exclusively on politicians with 15+ years of EXPERIENCE (potentially revisiting the same subset of experienced politicians), the attribute distribution metric for EXPERIENCE would be high (more biased). Alternatively, if the user's interactions are proportional to the dataset, the metric value would be low (less biased). See [46] for the precise formulation of the bias metrics. These metrics drive the visualization design in this paper that shows a user's interaction traces as they make their decisions.

4 Methodology

To study the effect of visualizing interaction traces toward mitigating bias, we conducted a series of in-lab studies and a crowd-sourced experiment to test four interface variations (CTRL, SUM, RT, RT+SUM). In this section, we describe two tasks and datasets in the domains of politics and movies (Section 4.1) and the implementation of a visualization system that realizes interaction traces to serve as the testbed for subsequent experiments (Section 4.2).

4.1 Tasks & Datasets

We selected two complementary tasks (counterbalanced within subjects) to observe how people would respond to interaction traces in the presence of a variety of potential biases, described below for each task.

4.1.1 Politics

Task. The USA has a two-party political system: Democrats and Republicans [7]. In Georgia's General Assembly, committees may be formed to explore complex issues, draft legislation, and make recommendations [3]. Many such committees, particularly subcommittees focused on specific issues, may be formed by top-down appointment [3]. With membership in committees often decided by an individual or by few, the decision can be subject to an individual's biases.

In May 2019, Georgia's incumbent Governor Brian Kemp signed Georgia House Bill 481 (Georgia HB481) banning abortion after 6 weeks (earlier than the previous state law of 20 weeks) [39]. Scheduled to take effect in January 2020, the bill was received by the public with

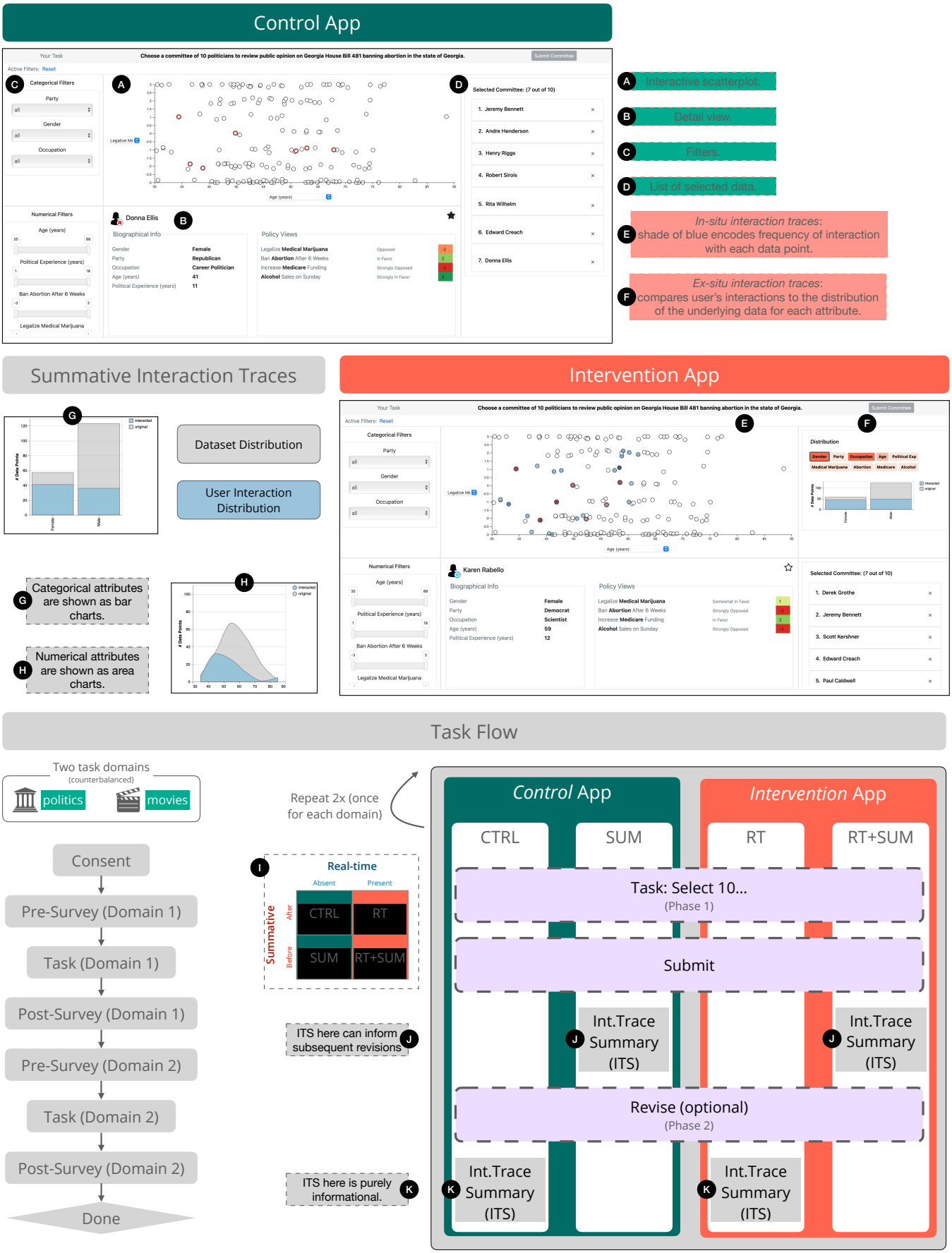

Fig. 1: A summary of the interfaces and procedure for this experiment.

significant controversy [1]. Supporters of Georgia HB481 (colloquially referred to as a "Heartbeat Bill") hoped it would lead to overturning of Roe v. Wade, 410 U.S. 113 (US federal court decision protecting a woman's right to an abortion, 1973), while opponents hoped to challenge the bill before it became law.

Given a dataset of fictitious politicians, participants were given the following task: *Imagine you are engaged in political decision-making in the state of Georgia. The debate about abortion is ongoing, with variations of these bills cropping up across other states in the United States, which can potentially learn from the ongoing debate in Georgia. Select a committee of 10 candidates that you feel should review public opinion in Georgia on the controversial Georgia HB481.* We selected this task to simulate a realistic decision making scenario in American politics and evaluate our interventions in a politically and socially relevant context. Furthermore, this topic and dataset can elicit a number of factors that may influence an individual's decision making process, including personal preferences as well as multiple types of social biases (e.g., gender bias or political party bias), both conscious and unconscious.

