Analyzing Biases in Perception of Truth in News Stories and their Implications for Fact Checking

Mahmoudreza Babaei  
Max Planck Institute for Software Systems (MPI-SWS)  
Germany  
obabaei@mpi-sws.org

Abhijnan Chakraborty  
Max Planck Institute for Software Systems (MPI-SWS)  
Germany  
ochakr@mpi-sws.org

Juhi Kulshrestha  
GESIS - Leibniz Institute for the Social Sciences  
Germany  
juhi.kulshrestha@gesis.org

Elissa M. Redmiles  
University of Maryland  
United States  
eredmiles@cs.umd.edu

Meeyoung Cha  
Graduate School of Culture Technology, KAIST  
South Korea  
meeyoungcha@kaist.edu

Krishna P. Gummadi  
Max Planck Institute for Software Systems (MPI-SWS)  
Germany  
gummadi@mpi-sws.org

ABSTRACT

A flurry of recent research has focussed on understanding and mitigating the threat of “fake news” stories spreading virally on social media sites like Facebook and Twitter. In this work, we focus on how users perceive truth in viral news stories. To this end, we conducted online user-surveys asking people to rapidly assess the likelihood of news stories being true or false. Our goal is to quantify the extent to which users can recognize (perceive) the accurate truth-level of a news story (obtained from fact checking sites like Snopes and PolitiFact).

Our analysis of users’ implicit perception biases (i.e., inaccuracies in estimating truth-level of stories) reveals many interesting trends. For instance, we found that in the set of stories fact checked by Snopes, the perception biases are not correlated with the actual truth-level of the news stories. Our finding implies that there exist both true stories that are believed by users to be more false than they actually are as well as false stories that are believed to be more true than they actually are. We argue that the stories that are in need of being fact checked are the stories where users exhibit the largest perception biases. However, we show that existing fact checking strategies that rely on users to report stories they suspect to be false, would prioritize fact checking stories based on their actual truth-level rather than perception biases. We propose an alternative strategy to select stories with large perceived biases for fact checking.

KEYWORDS

Truth Perception Bias, False News, Fact Checking

1 INTRODUCTION

Recently, social media sites like Facebook and Twitter have been severely criticized by technologists, policy makers, and media watchdog groups for allowing fake news stories to spread unchecked on their platforms [7, 12, 13]. In response, Facebook1, Twitter2, Weibo3 and other social media sites are encouraging their users to report any news story they encounter on the site, which they perceive as fake. Stories that are reported as fake by a large number of users are prioritized for fact checking by (human) experts at fact checking organizations like Snopes4 and PolitiFact5. Stories deemed as false by fact checkers (following some principled method such as the Poynter’s Code of Principles6) are then prominently labeled as disputed. In essence, to counter the proliferation of fake news, social media sites are relying on their users’ perceptions of the truthfulness of news stories to select stories to fact check.

However, to date, few studies have focused on understanding how users perceive truth in news stories, or how biases in their perceptions might affect current strategies to detect and label fake news stories. Let us consider an example, to illustrate the need for fact checking systems to account for users’ truth perception biases. Figure 1 shows how users’ perceptions of the truth of a news story might differ from the actual truth (ground truth level) of that story. The stories that are likely to be reported (flagged) by most users for fact checking are the following two stories $S_1$ and $S_2$ that have the lowest perceived truth levels.

(1) **False Story $S_1$**: President Trump inherited a White House infested with cockroaches due to the careless behavior of his predecessor, Barack Obama.7

(2) **Mostly False Story $S_2$**: President Donald Trump changed the constitution to read ‘citizens’ instead of ‘persons’.8

However, note that these stories — which are false — are already perceived by users to be mostly or completely false. Most users have no misperception about these stories. We argue that in most scenarios, there is little to be gained by fact checking stories whose truth value is correctly perceived by most users, just as there is little

---


2https://www.snopes.com/ratings

3https://www.politifact.com/truth-o-meter/article/2013/nov/01/principles-politifact-pundifact-and-truth-o-meter

4https://www.poynter.org/international-fact-checking-network-fact-checkers-code-principles

5https://www.snopes.com/did-trump-alter-the-constitution

6https://www.politifact.com/truth-o-meter/article/2013/nov/01/principles-politifact-pundifact-and-truth-o-meter
In this paper, we present an in-depth analysis of how users perceive truth in news stories. Specifically, we present an exploratory analysis (characterization) of users’ truth perception biases for 150 stories fact checked by Snopes, and propose new approaches to leverage crowd wisdom to prioritize stories for fact checking to achieve three important goals:

- Goal 1: Remove false news stories from circulation.
- Goal 2: Correct the misperception of the users.
- Goal 3: Decrease the disagreement between different users’ perceptions of truth.

In summary, we make three primary contributions in this paper.

