The Effects of Crowd Worker Biases in Fact-Checking Tasks

Tim Draws
Delft University of Technology
Delft, The Netherlands
t.a.draws@tudelft.nl

David La Barbera
University of Udine
Udine, Italy
labarbera.david@spes.uniud.it

Michael Soprano
University of Udine
Udine, Italy
michael.soprano@uniud.it

Kevin Roitero
University of Udine
Udine, Italy
kevin.roitero@uniud.it

Davide Ceolin
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
davide.ceolin@cwi.nl

Alessandro Checco
University of Rome La Sapienza
Rome, Italy
a.checco@sheffield.ac.uk

Stefano Mizzaro
University of Udine
Udine, Italy
mizzaro@uniud.it

ABSTRACT

Due to the increasing amount of information shared online every day, the need for sound and reliable ways of distinguishing between trustworthy and non-trustworthy information is as present as ever. One technique for performing fact-checking at scale is to employ human intelligence in the form of crowd workers. Although earlier work has suggested that crowd workers can reliably identify misinformation, cognitive biases of crowd workers may reduce the quality of truthfulness judgments in this context. We performed a systematic exploratory analysis of publicly available crowdsourced data to identify a set of potential systematic biases that may occur when crowd workers perform fact-checking tasks. Following this exploratory study, we collected a novel data set of crowdsourced truthfulness judgments to validate our hypotheses. Our findings suggest that workers generally overestimate the truthfulness of statements and that different individual characteristics (i.e., their *belief in science*) and cognitive biases (i.e., the *affect heuristic* and *overconfidence*) can affect their annotations. Interestingly, we find that, depending on the general judgment tendencies of workers, their biases may sometimes lead to more accurate judgments.

CCS CONCEPTS

• **Information systems → Crowdsourcing**; • **General and reference → Estimation**.

KEYWORDS

Truthfulness, Crowdsourcing, Misinformation, Explainability, Bias

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3534629

ACM Reference Format:
Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. 2022. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3531146.3534629

1 INTRODUCTION

Human computation is a powerful tool to assess information quality and identify misinformation. Although experts are considered most reliable when it comes to truthfulness judgments, recent research has shown that crowd workers can also reliably perform such fact-checking tasks [23, 36–39] and assess information quality across multiple truthfulness dimensions or quality aspects [26, 43, 44]. Crowdsourced fact-checking is now widely used in academic research [30, 32, 40, 41, 48] and has already found applications in industry [1, 34]. However, because crowdsourcing often relies on contributions from large groups of laypeople with different backgrounds, expertise, and skills, systematic biases among those workers may reduce the quality of their annotations [9, 11, 21]. For example, in fact-checking tasks, factors such as workers' political affiliation or their general trust in politics may affect their ability to correctly identify misinformation.

Identifying systematic biases in crowdsourced fact-checking is a relevant matter. Because expert-provided assessments are expensive and slow to gather, crowdsourced truthfulness judgments are often used as training sets for supervised machine learning methods. The presence of bias in training data may lead to bias in the classification performed by these systems. Moreover, such biases might affect the accuracy (or even question the feasibility) of human-in-the-loop hybrid systems that try to identify misinformation at scale by combining experts, crowd, and automatic machine learning systems [7]. Unveiling these systematic biases would support a more reliable collection of crowdsourced training data and enable bias mitigation methods for existing data sets.

In this paper, we investigate which systematic biases may decrease data quality for crowdsourced truthfulness judgments. Our work is guided by the following research questions:

RQ1. What individual characteristics of crowd workers and statements may lead to systematic biases in crowd workers' truthfulness judgments?

RQ2. What cognitive biases can affect crowd workers' truthfulness judgments?

RQ3. Are different truthfulness dimensions affected by different biases?