Dataset. We generated a dataset of 180 fictitious politicians, representing the composition of the Georgia General Assembly [2]. Each row in the dataset represents a politician, described by the following attributes: GENDER, POLITICAL PARTY, OCCUPATION, AGE, and EXPERIENCE, along with numerical representations $\in [-3, 3]$ of the politician's view on topics such as BANNING ABORTION AFTER 6 WEEKS, LEGALIZING MEDICAL MARIJUANA, INCREASING MEDICARE FUNDING, and BANNING ALCOHOL SALES ON SUNDAYS (positive numbers indicate that the politician is in favor, while negative numbers indicate that the politician is opposed). Politicians' names are artificially generated from US census data [5]. The dataset contains 59% Republicans, of which 14% are female; and 41% Democrats, of which 57% are female, mimicking the distributions in the Georgia General Assembly [4]. The ages, political experience, and occupations were derived from data on the 115th U.S. House of Representatives [6]. The policy views were generated to represent general party voting trends (e.g., Democrats tend to be opposed to banning abortion, while Republicans tend to be in favor of the ban) with the strength of those views representing recent increasing polarity [38] in the USA political system (e.g., fewer politicians have neutral positions or positions against the party trend).

4.1.2 Movies

Task. Given a dataset of fictitious movies, participants were given the following task: *Analyze the data to pick 10 movies that you feel represent the collection of movies in the dataset as a whole.* We selected this task to complement the political scenario. It represents a parallel task (selecting a representative subset) in a domain that the general public is familiar with (movies). We hypothesize that this task may elicit an entirely different set of (less obviously dangerous) biases, based on idiosyncrasies in one's movie preferences. For instance, participants may make selections for movies by focusing on attributes of the data that are most familiar to them (e.g., ROTTEN TOMATOES RATING) while disregarding others that have a lesser impact on their own movie habits (e.g., RUNNING TIME). The instructions for both tasks were intentionally vague to avoid suggesting any particular criteria for selecting politicians / movies.

Dataset. The movies dataset was adapted [1] to match the general structure of the political dataset. We sampled 180 movies from the dataset and selected 9 attributes in total (3 categorical, 6 numerical): CONTENT RATING, GENRE, CREATIVE TYPE, WORLDWIDE GROSS, PRODUCTION BUDGET, RELEASE YEAR, RUNNING TIME, ROTTEN TOMATOES RATING, and IMDB RATING to match the dimensionality of the political dataset. In pilot studies, we found that (1) real movie titles were problematic because participants relied heavily on familiarity of titles rather than the data before them; and (2) anonymized identifiers (e.g., "Movie1", "Movie2", ...) led participants to be less engaged with the task. For consistency with the political scenario, we generated fictitious movie titles [2] so that participants would be more engaged with the task while not relying only on familiarity of titles. Complete datasets and analyses are included in supplemental materials[3].

4.2 System

Overview. For our experiments, we utilized a simplified version of Lumos [34], a visualization system to support data exploration while promoting reflection and awareness during visual data analysis. To assess the effectiveness of visualizing interaction traces, we produced two versions of the visualization system: a Control version of the interface, and an Intervention version of the interface, which was modified to visualize traces of the user's interactions with the data in real-time (Figure 1). Components A-D in Figure 1 are common across the Control and Intervention interfaces. The primary view is an interactive scatterplot (A), where the x- and y-axes can be set to represent attributes of the data via selection in a drop-down menu. Hovering on a point (politician / movie) in the scatterplot populates the detail view (B), which shows all of the attributes of that data point. Filters for categorical (e.g., GENDER, OCCUPATION, etc. in the political dataset; GENRE, CONTENT RATING, etc. in the movies dataset) and ordinal & numerical attributes (e.g., AGE, EXPERIENCE, etc. in the political dataset; RUNNING TIME, IMDB RATING, etc. in the movies dataset) can be adjusted on the left-hand side of the interface (C) using drop-down menus and range sliders. Clicking on the point in the scatterplot or on the star icon in the detail view adds the politician / movie to the selected list (D). Selected data points are shown in the scatterplot with a thick red border.

Interaction Traces. In the Intervention interface, user interaction traces are shown in *real-time* in the interface with respect to *data points* and with respect to *attributes*. First, the points in the scatterplot are given a blue fill color (in-situ interaction traces) once the user has interacted with the data point, with darker shades representing a greater number of interactions (DPD metric [46]; Figure 1E). The Control interface, by comparison, uses no fill color on the points (Figure 1A). Second, the top right view (Figure 1F) compares the user's interactions to the underlying distributions of the data for each attribute (ex-situ interaction traces). The attribute tags are colored with a darker orange background when the user's interactions deviate more from the underlying data and with a lighter orange or white background when the user's interactions more closely match the underlying distribution of data (AD metric [46]). Categorical attributes (GENDER pictured) compare user interactions to the underlying dataset using bar charts, where gray represents the underlying distribution of data (approximately 32% women, 68% men) and a superimposed blue bar represents the distribution of the user's interactions (approximately evenly split between women and men). Numerical attributes compare user interactions to the underlying data distributions using area curves.

Real-Time v. Summative. The interaction traces pictured in Figure 1(E-F) in the Intervention interface are shown in *real-time*. We also show interaction traces in a *summative* format, depicted in Figure 1(G-H), after the user has made a decision (choosing 10 politicians or 10 movies). We hypothesize that both real-time and summative formats may be beneficial in different ways. In real-time, interaction traces may help users maintain awareness throughout their analysis process about the distribution of their analytic focus across the data. In a summative format, interaction traces may be easier to process and adjust from in subsequent analyses without the additional simultaneous cognitive load of the decision itself. We test variations of both in our experiment.

5 FORMATIVE IN-LAB STUDY RESULTS

We conducted three (3) formative in-lab studies, described in turn below. These formative studies utilized a similar task as Section 4.1.1 about political decision making along with earlier variants of the same control and intervention interfaces, described in Section 4.2. Analysis from these formative studies was largely qualitative and exploratory [43]

[1] A federal judge permanently blocked Georgia HB481 in July 2020, finding it in violation of the U.S. Constitution [36]

[2] https://thestoryshack.com/tools/movie-title-generator/

[3] https://github.com/gtvalab/bias-mitigation-supplemental

in nature, informing the hypotheses and design of the confirmatory crowd-sourced experiment described in Section 6.

5.1 In-Lab Study 1

In the first formative study, 6 participants utilized the Control interface to choose a political committee. Our goal was to observe a baseline of user behavior and choices. Many participants intentionally balanced their political committee along several attributes (seeking "balanced representation" – P02). For example, four participants balanced by GENDER (5 men and 5 women). The same four also balanced by PARTY (5 Republicans and 5 Democrats).