1. **Methodological**: We developed a new method for assessing users’ truth perceptions of content (e.g., news stories). Our test asks users to rapidly assess (i.e., at the rate of a few seconds per story) how truthful or untruthful the claims in a news story are. We conducted our truth perception tests on-line and gathered truth perceptions of 100 Amazon Mechanical Turk (AMT) workers from the USA [1] for each story.

2. **Empirical**: Our exploratory analysis of users’ truth perceptions yielded several interesting findings. For instance, (i) for many stories, the collective wisdom of the crowd (average truth rating) differs significantly from the actual truth of the story, i.e., wisdom of crowds is inaccurate, (ii) across different stories, we find evidence for both false positive perception bias (i.e., a gullible user perceiving the story to be more true than it is in reality) and false negative perception bias (i.e., a cynical user perceiving a story to be more false than it is in reality), and (iii) users’ political ideologies influence their truth perceptions for the most controversial stories (those stories with high variance in truth perception between users), it is frequently the result of users’ political ideologies (i.e., whether they support Democrats vs. Republicans) influencing their truth perceptions.

3. **Practical**: Our predictive analysis of users’ perception biases reveals the limitations of current strategies for selecting a small set of news stories to fact check based on how many users report the story as fake. We provide a proof of concept simulation for how our truth perception test and classifier can be used to achieve the three goals stated above for prioritizing stories for fact checking. However, please note that design of mechanisms to signal the fact checked label to the users such that they are receptive to them is out of scope of this study.

## 2 BACKGROUNDS

Over the recent years, a growing amount of efforts have been put into detecting false information by analyzing large-scale digitally logged user behavioral and social network data on the web. This section first reviews the literature on false news in light of its information types and detection methods. We then discuss the status-quo of the current fact-check systems and possible mitigation strategies.

### 2.1 False news detection problem

False news include two kinds of information types. The first kind is misinformation (i.e., a piece of information that happens to be wrong). Here, researchers have long investigated online rumors, a term to describe claims that are yet to be verified as ‘true’ [13, 21]. Based on theoretical studies on characterizing online rumor behaviors [16, 18], computer science researchers have developed rumor...
detection algorithms using features across multiple categories. Machine learning models have been tested based on features describing linguistic characteristics and diffusion patterns of rumors [14, 26]. Recently, a study [12] compared classification capabilities across such multiple feature categories and built an algorithm that achieves a competitive accuracy at an early stage of rumor spreading. Another line of studies proposed deep learning approaches to detect rumors without a labor-intensive feature engineering. Ma et al. [15] proposed a RNN-based algorithm to learn sequential information on online rumor spreading. From experiments on Twitter and Weibo, their approach outperformed existing feature-based algorithms and further tackled early detection problems. Other newly proposed deep learning models combine temporal activity patterns of spreaders and source characteristics into existing features. In particular, a model called CSI [24] showed the state-of-the-art performance in detecting rumors on social media.

The second information type is disinformation (i.e., a piece of information that is intentionally manipulated or wrong). This refers to “fake news” that is intentionally and verifiably false [25]. Detecting news articles that contains false claims is a challenging task because human evaluators have shown marginal improvements (66%) over random guesses (50%) in a crowdsourced study [11]. Such findings justify the need for an automated fact-checking system. As a preliminary step, recent studies have focused on a fake news problem known as clickbait article or stance detection, where news headline and the associated body text have a discordance relationship [5]. One research [4] developed a SVM model that predicts clickbait articles based on linguistic patterns. On the same dataset, another group suggested a neural network approach that measures textual similarities between headline and the first paragraph [23].

### 2.2 False news mitigation strategies

The above studies discuss various methods to identify false news stories on the Web. Once detected, false stories may be treated in several different manners, which are non-orthogonal:

- **Strategy 1: Remove false news stories from circulation.** Upon detection, a hard-line policy is to remove them entirely to block their spread on social platforms. Alternative soft-line policies would be to down-rank contents or label them as “false.” Previous research has shown that contents labeled as a rumor will less likely to spread further, indicating efficacy of labeling [7]. Various independent fact-checking agencies such as snopes.com and politifact.com act as distributed data sources for news platforms.

- **Strategy 2: Correct the misperception of the users.** Beyond reducing the circulation of false news stories, a more active mitigation strategy is to “correct” for its impact on social networks. While there exists a number of reputable fact-checking sources and studies, no study has examined how false or true news stories may be differently perceived by people irrespective of their ground truth. For example, false urban legends may spread even when people are aware of their veracity, simply because they are amusing. Perception towards political news stories may vary depending on one’s underlying political belief. In correcting the misperception of users, one needs to decide which stories to prioritize (i.e., the current study) and also design an effective methods for correction, which is beyond the scope of this paper.