To address these research questions, we first conducted an exploratory study on an earlier collected data set containing crowdsourced truthfulness judgments for political statements (Section 3). These data also contain information on the political leaning of statements as well as individual worker characteristics (e.g., workers' level of education and political leaning). We used the findings from these exploratory analyses to formulate specific hypotheses concerning which individual characteristics of statements or workers (RQ1) and what cognitive worker biases (RQ2) may affect the accuracy of crowd workers' truthfulness judgments. To test these hypotheses, we subsequently conduct a new, preregistered crowdsourcing study (Section 4). Our findings suggest that crowd workers' degree of belief in science matters in this context (RQ1), that workers generally overestimate truthfulness, and that their annotations can be biased due to cognitive biases such as the *affect heuristic* and *overconfidence* [9] (RQ2). We also find exploratory evidence that different truthfulness dimensions may be affected by these biases to different degrees (RQ3, Section 5).

Supplementary materials related to this paper (e.g., task design, preregistration, data sets, and analysis code) are openly available: https://osf.io/8yu5z/. This study had been approved by an ethical committee at one of our institutions.

2 RELATED WORK

In this section, we summarize previous research on fact-checking performed by means of crowdsourcing (Section 2.2) and earlier work that investigated bias in user-generated data (Section 2.1).

2.1 Crowdsourced Fact-checking

To allow fact-checking tasks to scale and keep up with the large amounts of information posted online, previous work has studied methods to address the misinformation issue using crowd-powered systems. Many of those studies employed crowdsourcing to collect *truthfulness* judgments [32, 40, 45]. For example, La Barbera et al. [23], extending previous work by Roitero et al. [36], studied the effect of both judgment scale and assessor bias when fact-checking political statements. Their work demonstrated that coarse-grained scales are preferred by workers and that workers' political background is the main bias influencing workers' ability to effectively assess misinformation statements.

Roitero et al. [37] used crowdsourcing to collect thousands of truthfulness labels on multiple data sets for political fact-checking, employing different scales. They found that adjacent categories in the assessment scale can be grouped together to increase both worker effectiveness and agreement and that different scales lead to a similar agreement levels. More recently, Soprano et al. [43] re-assessed Roitero et al.'s [37] statements. Breaking down truthfulness on a multidimensional scale, they found that using multiple dimensions measures different aspects of the misinformation statement evaluated by the crowd workers. Roitero et al. [38, 39] focused on fact-checking statements related to the COVID-19 pandemic. Besides reporting results on crowd effectiveness and agreement, they performed a longitudinal study and presented an in-depth study on how the crowd's effectiveness changes when it is asked to perform fact-checking over different time spans. They also provide a failure analysis to investigate the statements that are mislabeled by crowd workers. Epstein et al. [13] deployed a survey to 1000 Americans to study their perceived trust in popular news websites, finding that mainstream sources are usually more trusted than fact-checking websites or hyper-partisan sources. Bhuiyan et al. [3] adopted a similar approach by surveying students enrolled in journalism or media programs about information dealing with climate change. Ghenai and Mejova [16] used crowdsourcing and machine learning to track misinformation on Twitter. Pennycook and Rand [31] crowdsourced news source quality labels. Giachanou and Rosso [17] developed a tutorial on online misinformation and fact-checking, with a focus on social media data.

2.2 Bias Investigation in User-Generated Data

Studying bias in user-generated data concerns multiple academic domains. Recent work has presented surveys on potential effects derived from biases on the web in general [2] and recommender systems specifically [5].

Other research has focused specifically on issues related to bias management in user-generated data. Yue et al. [51] investigated presentation bias in click-through data generated by a search engine, Love [25] studied different user biases in peer-assessment methods, Chandar and Carterette [4] estimated click-through bias in cascade models for information retrieval, and Muchnik et al. [29] focused on social influence bias. Yildirim et al. [50] and Lee [24] studied bias in user-generated data dealing with news media. Furthermore, several works investigated the role of commonly occurring cognitive biases (e.g., *exposure effects* and the *confirmation bias*) in web search on debated topics [10, 12, 33, 35, 49].

Several pieces of work focused on bias investigation in crowdsourced data: Eickhoff [11] investigated the effect of common cognitive biases when using crowdsourcing for relevance judgment tasks, Hube et al. [21] analyzed the effect of workers' opinions in subjective tasks, Draws et al. [9] created a checklist to fight common cognitive biases.