The ways that participants *biased* their committee selections were explicit but nuanced. For instance, while P05 balanced across GENDER and PARTY, they ultimately chose a committee with all 10 members opposed to the bill, explicitly prioritizing "members (who) were very opposed to the bill." We generally observed that **participants were able to *maintain awareness* about potential biases driving their decision making**, which we hypothesized was the result of the relatively small version of the political dataset used in this study (144 data points and 5 attributes). Subsequent formative studies increased data dimensionality, from which we observed greater difficulty in maintaining conscious bookkeeping of attributes that impacted decision making.

5.2 In-Lab Study 2

In the second formative study, 12 participants each utilized the Control and Intervention interfaces to choose a political committee (24 participants in total). Our goal was to observe the effects of interaction traces on users' behavior and subsequent decisions. Exploratory analyses revealed some notable differences between participants' behavior who used the Control v. Intervention interface. In particular, for the AGE attribute, Control participants tended to have higher Attribute Distribution (AD) bias metric values over time than Intervention participants, suggesting that **Intervention participants interacted with politicians whose ages were more proportional to the underlying dataset than Control participants** ($\mu_C = 0.857$, $\mu_I = 0.729$, $H = 3.360$, $p = 0.057$). Furthermore, participants who saw interaction traces **(Intervention) trended toward choosing more *proportional* gender composition of committees in the political task** (Figure 2a); however, this trend was not replicated in the third and final formative study (Figure 2b), potentially due to the introduction of confounding factors, described later.

We also observed instances where interaction traces may have led to altered behavior. For instance, after interacting with the interaction trace view, one participant's bias toward PARTY sharply decreased (as observed by the AD bias metric). One possible explanation is that the user observed bias in their interactions toward Democratic politicians in the interaction trace visualization and consequently went on to focus on Republicans to reduce the bias.

As captured by Likert ratings, participants found the *summative* metric visualization (4.5 / 5) more useful than *real-time* (4 / 5 for in-situ and 3 / 5 and 4 / 5 for categorical and numerical ex-situ representations, respectively). Participants expressed more surprise about how their interactions and selections mapped to the underlying dataset when considering the *summative* view, suggesting that the view increased their awareness of bias in their analysis process (e.g., P10-I said "I'm surprised I didn't choose a doctor").

5.3 In-Lab Study 3

In the third formative study, again 24 participants utilized the Control and Intervention interfaces to choose a political committee. This study focused on qualitative analysis of *awareness*, while also addressing some shortcomings of the previous experiment (namely, the previous experiment was completed primarily by male participants, and the dataset used had only one female Republican). We observed similar, yet weaker, effects as the previous in-lab study. It could be that there is weak or no effect (which is plausible given the exploratory nature of our analyses), or it could be the result of a confounding change to the interface in this study. In particular, for the study, we permitted categorical attributes to be assigned to axes of the scatterplot. The result

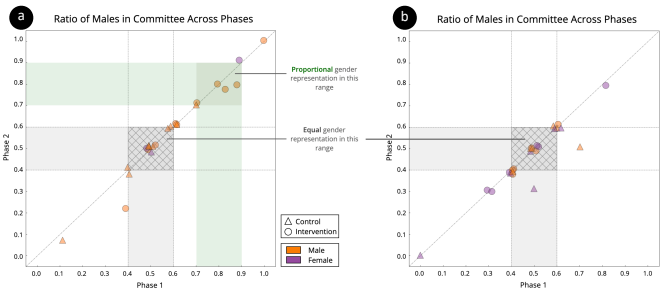

Fig. 2: GENDER balance in committees chosen by 24 participants in (a) Formative Study 2 and (b) Formative Study 3. Balance is shown as the ratio of men in each participant's committee in Phase 1 (x-axis) and Phase 2 (y-axis) (shape encodes condition; color encodes participant gender).

is that well-formed clusters appear on the scatterplot, which could itself help people more easily choose representative samples (e.g., pick a point from each cluster).

We again observed qualitative evidence of the efficacy of interaction traces toward increasing awareness of potential biases. Because the interaction trace view compares a user's interactions to the underlying distribution of the data, we hypothesized this would lead to changes in user decision making to *make the committee more proportionally representative of the underlying dataset*. For example, one participant's committee was comprised of 10 Democrats, until interacting with the interaction trace view. The participant then adjusted the committee from 10 Democrats to 4 Republicans and 6 Democrats. In fact, examination of the interaction trace view made the participant aware of a mistake in her analysis: "I forgot I had only filtered by Democrats."

Also consistent with previous findings, **participants indicated higher preference for *summative* interaction traces over *real-time* interaction traces**. Further, using a grounded theory approach to code participant utterances during think-aloud *summative* review of interaction traces, we found that **participants in the Control condition made more statements on average indicating heightened *awareness* than participants in the Intervention condition**. We hypothesize this may be due to the fact that Intervention participants already saw their interaction traces in *real-time* prior to the summative review phase.

5.4 Summary

As a result of these observations, we planned a fourth experiment to be conducted virtually with crowd-sourcing, in which we made the following adjustments:

1. We adjusted the size and dimensionality of the dataset to be sufficiently difficult such that unconscious biases may arise,
2. We corrected for confounds resulting from interface changes between the formative studies (in the crowd-sourced experiment, we permit numerical and ordinal attributes only to be assigned to axes, while categorical attributes can be used to apply filters; implications of which are described further in the Discussion).

We also added an additional task in the domain of movies to study the generalizability of our approach. Furthermore, across all of the in-lab formative studies, the vast majority of our participants identified as Democrats. Hence, conducting an experiment via the online crowdsourcing platform Amazon Mechanical Turk allowed us to broaden the political demographic of our users.

6 CROWD-SOURCED EXPERIMENT

6.1 Procedure

This study utilized a 2x2 design which manipulated real-time interaction traces (present, absent) x summative interaction traces (before revision, after revision). Participants in the user study were randomly assigned to one of four conditions: CTRL, SUM, RT, or RT+SUM (Figure 1I). The procedure is depicted in Figure 1. After providing informed consent, participants completed a background questionnaire. Participants were shown a demonstration video of the interface using a cars dataset, then

given the opportunity to practice by *choosing a shortlist of 5 cars they would be interested to test drive*.