- **Strategy 3: Decrease the disagreement between the users’ perceptions of truth.** In light of building a healthy public sphere that allows diverse ideologies, it is necessary to set a common ground on the intent and knowledge of news stories. Common ground can be achieved by helping to decrease the disagreement amongst users’ truth perceptions on news stories. For this, one needs to identify news stories of upmost disagreement to prioritize (i.e., the current study) and design an effective methods for mitigation (i.e., out of scope of this paper). This paper will discuss an effective method to identify news articles of upmost disagreement.

The remainder of this paper will introduce data and methodology (Section 3), expand on the needs for these mitigation strategies (Section 4) and suggest specific algorithms for each strategy (Section 5).

### 3 DATASET AND METHODOLOGY

As discussed in the earlier section, there has been no attempt to understand the effects of relying on the users’ truth perceptions for selecting the news stories for fact checking. To measure users’ truth perceptions, we design a test to assess how users perceive truth in news stories. We solicit users’ truth perceptions about a dataset of 130 stories that have already been fact checked, and thus have a known ground truth level to which we can compare users’ perceptions.

#### 3.1 Truth Perception Test

We sought to design a test methodology that can be used to assess people’s perceptions of the truth of news stories. We strongly encouraged the respondents of our tests to respond rapidly. At the beginning of our tests, the instructions given to the respondents were: “Please do not conduct any web search or use any online/offline resources for verifying or validating the claim presented to you. Please use your best judgment (your instinctive gut based guess within a few seconds) to label the claims.”

We then showed respondents a news claim, and asked them to label the claims as either ‘True’, ‘Mostly True’, ‘Mixture’, ‘Mostly False’ and ‘False’. In Figure 2, we show an example. We mapped the

![Figure 2: Example Amazon Mechanical Turk survey question we used to perform Truth Perception Tests.](image-url)
answer choices to a scale between -1.0 and +1.0, which we called the individual Perceived Truth-Level of user u (PTL_u). The aggregated value of PTL_u for a story S, called Perceived Truth-Level PTL(S), and is given by:

\[ \text{PTL}(S) = \frac{1}{N} \sum_{u=1}^{N} \text{PTL}_u(S) \]  

(1)

where N is the total number of users whose truth perceptions for the story S are being aggregated.

The news claims we showed users were drawn from the news stories that had been professionally fact-checked by Snopes.\(^3\) Snopes uses the same labels that we used in our test as answer choices to categorize news stories: False, Mostly False, Mixture, Mostly True, and True. We again mapped these values on a scale between -1.0 and +1.0, as shown in Figure 3. For each of these ground truth labels, we collected the 30 most recently fact checked news stories (from the Snopes’ Politics category) assigned that label, getting a total of 150 stories.

![Figure 3: Mapping ground truth labels of news stories on a scale between -1.0 and +1.0. The number of news stories collected corresponding to different ground truth labels for our dataset are also indicated in the figure.](image)

3.1.1 Test Development. We sought to ensure that our test was maximally robust to variations in test deployment and that we had accounted for a broad set of potential survey biases. As such, we conducted a number of micro-experiments to evaluate how, if at all, different test designs may influence results. Specifically, we evaluated:

- **Sample Effects**: Survey methodology literature supports that less-naive (e.g., expert) respondents may answer some types of survey questions in different ways than naive respondents [6, 20]. Additionally, the demographic composition of a survey sample is known to affect the generalizability of the results [2, 8]. As such, we compared the results of our test when run using MTurk Masters [1] vs. naive MTurkers and compared the results of our test when run using a census-representative sample of participants recruited by Survey Sampling International vs. MTurk Masters.

- **Answer Choice Effects**: Showing participants even- vs. odd-numbered options to answer (i.e., having a “middle” neutral option in the Likert scale answer choices or not) may effect the strength of the participants’ responses [22]. Thus, we compared the effect of using a 7 and 6 point Likert scale. Additionally, the text labels of the answer choices have also been shown to affect respondents’ answers to survey questions.\(^3\) Therefore, we compared the effect of using the texts used by Snopes (see Figure 2) with our 7 point scale ("I can confirm it to be true", "Very likely to be true", "Possibly true", "Can’t tell", "Possibly false", "Very likely to be false", and "I can confirm it to be false").

- **Satisficing and Incentive Effects**: Satisficing [10] is a commonly observed survey response effect in which respondents select what they consider the minimum acceptable answer, without fully considering their true feelings. Rapid response surveys like our truth perception tests may be at particular risk of satisficing because they encourage quick responses. Thus, we explored the effect of incentivizing participants to provide correct answers to evaluate whether satisficing may be affecting our test results.

In brief, we found no statistically significant differences across the survey variations for the proportion of correct answers. Additionally, we observed statistically significant high-correlation between our proposed measures (described in Section 4, including TPB), computed for our survey variations, with the Pearson correlation coefficients ranging from 0.90 to 0.96. Thus, we observe that our Truth Perception Test methodology appears to be robust to the variations in deployment and common survey biases. Further details about test validation can be found in Appendix 6.

