Bias Mitigation in Decision-Making with Expert Advice

Axel Abels[1,2], Elias Fernández Domingos[1,2], Tom Lenaerts[1,2,3], Vito Trianni[4], and Ann Nowé[2]

[1] Machine Learning Group, Université Libre de Bruxelles
[2] AI Lab, Vrije Universiteit Brussel
[3] Center for Human-Compatible AI, UC Berkeley
[4] Institute of Cognitive Sciences and Technologies, National Research Council

Abstract. Individual and social biases pose a significant obstacle to the use of human guidance in complex decision-making. These biases introduce pervasive judgment errors which can unfairly disadvantage groups along for example ethnic or gender lines. Preventive measures include educating experts on the potential for biases to influence their decision-making. Effective mitigation however should also include active measures to identify and counteract biases. In order to design techniques capable of handling the latter, data on how humans share expertise with learning systems is essential. Such data not only enables a deeper understanding of the prevalence of biases in diverse populations, but also permits the creation of benchmark problems on which algorithms addressing this problem can be evaluated. The aim of this demonstration is thus two-fold. First, we intend to collect data to enable us to answer questions about human biases in collective decision-making. And second, through their participation in this experiment, participants will be sensitized to the presence of biases in their own decisions. Concretely, this demonstration would consist of a 15 minute quiz on various problems involving sensitive features, followed by a personalized report. This report would guide participants through their responses to demonstrate how they influenced the centralized aggregator's behavior in a simulated collective decision-making task, highlighting the quality and biases of their responses.

Keywords: Bias · Human advice · Collective-intelligence · Decision-making

Cognitive biases are systematic errors in judgment and decision-making resulting from cognitive limitations, individual preferences, and/or inappropriate heuristics [5]. When human groups deliberate, individuals tend to transmit these biases to others, which may lead the group as a whole to make a sub-optimal choice. For instance, through herding [4], a judgment error caused by a senior physician's cognitive biases can lead a group of medical experts as a whole to misinterpret a patient's symptoms, resulting in a substandard treatment.

While biases can be mitigated to some extent by educating participants and by adapting the way questions are posed, these measures are not always suffi-

cient. In such cases, biases should be identified and counteracted algorithmically in order to minimize their impact on the collective decision-making process.

To achieve this, we propose that groups of experts deliberate through an online platform for collective decision-making. This platform handles the exchange of information between individuals, to guide them to a final, less biased decision. We hypothesize that by managing the exchange of information through a decision-making platform, it will be significantly harder for humans to impose their individual biases on the final collective decision. Therefore, our overarching objective is to identify methods that enable humans to make better decisions collectively, while ensuring that transparent explanations are returned to experts.

While methods tackling this problem have been proposed [1,2], their evaluation has been restricted to synthetic experts or human data adapted from other tasks. The central aim of this demonstration is thus twofold: i) to collect data of human participants on three key topics, where cognitive bias plays a key role on decision; and ii) provide participants with some insights into their own biases, and how algorithmic approaches can help mitigate them.

During the demonstration, participants will make a series of decisions through our web-based platform[5]. Participants are repeatedly asked to evaluate the likelihood of different events, such as whether a headline is false, whether a candidate should be approved for a loan, or whether someone is likely to re-offend. In addition to features predictive of the true outcomes (such as prior offenses), participants are also presented with sensitive features associated with the instance they are judging, e.g., gender, age or ethnicity. The demonstration concludes with a report on the participant's answers. Specifically, this report breaks down the participant's biases[6] along gender, ethnicity and age features, and compares it to previous participants. This report will make explicit to the participant how strongly their predictions deviate when everything but the sensitive feature is fixed. In addition, participants are shown how strongly their advice influenced the choices made by a centralized decision-maker, which learns how to make the best (unbiased) decisions based on the answers from different participants for each question. We hope that this will highlight possible biases in the participant's answers and provide an understanding of how the central aggregator functions.

Requirements Because our experiment is accessible through a website, we do not have significant requirements, aside from a space where we could install a number of devices with internet access (say 5). While this experiment typically takes approximately 15 minutes to complete, participants do not need to be synchronized, they would thus be able to participate as they come. Aside from these devices, we would will also display a QR code with a link to the experimental platform, so that anywone could participate through their phone.

[5] A demonstration of the experiment is available at `https://bias.ulb.be`

[6] Based on a counterfactual analysis of their responses, see [3].

References

1. Abels, A., Lenaerts, T., Trianni, V., Nowé, A.: Collective decision-making as a contextual multi-armed bandit problem. In: International Conference on Computational Collective Intelligence. pp. 113–124. Springer (2020)
2. Agarwal, A., Dudík, M., Kale, S., Langford, J., Schapire, R.: Contextual bandit learning with predictable rewards. In: Artificial Intelligence and Statistics. pp. 19–26. PMLR (2012)
3. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. Advances in neural information processing systems **30** (2017)
4. Lorenz, J., Rauhut, H., Schweitzer, F., Helbing, D.: How social influence can undermine the wisdom of crowd effect. Proceedings of the national academy of sciences **108**(22), 9020–9025 (2011)
5. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. science **185**(4157), 1124–1131 (1974)