Participants completed the first task (either politics or movies) followed by the second (movies or politics), with the order counterbalanced between subjects. For each task, participants first chose a set of 10 politicians / movies, then submitted their decision. Next, participants were either immediately given the opportunity to revise their selection (CTRL and RT) or were shown the summative interaction trace view (SUM and RT+SUM). The summative interaction trace view shown in Figure 1(G-H) depicted for each attribute of the dataset: the underlying distribution (gray), and the distribution of user interactions (blue). Then, based on any imbalances observed, participants were given the opportunity to reflect and revise their committee if desired. Lastly, those who did not see the summative interaction trace view before revision (CTRL and RT) were shown the view at the end after their decision was finalized.

Those who saw summative interaction traces *before* revision could incorporate any findings or realizations about their analysis process into subsequent revision, while those who saw summative interaction traces *after* revision could only use this information to reflect afterwards, without impacting any decisions. The study took participants 44 minutes on average, and they were compensated $10.

6.2 Participants

Based on a statistical power analysis from formative studies, we determined that at least 11 participants per condition would be required to detect an effect ($power = 0.8$, $\alpha = 0.05$). We recruited 56 participants via Amazon Mechanical Turk. We ultimately rejected 6 submissions due to a combination of failed attention checks, missing data, speeding through the study, and poor open-ended responses, leaving us with data from 50 participants who were randomly assigned to one of four conditions (13 CTRL, 14 SUM, 11 RT, 12 RT+SUM). Workers were restricted to only those located in the U.S. state of Georgia with 5,000+ approved HITs over their lifetime and a $\geq 97\%$ approval rating. By gender, participants identified as female (28), male (21), and agender (1). Participants were 24-69 years old ($\mu = 40$, 1 preferred not to say) and self-reported an average visualization literacy of $\mu = 3.08, \sigma = 0.9$ on a 5-point Likert scale. They had a wide range of educational backgrounds (6 high school, 12 some college, 7 associate's degree, 20 bachelor's degree, 3 master's degree, and 2 post-graduate degree); and a variety of fields of work, including e.g., art, bookkeeping, computer science, criminal justice, marketing, microbiology, office administration, political science, sales, social work, among others. We refer to participants from each condition as $\{P_{CTRL}1$ - $P_{CTRL}13\}$, $\{P_{SUM}1$ - $P_{SUM}14\}$, $\{P_{RT}1$ - $P_{RT}11\}$, and $\{P_{RT+SUM}1$ - $P_{RT+SUM}12\}$.

Politics Background. Most participants had voted in US Presidential (49), state (43), or local (35) elections. Participants identified as Democratic (35) and Republican (15), rating themselves on the political spectrum as conservative (2), moderate conservative (9), moderate (8), moderate liberal (14), and liberal (17).

Movies Background. Participants rated varying importance of movies in their lives (2 no importance, 15 little importance, 20 moderate importance, 12 large importance, 1 most importance). They watched movies daily (4), weekly (30), or monthly (16) and reported a diverse range of preferred genres.

6.3 Hypotheses

Based on findings from formative in-lab studies, our hypotheses for this experiment are as follows. We organize our hypotheses according to those regarding **B**ehavior, **D**ecisions, **A**wareness, and **U**sability.

- **B1** *Real-time* interaction traces will have an effect on users' analysis *process*.
- **B2** *Summative* interaction traces seen *before* revision will lead users to make more *revisions*.
- **D1** SUM, RT, and RT+SUM participants will make selections more proportional to the underlying data than CTRL participants.
- **A1** CTRL participants will exhibit greater *surprise* upon seeing *summative* interaction traces (Figure 1(G-H)) than participants in SUM, RT, and RT+SUM.

Fig. 3: Modeling user behavior using point and interval estimation of the mean value of interactions of different types performed, revisions made, and the DPD bias metric.

- **A2** The attributes for which participants indicate *surprise* about their interaction traces will correlate to lower *focus*.
- **A3** The attributes for which participants indicate *surprise* or *focus* about their interaction traces will correlate to AD metric values.
- **U1** Participants in the RT and RT+SUM conditions will *not* consistently use the real-time interaction trace view (Figure 1F).
- **U2** Participants will find the summative interaction trace visualization (Figure 1(G-H)) more useful than real-time interaction trace visualizations (Figure 1(E-F)).

Based on guidance for statistical communication [16], our analyses relied primarily on parameter estimation for all quantitative measures, using empirical bootstrapping with 1000 resamples to estimate the 95% confidence intervals around all sample means. We prioritize reporting results for attributes that users indicated as high focus (e.g., the top three are BAN ABORTION AFTER 6 WEEKS, PARTY, and GENDER for the political task; IMDB RATING, GENRE, and ROTTEN TOMATOES RATING for the movies task). Complete analyses are available in supplementary materials.

6.4 Behavior

We hypothesized that the presence of *real-time* interaction traces would impact user behavior as measured by (1) interaction counts, (2) bias metric values [46], and (3) revisions during Phase 2.

Interactions. Figure 3 (left, center) illustrates total number and type of interactions performed by users in each condition for both the politics and movie tasks. We find some notable distinctions as observed by lesser overlap in confidence intervals. Namely, CTRL participants performed fewer hover interactions and fewer total interactions, demonstrating less interactive behavior than those who saw interaction traces. However, other specific interaction types showed less obvious trends. This result provides **some support for hypothesis B1**.

Bias Metrics. The AD bias metric values provide one way of quantifying how a user's interactive behavior aligns with the distributions of the underlying data per attribute [46]. Lower metric values indicate interaction distributions that are more similar to the distribution of a given attribute, while higher metric values indicate dissimilarity. Figure 4 (left) shows the average AD bias metric values for the top three attributes that participants focused on in each task. We observe that CTRL participants exhibited less bias towards some attributes (e.g., GENDER and PARTY in the political task and GENRE in the movies task) while other attributes display less clear trends. We hypothesize that, with increased *awareness of unconscious biases*, the interaction trace interventions may have ultimately *amplified conscious biases*, discussed further in Section 7.

The DPD bias metric similarly quantifies how evenly a user's interactions are divided among individual data points. Figure 3 (top, right) shows a slight trend toward lower DPD metric values for CTRL participants. These results provide **mixed support for hypothesis B1**.

Revisions. All participants had the opportunity to revise their initial selections. We hypothesized that those participants who saw *summative*

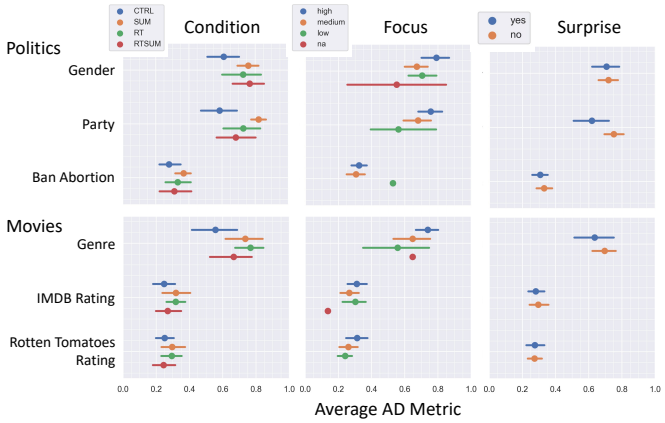

Fig. 4: The effect of condition, focus, and surprise on estimating the average AD metric value for select attributes in each task.