3.2 Data Collection

We ran our final, validated test on MTurk, collecting 15,000 responses. Each MTurk worker saw 50 claims in a given survey, and no workers could take the survey more than once. Any MTurk worker over the age of 18 who resided in the US was eligible to participate in our survey.

3.3 Limitations

While we validate our truth perception tests extensively to ensure they are robust against design variations, our method does have some limitations which we discuss here. When users encounter and flag false news stories on the social media platforms, they are not only exposed to the claim or headline, but also to the source of the article, the images from the article, summary snippet or text of the article, and additional context for instance likes or shares for the story etc.

Our controlled experiments do capture the effect of the claim (or headline) of the news stories on the users, but they do not capture the effects of other factors as yet, and a promising direction of future work would be to design controlled experiments to measure the impact of the other factors.

4 USERS’ TRUTH PERCEPTIONS AND THEIR BIASES

As we argued previously, the major pitfall of current strategies for selecting stories for fact checking is that they ignore users’ truth perceptions of the stories. Thus in this section, by investigating users’ truth perceptions of news stories, we define measures to quantify perception bias of users along three dimensions: (i) bias in collective truth perceptions, i.e., aggregated wisdom of crowds, (ii) bias in individual truth perceptions, and (iii) disputability of individual truth perceptions.

\(^3\)https://www.snopes.com/

\(^\ast\)Prior work shows that it is always best practice to have text labels on Likert scale points, and thus we do not explore the omission of text labels [9].
4.1 Bias in Collective Wisdom of Crowds
Our investigation of the biases in users’ collective truth perceptions (i.e., wisdom of crowds) is motivated by the following two high-level questions:

1. Can wisdom of crowds be leveraged for assessing the truth level of news stories (i.e., fact checking stories)?
2. Is wisdom of crowds better at assessing the truth levels of true stories or false stories?

![Figure 4: CDF of the Perceived Truth Levels (PTL) for the news stories with different ground truth levels.](image)

Table 1: Comparison of Ground Truth Levels (GTL) and Perceived Truth Levels (PTL) of the news stories in the dataset.

<table>
<thead>
<tr>
<th>Truth-Levels</th>
<th>GTL True</th>
<th>PTL True</th>
<th>PTL False</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTL True</td>
<td>75%</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td>GTL False</td>
<td>32%</td>
<td>68%</td>
<td></td>
</tr>
</tbody>
</table>

To check whether collective truth perceptions are useful for assessing the truth levels of news stories, we begin by plotting the cumulative distribution of perceived truth levels (PTL) for the news stories with different ground truth levels (shown in Figure 4). The figure shows a high range in perceived truth levels for all stories, independent of their ground truth level – e.g., PTL values for true stories range from -0.22 to +0.5, while PTL values for false stories range from -0.71 to +0.32. While the distributions of perceived truth levels are different for false and true stories, they also exhibit a significant overlap – e.g., 34% of ‘true’ and 24% of ‘mostly true’ stories have negative PTL values, while 27% of ‘false’ and 30% of ‘mostly false’ stories have positive PTL values.

Our data suggests that while the collective wisdom of crowds have some predictive power in estimating the ground truth labels, their accuracy would be limited. Table 1 shows the limitations of using PTL values to predict GTL values for news stories at a coarse granularity. Here we consider a positive PTL (GTL) value to indicate the perception (ground truth) to be true and negative PTL (GTL) value to indicate the perception (ground truth) to be false. The table shows that the percentage of stories for which GTL and PTL values have same signs is 75% or lesser.

4.2 Bias in Individual Truth Perceptions
We next shift our focus to the bias in the individual truth perceptions of users. By perception bias (PB) of a user U for a news story S, we refer to the error or deviation between the ground truth level (GTL) of the story S and the user U’s perceived truth level (PTL_U) of the story S. Based on these truth levels of each story, we compute aggregated error or Total Perception Bias (TPB) as follows:

**Total Perception Bias (TPB)** of a story S captures the total error (gullibility or cynicality) in the users’ perceptions of truth levels of the story, and is given by

\[
TPB(S) = \frac{1}{N} \sum_{u=1}^{N} |PTL_u(S) - GTL(S)|
\]

where N is the total number of users whose truth perceptions of the story S are being aggregated. The entire area under the curve in Figure 5 depicts the total error (TPB) for the news story – "Bill Clinton was disbarred and fined over actions related to the Monica Lewinsky scandal".

![Figure 5: Users’ truth perception for the news story – “Bill Clinton was disbarred and fined over actions related to the Monica Lewinsky scandal”. The purple and black hachure areas corresponds to FNB and FPB respectively. The entire area shows the TPB.](image)

To capture the scenarios where the users may be gullible or cynical in their perceptions of truth in news stories, we compute the measures of False Positive Bias and False Negative Bias.

