

Be the Expert - Experimental Design

October 2022

1 Introduction

When a group of humans deliberate freely, they often fail to identify the best decision. One of the reasons humans fail to do so is that they are inherently biased in their beliefs. In other words, their perception of what is true is affected by individual and social factors, for example stereotypes. When they can deliberate about a decision, they have a tendency to transmit these beliefs to others, potentially moving the whole group away from the best decision. For example, a senior doctor could spread a misinterpretation of symptoms to junior doctors and lead to a bad treatment being administered. The aim of this research is to prevent biases from affecting the group's decision-making in settings such as medical diagnostics, policy making and market forecasts. We propose to achieve this by having humans interact through a collective decision-making platform in charge of handling the exchange of information and the decision-making. The key hypothesis here is that this will make it substantially harder for humans to impose their biases on the final decision. Our overarching goal is therefore to identify methods that allow humans to reach collectively better decisions, while ensuring that transparent explanations are returned to the experts. To achieve this, we proposed and implemented algorithmic approaches capable of identifying and countering the biases (such as for example confidence bias, biases induced by a resistance to changes over time, or illusory correlations) which affect the judgment of participants in the collective decision-making task. The empirical performance of these algorithms has so far been established on synthetic experts, albeit ones designed to mimic human biases.

The primary goal of these experiments is to test the validity and performance through an evaluation with human participants. To accomplish this, we developed an online platform through which participants are able to interact with our decision-making system. In particular, participants are presented with a series of problems for which they must provide advice. Note that throughout this document, the use of the “expert” reflects the terminology used in the formalization of the problem which interests us. In practice, we don't expect participants to be experts on the problems we present them.



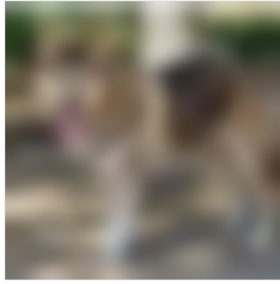
Because no information exchange between participants is considered, humans can participate independently from each other and we can sample subgroups later on which to run the alternative algorithms. Note that our algorithms assume that experts used in the collective decision-making solve the same problems. Participants will thus be presented one of 5 question sets in the same order.

We aim to recruit paid participants through prolific¹.

¹<https://www.prolific.co/>

Introductory Example 1

How likely is it that each of these pictures represents a dog?

☐ Very Unlikely
☐ Unlikely
☐ Undecided
☐ Likely
☐ Very Likely

☐ Very Unlikely
☐ Unlikely
☐ Undecided
☐ Likely
☐ Very Likely

☐ Very Unlikely
☐ Unlikely
☐ Undecided
☐ Likely
☐ Very Likely

Next

Careful, once you click on next, you can not go back.

Figure 1: Introductory example, typical answers would be "very likely", "very unlikely" and "undecided".

Headline 1

"White British pupils least likely to go to university, says research"

Figure 2: Example headline.

2 Problems

Each round in the experiment presents the participant with 3 instances on which they have to advice.

For example, the toy problem of Figure 1 is presented as an introduction to participants. The participant is presented with 3 photos of animals, and for each photo they have to estimate the likelihood that the photo represents a dog. Participants are presented with a choice between "very unlikely", "unlikely", "uncertain", "likely", and "very likely" for each of the photos.

The actual problem on which we collect advice is false headline detection. Figure 2 presents one example prompt for this problem. Participants will be shown such prompts and again be asked to estimate the likelihood. The following sections further detail the headline problem.

2.1 Fake news

The rise of social media has given fake news a unique channel to proliferate. One way of dealing with fake news is to ask human fact checkers for their opinion. In this problem we ask participants to play the role of such a fact checker. We present them with headlines and ask how likely it is that they are fake news.

We collected headlines matching the following criteria:

- The headline (implicitly or explicitly) contrasts two sensitive groups (e.g., *"Men more likely than women to say they are financially better off since last year"*)

- The headline should present a clear negative or positive outcome. For example, it is not clear that *"Poll: Kanye more popular with whites than nonwhites"* is positive or negative, but *"African-Americans, Hispanics, dying at faster rate of fentanyl overdoses than whites: analysis"* is clearly a negative outcome for African-Americans and Hispanics.
- The headline should be relatively universal, in the sense that most participants should understand the headline, regardless of whether they correctly assess its truthfulness.

Bias against groups in this setting would be captured by a higher likelihood of associating certain outcomes (negative or positive) with certain groups.

The following variables should be balanced across headlines: gender, ethnicity, age, outcome (positive or negative), truth.

We would thus need $(\# \text{ sensitive features} \times \# \text{ values per feature}) \times (\# \text{ number of outcomes} \times \# \text{ possible truth values}) = (3 \times 2) \times (2 \times 2) = 24$ headlines per treatment in order to balance the question set. For 5 treatments we thus need a multiple of 120 headlines.

We selected the headlines given in the headlines.tsv file. In particular, these headlines contain 40 instances for each of the 6 protected groups. Distributed over 5 treatments this should allow us to present 4 false and 4 true headlines of each protected group to each participant (thus assuming 16 iterations of 3 questions). Note however that any positive headline about one group is implicitly a negative headline about the complementary group. We thus have 8 false (4 positive and 4 negative) and 8 true (again 4 positive and 4 negative) headlines per group. For example, *"Across Age Groups, Whites Fared Worse in Employment Rates"* is both negative for Whites and implicitly positive for African-Americans.

3 Experimental Platform

We use otree (<https://www.otree.org/>) to implement the platform, and provide some screenshots in Figures 3 to 4 to illustrate the experiment's interface.