3 EXPLORATORY STUDY

To identify specific hypotheses concerning our research questions, we conducted an exploratory study using a publicly available data set. This section details this exploratory study and describes the hypotheses we formulated as a result.

3.1 Data

We conducted our exploratory study on a data set collected and published by Soprano et al. [43].[1] The data set is composed of crowdsourced truthfulness judgments for 180 statements from two

[1] The data set is publicly available at https://github.com/KevinRoitero/crowdsourcingTruthfulness.

political fact-checking websites: *Politifact* [47] and *ABC*.[2] *Politifact* is a collection of more than 10000 statements from mainly US politicians, labeled by experts on a six-level truthfulness scale containing the categories *pants-on-fire*, *false*, *barely-true*, *half-true*, *mostly-true*, and *true*. *ABC* is a collection of more than 500 statements on Australian politics that are first labeled by experts on a fine-grained semantic scale with more than 30 levels and then mapped into a three-level scale with the labels *negative*, *in-between*, and *positive*. The 180 statements in the data set had been selected by sampling, per truthfulness level, 10 statements for each of the main two political parties present in the *Politifact* and *ABC* data sets (i.e., Republican and Democrat for *Politifact*; Liberal and Labor for *ABC*). This resulted in $10 * 2$ (political parties) $* 6$ (truthfulness levels) $= 120$ *Politifact* statements, and $10 * 2$ (political parties) $* 3$ (truthfulness levels) $= 60$ *ABC* statements.

Soprano et al. [43] asked crowd workers to reassess the 180 statements in a set of Human Intelligence Tasks (HITs). In addition to the *Overall Truthfulness* of the statement, they employed a multidimensional truthfulness scale composed by the following dimensions: *Correctness*, *Neutrality*, *Comprehensibility*, *Precision*, *Completeness*, *Speaker's Trustworthiness*, and *Informativeness*.[3] They recruited 200 US-based crowd workers from *Amazon Mechanical Turk* (MTurk)[4] who subsequently performed the HITs on a private server. Each HIT required workers to assess the truthfulness of 11 statements: six from Politifact (i.e., one for each truthfulness level), three from ABC (i.e., one for each truthfulness level), and two additional handcrafted statements that the authors used for quality control (i.e., to identify malicious or low-quality workers). The statement sets were also balanced in terms of political parties (i.e., all political parties were equally represented). Workers assessed the truthfulness of statements using a set of five-point Likert scales ranging from "strong disagreement" (-2) to "strong agreement" (2). Each statement was evaluated by 10 distinct workers. Before judging the truthfulness of statements, each worker completed a mandatory questionnaire (i.e., to record their age group, level of education, income, general political view, favored political party, opinion on the US southern border, and opinion on US environmental regulations) and a *Cognitive Reflection Test* (CRT) [15] to assess their cognitive reasoning abilities. We used these bits of worker-specific information as independent variables for our exploratory study.

3.2 Data preprocessing

We performed several preprocessing steps on the data described in Section 3.1 so that they fit our purposes. Specifically, we transformed several scales and computed worker-related bias metrics.

3.2.1 Scale Transformations. Each statement in the data set described in 3.1 contains a truthfulness judgment from either *Politifact* or *ABC*, as well as several truthfulness judgments from crowd workers. However, these different types of judgments all adhere to different (ordinal) scales: whereas *Politifact* judgments are made on a six-level scale, *ABC* judgments are made on a three-level scale and worker judgments are made on a five-level (Likert) scale. Comparing the different assessments required that we align all of those scales. Assuming that all the *Politifact*, *ABC*, and Likert scales are linear equally spaced scales,[5] we converted the *Politifact* and Likert scales to the three-level scale used by *ABC*. This meant transforming each judgment to one of three labels: *negative* (-1), *neutral* (0), and *positive* (1):

- Politifact: we mapped *pants-on-fire* and *false* into *negative* (-1), *barely-true* and *half-true* into *neutral* (0), and *mostly-true* and *true* into *positive* (1).
- ABC: *negative* and *positive* maintain the same semantic meaning, while *in-between* was mapped into *neutral* (0).
- Likert scale: we mapped -2 and -1 into *negative* (-1), 0 into *neutral* (0), and $+1$ and $+2$ into *positive* (1).