**False Positive Bias (FPB)** of a story S measures the gullibility of users in their perception of the truth level of the story, i.e., how much...
the users have over-estimated the truth level of the story by rating it to be more true than it is according to ground truth. Black hachure area in Figure 5 depicts the FPB for the story. The False Positive Bias of a story $S$ is computed as follows:

$$\text{FPB}(S) = \begin{cases} \frac{1}{N} \sum_{u=1}^{N} (PTL_u(S) - GTL(S)), & \text{if } PTL_u(S) > GTL(S) \\ 0, & \text{otherwise} \end{cases}$$

(3)

False Negative Bias (FNB) of a story $S$ measures the cynicality of users in their perception of the truth level of the story, i.e., how much the users have under-estimated the truth level of the story by rating it to be less true than it is according to the ground truth. In Figure 5, purple hachure areas shows the FNB for the story. The False Negative Bias of a story $S$ is computed as follows:

$$\text{FNB}(S) = \begin{cases} \frac{1}{N} \sum_{u=1}^{N} (GTL(S) - PTL_u(S)), & \text{if } PTL_u(S) < GTL(S) \\ 0, & \text{otherwise} \end{cases}$$

(4)

In Figure 6, we examine the relative contributions of FNB and FPB to the TPB of stories. On examining the distribution of TPB of stories (indicated by the red curve in Figure 6), we observe that a substantial fraction of stories have a perception bias of more than 0.5. To examine whether FPB or FNB dominate in their contribution to TPB, we separately plotted their distributions for the news stories (indicated by the orange and cyan curves in Figure 6). We found that many stories have values of FNB and FPB higher than 0.5, indicating that FPB is as large a concern as FNB for the stories our dataset.

In summary, we find that even though false positive and false negative biases may affect very different sets of stories, their overall impact on the total perception bias is comparable across all stories. Since users make the largest errors in judging the truth values of such stories (with high biases), in the next section, we argue that stories with high TPB (i.e., high FNB and high FPB) values can be prioritized for fact checking, if the goal for fact checking is to correct the misperception of users.

4.3 Disputability in Individual Truth Perceptions

To understand the extent to which the perceptions of users differ from each others’, in this section we analyze the disputability of a news story as the variance in the individual truth perceptions of users. The disputability of a story can be computed by considering the overall disagreement between all users, or disagreement between groups of users, as described next.

Variance in Perception Biases (VPB) or Disputability of a story $S$ captures the disputability in users’ truth perceptions of the story and is measured as follows:

$$\text{VPB}(S) = \frac{1}{N} \sum_{u=1}^{N} (PTL_u(S) - PTL(S))^2$$

(5)

The higher the value of VPB, the more the disagreement or dispute amongst the users about the story’s truth level.

I ideological Mean Perception Bias (IMPB) or Ideological Disputability of a story $S$ captures the disagreement in the truth perceptions of users with different political ideologies (i.e., Democrats and Republicans), and it is computed as:

$$\text{IMPB}(S) = |\text{MPB}_\text{Dem}(S) - \text{MPB}_\text{Rep}(S)|$$

(6)

Here the Mean Perception Bias (MPB) of story $S$ is defined as:

$$\text{MPB}(S) = \frac{1}{N} \sum_{u=1}^{N} (PTL_u(S) - GTL(S))$$

(7)

where PTL(S) is computed as the average truth perception of all users (or group of users) for the story $S$.

Next, we utilize these measures for analyzing the disputability in the truth perceptions of users to answer two high level questions:

1. How disputed are the truth perceptions of news stories and are true stories more or less disputed than false stories?
2. Are highly disputed stories the result of truth perceptions influenced by users’ political ideologies?

Significant variance in perception biases for many stories: When we compare the distributions of disputability for stories with different ground truth levels (shown in Figure 8), we observe that their disputabilities have similar distributions, i.e., true stories are as disputed as false stories. Thus, we observe that not only are the individual user perceptions of a story biased (as shown in earlier sections), but the biases also vary significantly between users.

![Figure 8: Distribution of disputability (variance in perception bias) of news stories across different ground truth levels.](image.png)
Next, to understand the extent to which truth perceptions in highly disputed stories are ideologically polarized, in Figure 9, we plot the Ideological Mean Perception Bias (IMPB) for all news stories in our dataset ranked in the increasing order of their disputability (VPB). We observe that disputability is correlated with ideological perception biases (Pearson correlation coefficient is observed to be 0.38). Many stories with high disputability also exhibit high ideological perception biases between Democrats and Republicans and vice-versa.

Our findings make the case for preferentially selecting such stories with high disputability for fact checking – they are the stories where users’ truth perceptions reveal a high degree of ideological polarization. Fact checking and labeling such stories could help establish ground truth for stories whose truth values are most contested, and help decrease the disagreement amongst users’ perceptions of truth.