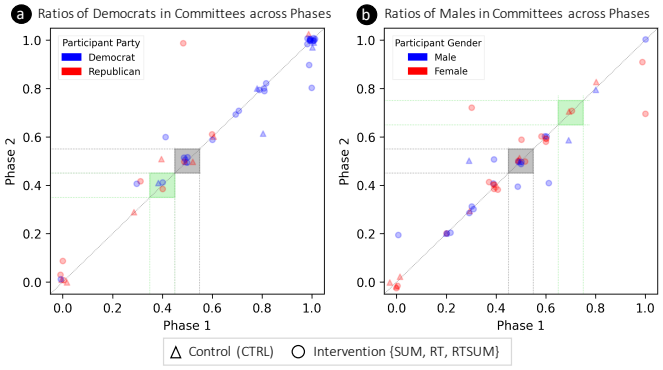

Fig. 5: Ratio of PARTY (a) and GENDER (b) composition of committees in Phase 1 and Phase 2 of the political task, colored by the participant's corresponding party affiliation and gender.

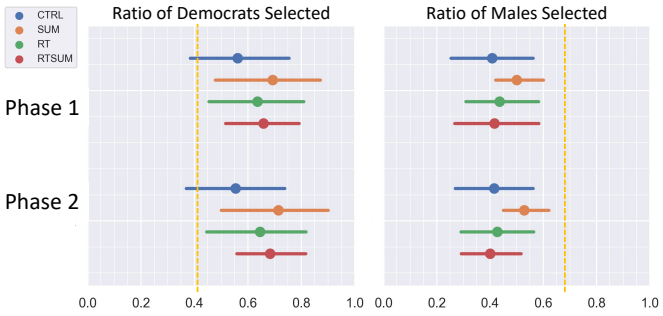

Fig. 6: Mean estimates for the ratio of Democrats (left) and Males (right) in participants' chosen committees, selected during each phase for each condition. Vertical dashed lines indicate ratio of each candidate type present in the underlying dataset.

interaction traces *before* revision would make more revisions (i.e., number of edits to their initial selection after Phase 1). Figure 3 (bottom right) shows that SUM and RT+SUM participants who saw summative interaction traces before revision tended to make more revisions compared to CTRL and RT participants during the political task, but not during the movies task. Similarly, the same groups tended to perform more interactions in general in the phase 2 revision (Figure 3 (bottom left)), demonstrating that summative interaction traces tended to correspond to more exploration and decision changes during revision. This result **confirms hypothesis B2**.

6.5 Decisions

We hypothesized that participants in SUM, RT, and RT+SUM (i.e., those who were influenced by interaction traces in some format, real-time or summative, before finalizing their decisions) would ultimately make choices that were more proportional to the underlying data compared to participants who did not (CTRL). We quantify this effect for binary attributes in the datasets (e.g., PARTY and GENDER in the politics dataset), by considering the ratio of values *chosen* with respect to the ratio that appears in the *dataset*. In the movies task, no attributes are binary; hence this analysis only applies to PARTY and GENDER in the politics task.

For PARTY, the dataset contains 59% Republicans and 41% Democrats. For GENDER, the dataset contains 32% females and 68% males. As shown in Figure 6, *all conditions* chose committees that were relatively dissimilar from the underlying distributions of GENDER and PARTY in the dataset (annotated with a vertical dashed line). We discuss this result further in Section 7. In fact, for participants who saw any intervention (RT, SUM, RT+SUM), they trended toward choosing *more dissimilar* distributions of PARTY compared to CTRL participants. However, there are no clear distinctions in the ratios of GENDER or PARTY between conditions. Hence, we find **no support for hypothesis D1**.

While we observe no clear effects on committee composition based on intervention conditions, we do see some distinctions in the way participants chose their committee based on their own political party affiliation. For instance, Figure 5a shows the ratio of Democrats that participants chose in Phase 1 (x-axis) and in Phase 2 (y-axis) committees. Points that fall on the diagonal represent participants who did not change the composition of their committee by PARTY during the revision. Points are colored by the political party affiliation of the participant. There is clear delineation, where participants who most identified with Democrats (blue) appear in the top right (choosing Democrat-dominant committees) and participants who most identified with Republicans (red) appear in the bottom left (choosing Republican-dominant committees). Three notable examples emerge e.g., a blue circle in the bottom left, a red circle in the top middle, and a red triangle in the top right. These participants chose final committees that were entirely composed of politicians from the opposite PARTY than their own affiliation. Upon further inspection, two participants identified as moderate or neutral on the spectrum of liberal → conservative, and one expressed divergent beliefs from their affiliated party. For GENDER (Figure 5b), we do not observe such a clear distinction. Participants' committees reveal some clusters (e.g., in the bottom left, many female participants chose committees of all female politicians; just above that cluster along the diagonal, many men chose committees with 70-80% female politicians; etc). However, there is less strict division in the overall trend ($\mu_{Female} = 0.45$, $\mu_{Male} = 0.44$).

6.6 Awareness

While formative studies indicated variable impacts of interaction traces on behavior and decisions, they had a more promising qualitative effect on people's *awareness* of potential unconscious biases that drive data analysis and decision making. In particular, we assess awareness by asking two questions for each attribute of the data, at the time that the user viewed the summative interaction traces (*before revision* for SUM and RT+SUM, and *after final selections* for CTRL and RT).

