Designing and Optimizing Cognitive Debiasing Strategies for Crowdsourcing Annotation

CHIEN-JU HO*, Washington University in St. Louis, United States
MING YIN*, Purdue University, United States

As artificial intelligence (AI) gets increasingly involved in our daily life, the biases in AI and the downstream negative social impacts have also become a pressing concern. In this position paper, we focus on one important source of AI biases—the biases in crowdsourcing annotations that AI is trained on—and advocate for leveraging cognitive debiasing strategies developed in the psychological literature to mitigate biases in crowdsourced annotations.

1 INTRODUCTION

Data has been the secret sauce for the rapid progress of artificial intelligence (AI), and crowdsourcing—the act of outsourcing a task to the crowd—has been one of the most ubiquitous paradigm for obtaining data from humans to enhance machine intelligence in a scalable and cost-effective manner [1, 11, 15, 17, 24, 25, 32, 36, 37]. Meanwhile, humans are notorious for being prone to various kinds of biases, which may lead to systematic deviations between the data collected from them and the ideal [10, 14, 19, 29, 35]. Even worse, these biases could have downstream effects and lead to negative and discriminatory outcomes that hurt the society [2, 3, 40]. Given the critical role that data plays in AI, the need for developing effective and practical methods to mitigate the biases in crowdsourced data is pressing.

In this position paper, we advocate for addressing this challenge by leveraging the cognitive debiasing strategies developed in psychological literature to mitigate the biases in crowdsourced annotation. In particular, we highlight two important research themes: (1) designing cognitive debiasing strategies for crowdsourcing annotation and understanding their empirical effects, and (2) optimizing the use of cognitive debiasing strategies with algorithmic frameworks. In addition to the research themes, we also advocate for the importance of having public, anonymized annotation datasets for performing future research on biases in crowdsourced data, as well as tools for researchers and practitioners to easily incorporate cognitive debiasing strategies during the data collection process.

2 THEORETICAL BACKGROUND: ORIGIN OF BIASES AND COGNITIVE DEBIASING

Decades of psychological studies have identified a wide variety of human biases that would lead to deviation from rationality and result in suboptimal decision-making [23]. The dual process theory (DPT) of reasoning provides a plausible account of the origination of these biases [5, 12, 22]. In particular, DPT specifies two processes through which thoughts may arise—Type 1 and Type 2 processing. Type 1 processing is fast, automatic, instinctive, and unconscious. On the other hand, Type 2 processing is slower, deliberate, rule-based, and conscious. While people usually utilize some combination of both intuitive and analytical processing during their decision making, it is believed that the default processing mode human brains would select is Type 1 processing. However, Type 1 processing is largely associated with the use of heuristics. Thus, excessive reliance of Type 1 processing would override Type 2 processing, trigger bias from humans, and lead to insufficient deliberation and unexamined decisions. Moreover, the risk of overusing Type 1 processing is particularly high when humans suffer from fatigue, sleep deprivation, and cognitive overload [5].
Based on DPT, a premise for “debiasing” is to enable people to decouple from their own automatic responses in decision-making that are resulted from Type 1 processing. In other words, the key to mitigate human biases is to have people actively engage in Type 2 processing and override Type 1 processing as needed. Addressing this key requirement, the concept of “cognitive debiasing” [33, 39] is proposed in the clinical and forensic domains. A variety of cognitive debiasing strategies have been proposed and evaluated, including raising people’s awareness of bias and motivating people to correct bias, enabling people to use situational cues to recognize the need of debiasing, instructing people to inhibit heuristic responses and analyze alternative solutions, etc [4, 5, 13, 20, 21, 28, 31, 34].

3 DESIGNING COGNITIVE DEBIASING STRATEGIES FOR CROWDSOURCING ANNOTATION

As a first step towards mitigating biases in crowdsourced data, established cognitive debiasing strategies can be adapted into the crowdsourcing contexts so that their effectiveness can be empirically evaluated. Based on when these strategies will be applied during a data annotator’s annotating process, a design space can be defined as the following:

- **Pre-annotation debiasing**: Debiasing elements can be designed before an annotator starts the annotating process. These elements could serve two main goals: First, increase annotators’ awareness of the existence and risks of their own biases, and promote their initiation in combating these biases (e.g., [19]). Second, help annotators to establish a physical and mental condition that is less vulnerable to biases (e.g., via short breaks and meditation [16, 27]).

- **In-annotation debiasing**: Debiasing elements can be designed to influence the annotators while they are determining the annotation. The main goal of these elements is to nudge annotators to consciously adopt Type 2 processing by, for example, formalizing their thinking process (e.g., as a checklist of actions or if-then rules) and grounding their annotations on sound data [26, 30].