5 IMPLICATIONS FOR FACT CHECKING SYSTEMS

In this section, we explore the implications of our findings about users’ perception biases on the fact checking of news stories. Today, fact checking sites like Snopes and PolitiFact employ human experts to assess the truth-level of news stories. While the experts can produce high quality and reliable trust assessments, they are limited in terms of the number of stories they can fact check. To select a small set of stories for fact checking, social media sites like Facebook and Twitter rely on their users to report stories that they suspect to be false. In other words, they are presently using TBPs for ranking the stories. However, our discussion earlier highlights the pitfalls of ignoring the biases in user truth perceptions when selecting stories for fact checking, and demands definition of clear goals for selecting stories. In this section, we present (a) how we can operationalize the different goals we introduced earlier, and (b) whether the current strategy used by social media sites can act as a suitable proxy for the three goals.

5.1 Operationalizing the Goals

Here, we describe how we can leverage users’ truth perceptions to prioritize stories for fact checking to achieve the three aforementioned goals.

5.1.1 Goal 1: Removing false news from circulation.

To ensure that false stories do not spread virally on social media platforms, stories which are flagged by many users as false could be selected for fact checking with a higher probability than true stories. We demonstrate in this work, for selecting false stories for fact checking we can use users’ truth perception values (PTL) as a suitable proxy. There may be other approaches for ranking stories to achieve this goal, for example, assigning higher weight to those users’ perceptions who are more experts, knowledgeable, or have a specific demographic feature. In this work, we consider PTL as a simple proxy for addressing this goal. As Table 1 depicts, although the accuracy is limited, the collective wisdom of crowds (PTL) have some predictive power of estimating GTL.

Footnote: Fact checkers are bound by the Poynter Code of Principles (https://www.poynter.org/international-fact-checking-network-fact-checkers-code-principles)
5.1.2 **Goal 2: Correct the misperception of users.**

The goal here is not to remove false stories (for instance, the social media platform may want to flag the stories rather than censor them to support freedom of speech), but to reduce the bias in truth perceptions of users. Therefore, as we discussed in section 4.2, to reduce the misperception, we can use total error that users make while judging a story. In addition, to quantify the extent to which a story S captures the total error (gullibility or cynicality) in the users’ perceptions of truth levels of the story, we defined TPB that we can rank based on it. Stories where users’ perceived truth levels potentially differ significantly from ground truth levels would be selected for fact checking. In practice, we would not have access to the GTL to compute total perception bias, thus here we proposed three methods to address this issue as follows:

I. Firstly we estimate the GTL of stories using PTL as discussed in the method for achieving the first goal call it $\text{GTL}_{\text{estimated}}$. Then we compute TPB of each stories as follows and then rank them.

$$TPB_{g1}(S) = \frac{1}{N} \sum_{u=1}^{N} |PTL_u(S) - GTL_{\text{estimated}}(S)|$$ (8)

The correlation between actual TPB and total error estimated to achieve goal one $TPB_{g1}$ is 0.37. If we simplify our ranking to just assigning high(TPB value higher than the median value of TPB over all stories) or low(TPB value less than the median value) TPB value to each story, then as table 2 depicts this strategy give us 62% accuracy.

II. We find that disputability of a story correlates highly with perception bias of a story, therefore by ranking according to disputability, we get a high fraction of high impact (large perception bias) stories in the top ranks. Figure 10 depicts the high correlation between disputability and TPB (the pearson correlation for our dataset is 0.42). As we show in table 2 assigning high and low TPB to stories considering disputability has 67% accuracy.

III. Using total perception bias (TPB) and combining these with the user demographics (e.g., users with different genders, educational qualifications, and political views), we have designed a prediction algorithm for identifying stories with high and low TPB that performs with an accuracy of 82%. Using the features described earlier, we apply Linear SVM and Logistic Regression classifiers (there is no significant difference between them) for assigning high and low TPB to each story and regression for ranking the stories and calculate the correlation. For training our classifiers, we use 5-fold cross-validation. In each test, the original sample is partitioned into 5 subsamples, out of which 4 are used as training data, and the remaining one is used for testing the classifier.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Correlation</th>
<th>Disputability</th>
<th>prediction algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimating GTL using PTL</td>
<td>0.37</td>
<td>0.42</td>
<td>0.52</td>
</tr>
<tr>
<td>Accuracy</td>
<td>62%</td>
<td>87%</td>
<td>82%</td>
</tr>
</tbody>
</table>

Table 2: Correlations and accuracies for different approaches used for ranking news stories to achieve second goal.

In summary, we proposed three approaches for achieving this goal, i.e., correcting misperception of users. We examine the performance of our prediction algorithm using demographic features and find it to perform the best amongst the three methods we argued. Table 2 demonstrate the difference performance among different methods.