1. Are you **surprised** by your interactions with this attribute? {*yes*, *no*}
2. How much **focus** did you give this attribute during your task? {*high*, *medium*, *low*, *NA*}

Figure 7 compares the average number of times a particular combination of focus and surprise was recorded between all conditions and tasks. We hypothesized that CTRL participants would express surprise more often upon seeing the *summative* interaction traces than participants in the other conditions. Participants in all other conditions had some form of signal from interaction traces (real-time or summative) *before* their final decisions, whereas CTRL participants only saw the summative view after their final selections were locked in. In the political task, we observe that CTRL participants expressed surprise for high-focus attributes more often than the other conditions, and conversely reported the lowest numbers of no surprise for high-focus attributes (Figure 7

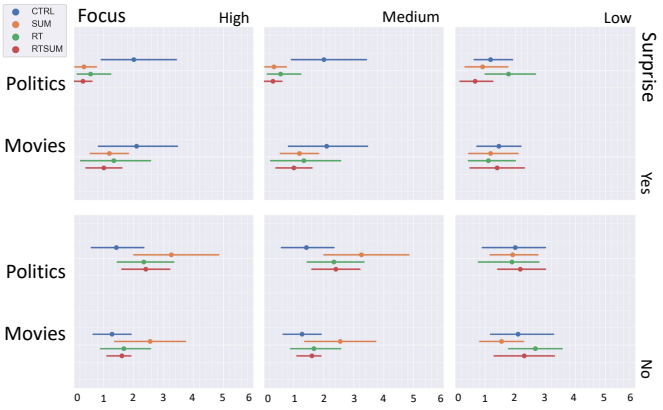

Fig. 7: Estimating the mean number of times a combination of focus and surprise was recorded for each task and condition.

left). This provides some support for the idea that seeing summaries, either in real-time (RT), before revisions (SUM) or both (RT+SUM), reduces the surprise reported when participants expressed high focus in more divisive analysis scenarios. The effect was less pronounced in the movies task. In general, as focus decreased on any given attribute, the number of surprised responses decreased uniformly as well across conditions and tasks. This pattern further supports the idea that summaries can affect whether participants are aware of their analysis strategies at the end of a task. This result **confirms hypothesis A1**.

We also hypothesized that expressing lower focus on an attribute (e.g., BAN ABORTION AFTER 6 WEEKS, PARTY, and GENDER for the political task; IMDB RATING, GENRE, and ROTTEN TOMATOES RATING for the movies task) would be correlated to more instances of surprise. That is, we believed attributes that were unattended may have a surprising distribution of user interactions, since the user did not focus on them. Comparing each column of Figure 7 top v. bottom demonstrates that when CTRL participants expressed surprise (top, blue), fewer people expressed that level of focus compared to when participants were not surprised (bottom, blue). For other conditions, there does not appear to be any substantial difference in surprise (top) v. no surprise (bottom) w.r.t. focus. That is, surprise and focus do not appear to be correlated. All means and intervals are roughly between 1 and 3. This result **disconfirms hypothesis A2**.

Lastly, we hypothesized that there would be a correlation between attributes that participants' *focused* on (Figure 4 center) or were *surprised* by (Figure 4 right) and the average AD bias metric values. Some attributes (e.g., PARTY in the political task, ROTTEN TOMATOES RATING in the movies task) roughly corresponded to greater focus related to higher AD bias metric values. On the other hand, some attributes for which participants expressed surprise about their interactions (e.g., PARTY in the political task) corresponded to lower AD bias metric values However, these trends were not true for all attributes, hence **support for hypothesis A3 is inconclusive**.

6.7 Usability

Based on formative studies, we formulated two general hypotheses about the usability of interaction trace interventions. First, in formative studies we observed relatively little use of the *real-time* interaction traces in the form of interactions with the Distribution panel (Figure 1F), which we believe to be due to high cognitive load during the task itself (i.e., participants were unable to attend to an additional view in the system while trying to explore the data). Indeed, participants interacted minimally with the Distribution panel in both the politics ($\mu_{RT} = 2.09$, 95% CI $[0.64, 3.82]$, $\mu_{RT+SUM} = 1.33$, 95% CI $[0.33, 2.42]$) and movies tasks ($\mu_{RT} = 3.18$, 95% CI $[0.45, 6.28]$, $\mu_{RT+SUM} = 2.08$, 95% CI $[0.17, 4.67]$). This result **confirms hypothesis U1**.

For similar reasons (high cognitive load), we hypothesized that participants would prefer the *summative* interaction traces over the *real-time* interaction traces. This result was somewhat variable. CTRL participants saw only summative interaction traces (Figure 1G-H, after revision) and rated their utility on a Likert scale a median 4 / 5 across both politics and movies tasks. SUM participants who saw only the summative interaction traces (before revision) likewise gave a median 4 / 5 Likert rating for movies and 3.5 / 5 for politics. RT participants rated in-situ interaction traces (Figure 1E), ex-situ interaction traces (Figure 1F), and summative interaction traces (after revision) all the same with a median Likert rating of 4 / 5 across both politics and movies tasks. These participants all viewed *summative* interaction traces very positively. Only RT+SUM participants differed, rating in-situ interaction traces 4 / 5 (politics and movies), ex-situ interaction traces 2 / 5 (politics) and 3 / 5 (movies), and summative (before revision) 3 / 5 for both politics and movies. The participants in the only condition that could compare both real-time and summative interaction traces surprisingly preferred *real-time* interaction traces. Overall, RT+SUM participants rated all forms of interaction trace lower than participants in the other conditions, perhaps due to this condition having the highest cognitive load of all the conditions, showing both real-time and summative interaction traces. This result **disconfirms hypothesis U2**.

6.8 Qualitative Feedback

The survey after each task included open-ended questions about the participant's decision criteria and, in the political scenario, their desired outcome of the committee. Below we discuss themes that emerged.

Politics. Some focused on choosing diverse committees (e.g., $P_{CTRL}2$ said "[they] tried to choose politicians with a wide range of views, to represent most people's views and not necessarily one side"). Many expressed firm goals about choosing politicians differently than the underlying data distribution (e.g., $P_{CTRL}1$ expressed they wanted their committee to be mostly women because "[they] truly feel abortion is an issue only women can really comment on"), while also simultaneously balancing other attributes (e.g., "[they] also tried to get a mix of Republicans and Democrats with many years experience in politics to get a fair showing to both sides"). Others sought committee members based on "who [they] thought might share [their] values" ($P_{CTRL}10$), instances of which were observed from both sides of the political spectrum (e.g., $P_{SUM}8$ chose "[Republican] politicians who were in favor of the abortion ban", while $P_{SUM}13$ intentionally "chose all Democrats").

Ultimately, people hoped their chosen committee would lead to outcomes such as "uphold the ban on abortion after 6 weeks, except in extreme cases where there is a life-threatening decision that needs to be made" ($P_{RT}6$), that "the abortion ban would become or stay law" ($P_{SUM}12$), or "assess the public opinion in an unbiased way" ($P_{RT+SUM}2$). Many of the biases that emerged as a result of these goals were very conscious. For instance, $P_{RT}3$ expressed, "I was really biased, to be honest. I wanted people who were in favor of abortion, and most, if not all, Republicans did not fit the bill." In some cases, participants adjusted their strategy for Phase 2 (Revision); e.g., $P_{RT+SUM}6$ said, "my revised committee was solely focused on trying to get vast ideological perspective without necessarily focusing on men or women. I did more of a 'blind' choosing without focusing so much on gender the second time around ... I only wanted different viewpoints without focusing too much on the extreme of either side."