3.2.2 Annotation Bias Metrics. We computed three different metrics to quantify and evaluate annotation bias. We considered both *external* errors (i.e., when comparing crowd annotations with the ground truth) and *internal* errors (i.e., when comparing crowd annotations with other crowd annotations for the same set of items).

- **External Error** (eE): the difference between a worker's Overall Truthfulness judgment and the respective item's ground truth label as assessed by the expert. This metric assesses the degree to which a crowd worker overestimates or underestimates the Overall Truthfulness of a particular statement. Its values range in $[-2, 2]$: for example, if the ground truth label (i.e., from *Politifact* or *ABC*) for an item is positive (1) but the crowd worker's annotation is negative (-1), eE for this particular annotation is equal to -2.
- **External Absolute Error** (eAE): the *absolute* difference between a crowd worker's Overall Truthfulness judgment and the respective item's ground truth label. Its values range in $[0, 2]$. In contrast to eE, this metric quantifies the *magnitude* of bias. It is the absolute value of eE.[6]
- **Internal Error** (iE): the difference between a worker's judgment and the average judgment of other crowd workers for the same statement. Its values range in $[-2, 2]$. We computed nine such metrics in total, i.e., one for Overall Truthfulness, one for workers' confidence, and one for each of the seven truthfulness dimensions. These nine metrics quantify the degree to which a specific annotation was above or below other crowd workers' judgments on a particular dimension.

3.2.3 Worker Bias Metrics. We computed aggregate bias metrics that evaluate each worker's individual degree of bias based on the annotation bias metrics described in Section 3.2.2. Specifically, we compute each worker's mean eE (eME), mean eAE (eMAE), and – for Overall Truthfulness, confidence, and each of the seven dimensions – mean iE (iME). These 11 worker-specific metrics are used as dependent variables for the exploratory study.

[2] See https://www.abc.net.au/news/factcheck/.

[3] For the rationales behind and detailed discussion of these extra dimensions, we refer to [43, Section 4.3].

[4] https://www.mturk.com/

[5] The same assumption has been made in previous studies and discussed in more detail by Roitero et al. [37, Section 3.3].

[6] We did not use the mean squared error here to avoid penalizing larger errors (e.g., an error of 2 should not be more than the double the error of 1).

3.3 Exploratory Analyses

We performed a series of exploratory analyses on the public data set described in Section 3.1 to identify potential systematic biases in crowd workers' truthfulness judgments. Specifically, we used different worker-related attributes (e.g., political views and average time per judgment) as independent variables and the aggregate worker bias metrics described in Section 3.2.3 as dependent variables. We found the workers in the data set to be quite balanced in terms of demographics (e.g., age group and income) and political views (e.g., conservative versus liberal orientation). Note that the results we report in this subsection (e.g., p-values from hypothesis tests) are exploratory. We only conducted these analyses to identify concrete hypotheses that we would test on novel data (see Section 3.4).

3.3.1 Exploring Worker's eME. We began our exploratory analysis by computing workers' eME, corresponding to the average difference between a crowd worker's judgment and the respective item's ground truth label. We found that workers overall tended to overestimate truthfulness (mean eME = 0.32, sd = 0.42, $t = 10.93$, $p < 0.001$; result from a one sample t-test; test value = 0). Looking at specific worker characteristics using linear regression and ANOVA models (incl. post-hoc tests), we found that workers who identified as *very conservative* and/or *Republican* tended to overestimate truthfulness more than other worker groups (i.e., Tukey-adjusted $p = [0.006, 0.050]$ compared to other political views for *very conservative* workers; Tukey-adjusted $p = [0.012, 0.089]$ compared to other party affiliations for Republican workers). The results further showed that workers who agreed to the southern border question (see the full survey on our repository) overestimated truthfulness more than workers who disagreed (Tukey-adjusted $p = 0.004$); although this effect seemed to be explained by workers' political affiliation, as 78% of those workers also identified as Republicans.