![Figure 10: Comparing Total Perception Bias with disputability of news stories from our dataset. The pearson correlation coefficient is observed to be 0.42.](image)

5.1.3 **Goal 3: Decrease the disagreement amongst users’ perceptions of truth.**

For the society to have fruitful debates in the public sphere, it is essential for there to exist a common ground for the different, possibly disagreeing sections of the society. To ensure the existence of this common ground it may be desirable to decrease the disagreement amongst users’ truth perceptions by prioritizing those stories for fact checking where people disagree most about the truth value of the stories (i.e., there is highest variance in the truth perceptions reported by users). Several efforts have proposed ideas to reduce the polarization or disagreement between different group either by introducing diversity in the news that users are consuming [17, 19, 27] or by highlighting posts that evoke similar reactions from opposite political views [3].

In section 4.3, we argue that there are many stories with high disputability, in which there are a significant disagreement between users’ truth perceptions and also reveal a high degree of ideological polarization. Thus for addressing this goal of prioritizing stories which have high disagreement between users’ perceptions, we use two metrics of Aggregate Disputability (i.e., variance in users’ truth perceptions of a story) and Ideological Disputability (i.e., different between the truth perceptions of Democrat-leaning and Republican-leaning users). For achieving the this goal, we prioritize stories with high values of these two disputability measures.

To select stories with high ideological perception bias, ideally, we would need background knowledge about the political leanings of the users. We observed in the previous section, disputability (variance in perception biases) is highly correlated with ideological perception bias and disputability can be computed from users’ truth perceptions without any knowledge of their political leanings. So, another practical strategy would be to use stories’ disputability as a proxy measure for their ideological perception bias, when selecting stories. Figure 9 depicts that there is a high correlation between disputability and Ideological perception bias.
5.2 Pitfalls of the Current Strategy

As we discussed, the current strategy ranks stories based on PTL. The correlation between the stories ranked based on PTL and their ground truth level (GTL) is almost 0.4. Table 1 depicts that PTL have some predictive power in estimating the ground truth labels, however, their accuracy would be limited. We also compute the correlation between stories ranked by PTL and TPB (achieving the second goal) as well as Disputability and ideological disputability (achieving the third goal). We can see in Table 4 that the correlations are low using PTL as a proxy for achieving the different goals. Rather, using disputability to rank stories would work better than relying on PTL to achieve the second and third goals.

Table 3: Summary of the three goals for fact checking, along with how they can be achieved using measures based on ground truth information as well as our proposed operationalization strategies for achieving them.

<table>
<thead>
<tr>
<th>Goals</th>
<th>Ground Truth Measures</th>
<th>Proposed Operationalization Strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal 1: Removing false news from circulation</td>
<td>GTL (Ground Truth Level given by Snopes)</td>
<td>Estimating GTL using PTL then computing TPB</td>
</tr>
<tr>
<td>Goal 2: Correcting the misperception of users</td>
<td>TPB(S) = ( \sum_{u=1}^{N} PTL(S) - GTL(S) )</td>
<td>Disputability or VPB</td>
</tr>
<tr>
<td>Goal 3: Decreasing the disagreement amongst users' perceptions of truth</td>
<td>Disputability: VPB(S) = ( \sum_{u=1}^{N} (PTL(S) - PTL(S))^{2} )</td>
<td>Disputability or VPB</td>
</tr>
</tbody>
</table>

Table 4: The third and fourth rows show the correlation between stories ranked based on PTL and others as well as disputability and others respectively.

<table>
<thead>
<tr>
<th>Stories ranked based on</th>
<th>GTL</th>
<th>TPB</th>
<th>Disputability</th>
<th>Ideological disputability</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTL</td>
<td>0.4</td>
<td>0.1</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Disputability</td>
<td>0.42</td>
<td>0.12</td>
<td>0.11</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Figure 11 shows the ground truth levels of stories selected from Snopes by current PTL-based strategies that achieve the first goal and disputability-based strategy that is a suitable proxy for achieving both second and third goals. We can see that our new strategy selects stories with different ground truth levels in more balanced manner compared to current strategies that select stories that are predominantly false. Thus, current strategies are biased towards false stories and fail to select true stories to be fact checked, even in cases when the true story may (incorrectly) be perceived to be false.

Table 4: The third and fourth rows show the correlation between stories ranked based on PTL and others as well as disputability and others respectively.

Figure 11: Composition of ground truth levels of stories selected by PTL-based strategy, disputability-based strategy.

A TEST DESIGN VALIDATION

As we mention in the main body of the paper, we sought to ensure that our test design was robust to variations in deployment and to common survey biases. We evaluated the effect of the variations described in Section 3.1.1 of the paper as follows:

- We evaluate the similarity between the distribution of answers to each survey using a \( X^{2} \) test of independence. For comparing the distribution of answering, we consider distribution of...
answering by workers that both survey variations receive and using chi square test we investigate if they are independent or not.