Movies. Participants expressed diverse criteria for selecting movies. Some focused on making choices that "were a good representation of the given dataset as a whole" by "find[ing] and select[ing] films that were spread out across the graph" ($P_{CTRL}7$). Several other participants expressed a focus on finding variety for one factor, e.g., highly rated ($P_{CTRL}6$) or diverse genres ($P_{CTRL}5$). Others had criteria that were too abstractly expressed to capture in the data (e.g., "selecting movies that [they] feel would be interesting to watch" -$P_{CTRL}11$). Some participants expressed a focus on the dummy TITLE attribute ($P_{RT}7$).

Some participants expressed how interaction traces influenced their behavior. For instance, in Phase 2 (Revision), $P_{SUM}6$ expressed that "the 2nd time through, [they were] careful to check the different areas such as genre" to "[assess] which [they] thought were representative in each case." Similarly, $P_{SUM}13$ was "glad [they] got to go back and revise because [they] missed the creative type selection which [they]

corrected to contemporary fiction." Others relied on different features to facilitate their analysis, e.g., filters ($P_{RT}3$).

7 Discussion

Conscious v. Unconscious Biases. We observed both conscious and unconscious biases throughout the studies in this paper. While some results were as we hypothesized, there were a number of surprising findings (e.g., CTRL participants exhibited lower DPD metric values and AD metric values for several attributes; CTRL participants tended to pick more proportional political committees w.r.t. gender and party; surprise at seeing distributions in interaction traces corresponded to higher AD metric values; etc). We speculate that some of these findings may be the result of interaction traces leading to **amplified conscious biases**. That is, while awareness of unconscious biases may have been improved, the same intervention may have led to exaggeration of conscious biases. Additional studies are needed to understand the nature of the relationship. Regardless, these unconscious biases may be the result of lack of attention and unknown correlations in the data, or they could be the result of more dangerous implicit attitudes and stereotypes. From a behavioral perspective, the interactions users perform related to conscious or unconscious bias may look similar. Thus in future work when in-lab experiments are again feasible (outside of the COVID-19 pandemic), eliciting user feedback can be helpful to refine models of bias by users directly indicating if their focus was intentional or not [49] and by correlating outcomes with results of implicit association tests [26].

False Positives v. False Negatives. Given the imperfection of quantifying bias from user interactions, it begs the question: what is the harm of inaccuracy? A false positive (i.e., the system believes you are biased when you are not) could be frustrating to users, but we posit is relatively harmless apart from possible damaged ego. On the other hand, a false negative (i.e., the system believes you are not biased when you are) could be much more harmful, leading to unchecked errors. Furthermore, "false negative" circumstances are essentially the present norm, given that most systems do not attempt to capture bias in the analysis process.In such circumstances, biases would have propagated unchecked regardless. Hence we argue that a system that characterizes bias, even with low or unknown accuracy, can provide benefit in situations where bias may cause urgent problems.

Implications of Bias Definition. The bias metrics [46] used in these studies are formulated based on comparing user interactions to a **proportional** baseline. Given that visualizing these metrics in *real-time* sometimes resulted in changes in behavior and decisions, it begs the question whether this was the *right* way to nudge participants. Some participants expressed explicit goals to choose representative (or proportional) samples of political committees or movies, some aimed to choose equal samples (e.g., one movie from each GENRE or equal numbers of Democrats and Republicans), while still others intentionally biased their selections in other ways (e.g., all female politicians). Future work can explore the contexts in which different baselines of comparison are appropriate given Wall et al.'s metrics [46] or by introducing alternative metrics (e.g., that incorporate a user's prior beliefs using a Bayesian model). This has further ethical implications, in that designers must take on the social responsibility to choose visualization designs and bias computation mechanisms that reflect social values without unduly compromising user agency.

Design Implications. Based on feedback in formative studies, with or without *real-time* interaction trace visualizations, some participants found indirect ways to assess balance or bias in their committee choices (e.g., by applying filters and cycling through combinations of scatterplot axes to see the distribution of selected points). Hence, the affordances within the interface design can itself serve as a potentially powerful bias mitigation approach, promoting user awareness and enabling self-editing. Another example is enabling categorical attributes to be assigned to axes to see categorical distributions of selected points, which can offload oft complex management of cognitive decision making to a perceptual task. For instance, we varied this feature across formative studies and observed that categorical assignments enabled participants to easily identify clusters of points where they may not have selected any data points. Participants were able to choose *equally* across clusters or intentionally *bias* across clusters by visually inspecting for selected points. Particularly in situations where cognitive overload may prevent users from managing secondary views, designing the interface to afford indirect assessment of their choices may be a better alternative.

Study Limitations. Our formative studies suffered from a biased sampling of participants (mostly male Democrats). In the third and final formative study, we were able to correct for gender bias; however, due to sampling within our university, we were unable to recruit participants with diverse political party affiliations. We addressed these concerns in our fourth study using the crowdsourcing platform, Amazon Mechanical Turk. In spite of our best efforts, there was still a political imbalance (leaning Democrat), which we could not selectively recruit for due to our constraints of recruitment (within Georgia, high MTurk approval ratings). However, this experiment came with its own set of tradeoffs. While we were generally satisfied with participant engagement in the task as observed by their open-ended feedback, we were nonetheless limited in our ability to make rich observations. In addition, the task phrasing was intentionally vague to not bias participants toward any particular selection strategy. The cost, however, is noise in our data due to variable interpretations of the task. Future experiments may explore refining the task phrasing or exploring performance-based incentives to reduce the noise in collected user data.

8 Conclusion

In this paper we explored the effect of *real-time* and *summative* visualization of user interaction traces toward mitigating human biases in decision making tasks in the domains of politics and movies, where success was measured by changes in (1) behavior, (2) decisions, and (3) awareness. To study this effect, we conducted three formative in-lab experiments and a virtual crowdsourced experiment. We found that when both interventions were combined (real-time and summative), participants tended to perceive *both* to be less useful. Hence, the impact of heightened awareness may come at the expense of user experience. Furthermore, while we find some support for the impact of interaction traces (e.g., towards behavioral changes in interaction, for increasing awareness), we also find some surprising trends (e.g., CTRL participants' lower bias metric values, slightly more skewed political committees). These mixed results suggest that while interaction traces may lead to increased awareness of unconscious biases, they may also lead to amplification of conscious biases. Thus, while we find some promising support that interaction traces can promote conscious reflection of decision making strategies, additional studies are required to reach more conclusive results.