When looking for explanations for the aforementioned systematic biases, we found a slight trend that workers (especially those who identified as Republicans) particularly overestimated the truthfulness of those statements that confirmed their political views (see the left-hand panel of Figure 1). Ironically, due to the general trend toward overestimating truthfulness, this led the average worker to judge the truthfulness of statements affiliated with other parties more accurately than their own. This phenomenon could be explained by different cognitive biases [9], i.e., the *affect heuristic* (crowd workers may overestimate truthfulness when they like the statement speaker) or the *confirmation bias* (crowd workers may overestimate truthfulness when they support the underlying political message).

3.3.2 Exploring Worker's eMAE. We also considered eMAE, which corresponds to the mean absolute difference between a crowd worker's judgment and the respective item's ground truth label. The mean eMAE in the data is 0.42 (sd = 0.31), reiterating that the average worker was somewhat biased in their annotations (i.e., eAE ranged from 0 to 1.11). Moreover, in line with the findings above, we found that workers who identified as *very conservative* (Tukey-adjusted $p = [0.012, 0.200]$), Republican (Tukey-adjusted $p = [0.031, 0.129]$), or agreed on the southern border question (Tukey-adjusted $p < 0.001$) were more biased than others (i.e., had a higher eMAE; see the right-hand panel of Figure 1).

We also found that the more biased worker groups mentioned above generally took less time for their judgments compared to other workers. Although we did not find an effect of cognitive reasoning on eMAE when considering all independent variables at the same time, workers with lower cognitive reasoning also tended to do the task quicker. It could thus be that cognitive reasoning abilities explain some of the variance between worker groups but that the effect was too small to be detected in this exploratory study. Another explanation could be workers' *belief in science*: we found that 78% of the "disagree" answers regarding additional environmental regulations came from (very) conservative workers. Given the clear scientific stance regarding the environment, some workers may simply not trust scientific results and therefore distrust statements in which scientific results are brought up as evidence. Although there are too few of these "disagree" answers overall to detect a direct effect here, *belief in science* may be an underlying variable that influences the accuracy of crowd workers' truthfulness judgments.

Interestingly, the analyses also revealed a positive relationship between workers' average confidence in their judgments and eMAE ($\beta = 0.14$, $p < 0.001$), which might be an indication of *overconfidence*, a cognitive bias whereby workers with too much confidence in their abilities make more inaccurate judgments than others [9].

3.3.3 Exploring Worker's iME. Finally, we investigated iME, which corresponds to the mean difference between crowd workers' judgments and other crowd workers' judgments on the same statements. We found that workers with some postgraduate or professional schooling (no postgraduate degree) had higher confidence in their abilities to judge truthfulness compared to most workers with lower or higher education status (Tukey-adjusted $p = [< 0.001, 0.018]$). Our analyses also revealed that the more a worker identified as being conservative, the higher their self-reported confidence compared to other workers who annotated the same items. In general, confidence was higher in worker groups with greater bias, which further pointed to a potential *overconfidence* bias in some workers. This could also indicate that the confidence dimension acts as a proxy for explaining the political skewness of the results.

By far the strongest predictor of eME among the iME measures was the Correctness dimension ($\beta = 0.51, p < 0.001$). This suggests that workers might see the Correctness dimension as commensurable to Overall Truthfulness (as previously identified by Soprano et al. [43]), and indicates that workers who judge Correctness higher than others are likely also overestimating Overall Truthfulness.

Furthermore, we found that workers who identified as Democrats or Republicans judged truthfulness higher on most dimensions than workers who identified as independent or something else, which usually led to more accurate judgments for the latter group due to the general tendency toward overestimation of truthfulness. Even though these differences were small, this might be an indication that workers with higher *trust in politics* (as here represented by Republicans and Democrats) exhibit more overall bias because they overestimate truthfulness to a greater degree than workers with lower *trust in politics* (as here represented by other workers). This suspicion is underlined by the finding that workers who answered with "no opinion" to the southern border question tended to judge the speaker's trustworthiness lower than other workers (see the right-hand panel of Figure 1).