4 Research questions

The main goals of the experiment are to 1) measure whether learning algorithms can improve the performance of human collectives and 2) characterize biases in human participants. The performance of our algorithms can be compared to static (e.g., Majority Vote) or other learning algorithms (EXP4 [1] for example). Depending on the distribution of expertise, the best baseline varies. Our algorithm should in most cases surpass these baselines. This is the main effect we want to measure. In other words: given the same set of experts, does our algorithm outperform the average/exp4/... We first list variables we measure grouped by type

within factor: category, algorithm, question, sensitive group, iteration

between factor: treatment, question, and demographic information, group diversity

subjects: individual participants or groups of participants

dependent variables: reward, advice, response timing, and error

Given these factors, the following questions are our main focus:

- does the within-factor algorithm have an effect on the reward for a subject group (i.e., is one algorithm better)
- does the within-factor sensitive group have an effect on the advice or error of a subject individual (i.e., is an expert biased).

Note

Through this interface, you will be asked to make judgments involving sensitive characteristics. We do not expect any particular behavior from you. All your answers are collected anonymously.

Some of the questions presented in this study involve sensitive characteristics such as gender and ethnicity. Please be advised that exposure to scenarios involving these characteristics may potentially make you feel discomfort.

In the next few pages we will go through some example questions.

The aim of this short introduction is to familiarize you with the answer format that will be used throughout this questionnaire.


Each question will present you with 3 alternatives and ask you to rate the alternatives by how likely they are, from extremely unlikely to extremely likely.


Next


Careful, once you click on next, you can not go back.

Introductory Example 1

How likely is it that each of these pictures represents a dog?







☐ Very Unlikely

☐ Unlikely

☐ Undecided

☐ Likely

☐ Very Likely

☐ Very Unlikely

☐ Unlikely

☐ Undecided

☐ Likely

☐ Very Likely

☐ Very Unlikely

☐ Unlikely

☐ Undecided

☐ Likely

☐ Very Likely

Next

Careful, once you click on next, you can not go back.

Figure 3: Introduction page and example

Fake News Introduction

The rise of social media has given fake news a unique channel to proliferate. In order to deal with fake news, social networks rely on independent fact-checkers to warn readers of potentially misleading content. In the following questions we ask you to play the role of such a fact checker.

Note: in the context of this task we do not expect you to do any research, please provide answers based on your current knowledge.

Next

Careful, once you click on next, you can not go back.

Fake News Questionnaire

How likely is it that each of these headlines is real?

<div>Headline 1</div> <div><i>"African Americans and Latinos are more likely to be at risk for depression than Whites"</i></div> <div><div><input type="radio"/> Very Unlikely</div><div><input type="radio"/> Unlikely</div><div><input type="radio"/> Undecided</div><div><input type="radio"/> Likely</div><div><input type="radio"/> Very Likely</div></div>	<div>Headline 2</div> <div><i>"women have a higher opinion of themselves: study"</i></div> <div><div><input type="radio"/> Very Unlikely</div><div><input type="radio"/> Unlikely</div><div><input type="radio"/> Undecided</div><div><input type="radio"/> Likely</div><div><input type="radio"/> Very Likely</div></div>	<div>Headline 3</div> <div><i>"Old adults more likely to read than those who are under 30, says Pew report"</i></div> <div><div><input type="radio"/> Very Unlikely</div><div><input type="radio"/> Unlikely</div><div><input type="radio"/> Undecided</div><div><input type="radio"/> Likely</div><div><input type="radio"/> Very Likely</div></div>
--	--	---

Next

Careful, once you click on next, you can not go back.

Figure 4: Fake news introduction and example question

algorithm	treatment				
	1 (N=21)	2 (N=15)	3 (N=19)	4 (N=17)	5 (N=12)
exp4	6.0	5.0	7.0	5.0	5.0
metacmab	2.0	1.0	1.0	3.0	3.0
average	4.0	1.5	4.0	2.0	6.5
best expert	3.0	2.67	3.5	2.5	3.5
random	8.67	9.	9.33	9.	9.67

Table 1: Expected cumulative regret of different aggregators on different expert groups/treatments. Metacmab outperforms other aggregators in most cases, and achieves collective intelligence in all but one case.

In addition, the following questions could be of interest:

- do people consistently misevaluate some headlines (either individual headlines or headlines pertaining to a particular protected group)? In other words, does the headline or sensitive group have an effect on the error and/or advice? These results can also be further evaluated in terms of participant demographics.
- do people tend to be more conservative/assertive in their answers? I.e., are they more likely to select the "very" option? Again, results between demographic groups can be contrasted.
- do demographically diverse groups enhance the wisdom of the crowd? I.e., is the performance of heterogeneous groups better than that of homogeneous groups.
- Is response time a predictor of advice quality or bias?

4.1 Preliminary Results

We include some preliminary results involving voluntary participants ($N = 85$) as they anchor our estimates for the required number of paid participants.

Table 1 provides these preliminary results on the performance of all algorithms in terms of regret. Using a Wilcoxon test to test the significance of the improvements of metacmab, we obtain weak evidence for improvement against the average ($p = 0.072$) and the best expert ($p = 0.094$).

4.2 Sample size estimates

In order to collect further evidence for the performance difference between algorithms, we estimate necessary sample sizes based on results obtained with voluntary participants. We run our algorithms on sampled subsets of experts and use the results to estimate a partial η^2 which we then use to compute a sample size through G*Power [2] (ANOVA repeated measures, within factors; number of groups 5, number of measurements 2, correlation among rep measures 0.5, nonsphericity correction $\epsilon = 1$, effect size computed from partial η^2 , and α err probabilities of 0.10).

Based on a partial η^2 of 0.145 for groups of 10 experts, we deduce that approximately 20 groups (200 participants) would be appropriate to demonstrate the existence of the effect for moderate expert groups.

References

- [1] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.
- [2] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.