Acknowledgments

This work was supported in part by the National Science Foundation grant IIS-1813281 and the Siemens FutureMaker Fellowship. We thank the reviewers for their constructive feedback during the review phase. We also thank the Georgia Tech Visualization Lab for their feedback.

References

[1] https://github.com/nl4dv/nl4dv/blob/master/examples/assets/data/movies-w-year.csv. Accessed 2021-03-29.
[2] Georgia general assembly. https://www.legis.ga.gov. Accessed 2021-03-04.
[3] Georgia state senate committees. http://www.senate.ga.gov/committees/en-US/Home.aspx. Accessed 2020-04-02.
[4] Women in state legislatures in 2019. https://www.ncsl.org/legislators-staff/legislators/womens-legislative-network/women-in-state-legislatures-for-2019.aspx. Accessed 2021-03-04.
[5] Random name generator. https://github.com/treyhunner/names, 2014. Accessed 2019-07-25.
[6] Membership of the 115th congress: A profile. https://www.senate.gov/CRSpubs/b8f6293e-c235-40fd-b895-6474d0f8e809.pdf, 2018. Accessed 2019-07-25.
[7] J. H. Aldrich et al. *Why parties?: The origin and transformation of political parties in America*. University of Chicago Press, 1995.

[8] M. Bertrand and S. Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

[9] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1663–1672, 2014. doi: 10.1109/TVCG.2014.2346575

[10] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski. Characterizing guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):111–120, 2017.

[11] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou. The anchoring effect in decision-making with visual analytics. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

[12] E. Dimara, G. Bailly, A. Bezerianos, and S. Franconeri. Mitigating the attraction effect with visualizations. *IEEE transactions on visualization and computer graphics*, 25(1):850–860, 2019.

[13] E. Dimara, A. Bezerianos, and P. Dragicevic. The attraction effect in information visualization. *IEEE transactions on visualization and computer graphics*, 23(1):471–480, 2017.

[14] 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*, 2018.

[15] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang. Recovering Reasoning Process From User Interactions. *IEEE Computer Graphics & Applications*, May/June(March):52–61, 2009.

[16] P. Dragicevic. Fair statistical communication in hci. In *Modern statistical methods for HCI*, pp. 291–330. Springer, 2016.

[17] J. S. B. Evans and K. E. Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013.

[18] M. Feng, C. Deng, E. M. Peck, and L. Harrison. Hindsight: Encouraging exploration through direct encoding of personal interaction history. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):351–360, 2017.

[19] M. Feng, E. Peck, and L. Harrison. Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. *IEEE transactions on visualization and computer graphics*, 25(1):501–511, 2019.

[20] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[21] G. Gigerenzer. Fast and frugal heuristics: The tools of bounded rationality. *Blackwell handbook of judgment and decision making*, 62:88, 2004.

[22] G. Gigerenzer and H. Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143, 2009.

[23] G. Gigerenzer and W. Gaissmaier. Heuristic decision making. *Annual Review of Psychology*, 62:451–482, 2011.

[24] D. Gotz, S. Sun, and N. Cao. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pp. 85–95. ACM, 2016.

[25] A. G. Greenwald and L. H. Krieger. Implicit bias: Scientific foundations. *California Law Review*, 94(4):945–967, 2006.

[26] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

[27] T. Grüne-Yanoff and R. Hertwig. Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, 26(1-2):149–183, 2016.

[28] T. Jankun-Kelly and K.-L. Ma. A spreadsheet interface for visualization exploration. In *Proceedings of the Conference on Visualization'00*, pp. 69–76. IEEE Computer Society Press, 2000.

[29] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[30] D. Kahneman and S. Frederick. A model of heuristic judgment. *The Cambridge Handbook of Thinking and Reasoning*, pp. 267–294, 2005.

[31] P.-M. Law and R. C. Basole. Designing breadth-oriented data exploration for mitigating cognitive biases. In *Cognitive Biases in Visualizations*, pp. 149–159. Springer, 2018.

[32] T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*, 2019.

[33] S. Monadjemi, R. Garnett, and A. Ottley. Competing models: Inferring exploration patterns and information relevance via bayesian model selection. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[34] A. Narechania, A. Coscia, E. Wall, and A. Endert. Lumos: Increasing awareness of analytic behavior during visual data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2021. To appear.

[35] C. North, R. May, R. Chang, B. Pike, A. Endert, G. A. Fink, and W. Dou. Analytic Provenance: Process + Interaction + Insight. *29th Annual CHI Conference on Human Factors in Computing Systems, CHI 2011*, pp. 33–36, 2011. doi: 10.1145/1979742.1979570

[36] A. Press. Federal judge strikes down georgia abortion restrictions. *Georgia Public Broadcast (GPB)*, July 2020.

[37] B. F. Reskin, D. B. McBrier, and J. A. Kmec. The determinants and consequences of workplace sex and race composition. *Annual review of sociology*, 25(1):335–361, 1999.

[38] M. M. Ringel, C. G. Rodriguez, and P. H. Ditto. What is right is right: A three-part account of how ideology shapes factual belief. *Belief systems and the perception of reality. Oxon: Routledge*, 2019.

[39] V. Romo. Georgia's governor signs 'fetal heartbeat' abortion law. *NPR*, May 2019.

[40] P. Sengers, K. Boehner, S. David, and J. Kaye. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, pp. 49–58, 2005.

[41] P. T. Sukumar and R. Metoyer. A visualization approach to addressing reviewer bias in holistic college admissions. In *Cognitive Biases in Visualizations*, pp. 161–175. Springer, 2018.

[42] R. H. Thaler and C. R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[43] J. W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25, 1980.

[44] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

[45] A. C. Valdez, M. Ziefle, and M. Sedlmair. Priming and anchoring effects in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):584–594, 2018.

[46] 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. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

[47] E. Wall, L. M. Blaha, C. Paul, and A. Endert. A formative study of interactive bias metrics in visual analytics using anchoring bias. *Proceedings of the 17th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT'19)*, 2019.

[48] E. Wall, L. M. Blaha, C. L. Paul, K. Cook, and A. Endert. Four perspectives on human bias in visual analytics. In *Cognitive biases in visualizations*, pp. 29–42. Springer, 2018.

[49] E. Wall, J. Stasko, and A. Endert. Toward a design space for mitigating cognitive bias in vis. *IEEE Conference on Information Visualization (VIS)*, 2019.

[50] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, 2007.