- We evaluate similarities in answering accuracy by computing the sign of PTL and GTL. If the sign is the same then we consider the claim to be judged correctly. We then compare the proportion of correct answers in each survey using a $X^2$ proportion test. Using chi square test we assess if they are dependent or not.
- We evaluate similarities in the TPB and MPB values between the two surveys using Pearson correlation. We compute the correlation between TPB or MPB values of both survey variations to see whether our tests are robust or not.

Next, we describe the results of each of our design effect evaluations in more detail:

### Sample Effects
The survey variations we compare in the context of sample effects are:

- Running our test using MTurk Masters vs. naive MTurkers.
- Running our test using census-representative sample of participants recruited by Survey Sampling International (SSI participants) vs. MTurk Masters (experts participants).

Table 5 depicts that both distribution of answering and accuracy of judgments are independent (fail to reject H0) of types of survey respondents. Last column also shows that there is a significantly high correlation between TPB of claims of different surveys with different workers samples. This means that a particular claim has a close value of TPB in different surveys which confirm that our measure is robust against the sample effects.

### Answer Choice Effects
We evaluated the answer choice effects by comparing 6, 7 point scale with Snopes’ 5 point scale. Table 6 depicts that both distribution of answering and accuracy of judgments are independent (fail to reject H0) of the types of answer choices in the surveys. Significant high correlation between TPB of claims of different surveys with different answer choices is shown in last column.

### Satisficing and Incentive Effects
To investigate the impact of satisficing and incentives, we designed a survey in which we gave respondents incentives for answering correctly. At the beginning of the survey, we told the participants:

"In addition to the amount promised for the task, for each of your judgements which CORRECTLY matches the actual truth status of the claims, we will pay you 5 cents as a bonus. For example, if you judge a claim to be ‘True’, or ‘Mostly True’, and the claim is actually true, then you’ll get 5 cents for the claim. Similarly, to get the bonus for an actual false claim, it should be judged by you as ‘False’ or ‘Mostly False’. Finally if you judged the claim as ‘Mixture’ and the claim actually is mixture or mostly true/mostly false you will earn bonus.” To ensure that participants do not use online or offline resources to estimate the truthfulness of the claims we showed a timer in each page and told them: “If your judgment for each question takes more than 15 seconds then there would not be any bonus, even if you answer the question correctly.”

<table>
<thead>
<tr>
<th>Surveys</th>
<th>Chi square dependency of Dist-ANS</th>
<th>Chi square dependency of Acc</th>
<th>Correlation of TPB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTurk Masters &amp; MTurk naive</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>MTurk naive &amp; SSI workers</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>0.89</td>
</tr>
</tbody>
</table>

**Table 5: Sample effects**

<table>
<thead>
<tr>
<th>Surveys</th>
<th>Chi square dependency of Dist-ANS</th>
<th>Chi square dependency of Acc</th>
<th>Correlation of TPB</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-pt scale &amp; 8-pt scale</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>0.94</td>
</tr>
<tr>
<td>7-pt scale &amp; 5-pt scale</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>0.98</td>
</tr>
</tbody>
</table>

**Table 6: Answer choice effects**

<table>
<thead>
<tr>
<th>Surveys</th>
<th>Chi square dependency of Dist-ANS</th>
<th>Chi square dependency of Acc</th>
<th>Correlation of TPB</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-pt scale &amp; incentive and 5-pt scale</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>0.85</td>
</tr>
<tr>
<td>7-pt scale &amp; incentive and 7-pt scale</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>Chi-value:0.0 p-value=1.0</td>
<td>0.92</td>
</tr>
</tbody>
</table>

**Table 7: Satisficing and Incentive effects**

To see if incentivizing has any effect or not we compare the incentivized survey with the unincentivized survey. Table 7 depicts that incentivizing has no effect.

Figure 12: Accuracy of judgments (wisdom of crowds) for different design surveys.

In brief, we found no statistically significant differences across the survey variations for the proportion of correct answers. Additionally, we observed statistically significant high-correlation between our proposed measure of TPB, computed for our survey variations, with the Pearson correlation coefficients ranging from 0.90 to 0.96.
and 14 summarize the results of comparing TPB and MPB across various surveys. We thus conclude that our test is relatively robust and consequently the test variations, which show very similar results across variants.

Figures 12 depicts that wisdom of crowds (accuracy of judging by users) is very similar across different survey variations. Figures 13 and 14 summarize the results of comparing TPB and MPB across the test variations, which show very similar results across variants. We thus conclude that our test is relatively robust and consequently useful for application in industry settings and future research on content misperceptions.

REFERENCES