- **Post-annotation debiasing**: Finally, debiasing elements can be designed after the annotator provides an annotation in a task to help annotators reflect upon and critique their own annotations. These interventions aim to both enable annotators to identify any potential biases that they have been subject to in their annotations, and allow annotators to re-examine their annotations comprehensively and systematically (e.g., via examinations of competing hypothesis, feedback, interactions between annotators, etc. [8, 9, 37]).

We highlight a few steps to take in order to establish a comprehensive understanding of the effectiveness of various debiasing strategies on reducing biases in the crowdsourced annotations: (1) identify a few “model annotation tasks,” i.e., tasks that we are aware of that annotators tend to suffer from different kinds of biases; (2) for each model task, explore how to operationalize each element in the design space of debiasing strategies into its specific context; (3) conduct randomized controlled experiments on crowdsourcing platforms to understand the empirical effects of various combinations of debiasing strategies for each of the model tasks. We note that in the empirical evaluations, not only the “benefits” brought by the debiasing strategies (e.g., reduction in data bias) should be measured, but also the potential “cost,” such as the change on annotation expense, annotation time, and annotator burnout. The collection of these information will serve as the foundation for optimally controlling the use of cognitive debiasing strategies in crowdsourcing data collection. We also advocate for sharing the datasets of human annotations that are collected through these empirical evaluations to allow the research community to perform further research on analyzing biases in these annotations.

4 OPTIMIZING COGNITIVE DEBIASING STRATEGIES FOR CROWDSOURCING ANNOTATION

With the understanding of the effects of cognitive debiasing strategies for crowdsourced data, the natural next question is how to optimally decide when and which strategies to use. To address this question, we lay out a general framework for optimizing cognitive debiasing strategies for crowdsourced data and identify specific challenges that need to be
addressed. Formally, let $d_t \in D$ be the parameters of debiasing strategies deployed at time $t$ (e.g., $d_t$ could specify the type of the debiasing strategy, the parameters of the strategy, etc), $S(d_t)$ be the set of data collected with this strategy (which could be a set of answers from workers), and $c(d_t)$ be the cost of deploying this strategy. The requester has a budget $B$ and time $T$ to make decisions. Let $L(\{S(d_t)\}_{t=1}^{T})$ be the loss the requester suffers from data $\{S(d_1), \ldots, S(d_T)\}$, collected with debiasing strategies $\{d_1, \ldots, d_T\}$. The goal of the requester can be formulated as the following constrained optimization problem.

$$
\begin{align*}
\text{minimize} & \quad d_1, \ldots, d_T \quad L(\{S(d_t)\}_{t=1}^{T}) \\
\text{subject to} & \quad \sum_{t=1}^{T} c(d_t) \leq B
\end{align*}
$$

We identify the following challenges in optimizing biasing strategies for mitigating biases in crowdsourced data.

- Aggregate data collected with cognitive debiasing strategies: The loss function in the optimization objective often depends on the ground truth of annotations, which are not known a priori. One approach is to leverage the techniques in truth discovery [6, 7, 18, 38, 41] to simultaneously infer the ground truth of annotations and the biases associated with the process from the collected data. To achieve this, in the literature, it is often assumed that data is independently drawn from some distribution characterized by given generative models. However, when we deploy debiasing strategies, we might alter the generative model and might even break the independence assumption. Therefore, to address the optimization problem, it is important to develop novel algorithms for aggregating data collected with debiasing strategies.

- Design online optimization algorithms: In practice, the requester often needs to decide whether and when to deploy debiasing strategies without having full access to the parameters in the optimization problem (e.g., the ground truths of labeling tasks are not known in advance). To approach this question, the requester needs to adaptively update the estimate of those parameters and make online decisions to optimize the overall loss. Therefore, developing online algorithms that can simultaneously optimize the objective and infer the latent parameters are important for solving this optimization problem.

- Determine the objective of the optimization with participatory design: There are various bias definitions, which are known to be incompatible with each other. In our optimization problem for bias mitigation, how should we decide on the optimization objective? Given the social-sensitive nature, we believe it is important to include relevant stakeholders in the loop to shape the objective of the problem. Therefore, developing participatory design approaches to elicit and aggregate stakeholders’ opinions in problem formulation is essential for this line of research.

- Develop tools for requesters to deploy the debiasing strategies: In order to maximize the outreach of the research outcomes, we need to make the research results easily applicable by requesters. Therefore, we argue developing easy-to-use tools for requesters to incorporate the debiasing strategies and the optimization algorithms during crowdsourced data collection is critical to maximize the impacts for this line of research.

5 CONCLUSION

In this position paper, we advocate to leverage cognitive debiasing strategies developed in psychological literature to mitigate biases in crowdsourced annotation. We highlight two important research themes on the design and optimization debiasing strategies. In particular, we highlight a few steps to take in order to establish a comprehensive understanding of the effectiveness of various debiasing strategies. We also layout an algorithmic framework for optimizing debiasing strategies and identify the technical challenges.
REFERENCES


Designing and Optimizing Cognitive Debiasing Strategies for Crowdsourcing Annotation