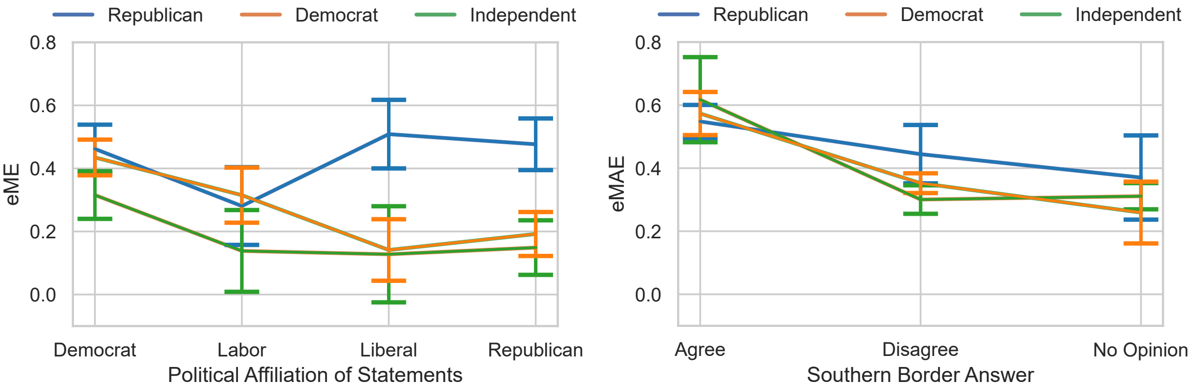

Figure 1: Mean eME per political affiliations of statements and workers (left) and mean eMAE per southern border answer and political affiliations of workers (right) in the public data set (see Section 3). Here, we excluded four workers who considered themselves something else other than Democrat, independent, or Republican.

Our analyses also revealed that iME for speaker's trustworthiness was the strongest predictor among the iME measures for eMAE ($\beta = 0.16, p = 0.040$). This again could point to a potential affect heuristic (see Section 3.3.1).

3.4 Hypotheses for the Novel Data Collection

From our exploratory study (see Section 3), we derived seven different hypotheses that we planned to test on novel data. We differentiated our hypotheses based on whether they refer to general worker traits (e.g., their *trust in politics*) or task-related cognitive biases (e.g., the *affect heuristic*).

3.4.1 General Worker Traits. These hypotheses refer to expectations about which worker groups may be more prone to biased judgments compared to others (RQ1).

- **Hypothesis 1a (H1a):** Workers with stronger *trust in politics* are less accurate in judging the Overall Truthfulness of statements compared to other workers.
  - Rationale: Workers who considered themselves Democrats or Republicans (i.e., the most "traditional" political parties) were less accurate in their truthfulness judgments than other workers in our exploratory study. Overly high *trust in politics* (i.e., the conviction that politicians and governmental bodies are trustworthy and aim to do the right thing) may lead some workers to strongly identify with political parties and could be the underlying reason for this bias. Such workers may not be skeptical enough when considering politicians' statements and therefore overestimate the likelihood of statements being true.
- **Hypothesis 1b (H1b):** Workers with stronger *belief in science* are more accurate in judging the Overall Truthfulness of statements compared to other workers.
  - Rationale: Workers who answered with "disagree" to the environmental regulations question (see the full questionnaire on our repository) tended to be more biased than others in our exploratory study. We hypothesize that the underlying responsible variable could be workers' *belief in science* (i.e., the conviction that scientific results are trustworthy and important for societal development). Workers with low belief in science may automatically doubt the truthfulness of statements that refer to scientific findings, e.g., related to climate change. This may undermine workers' ability to give accurate truthfulness judgments.
- **Hypothesis 1c (H1c):** Workers with better cognitive reasoning abilities are more accurate in judging the Overall Truthfulness of statements compared to other workers.
  - Rationale: In our exploratory study, we found that workers with lower cognitive reasoning abilities tended to perform the task quicker, which was generally associated with greater bias. Although we did not find a direct association of workers' cognitive reasoning abilities with their bias, we hypothesize that such a relationship could exist but that it might be hard to detect; especially given that many study participants have been exposed to the CRT before [19].

3.4.2 Cognitive Biases. These hypotheses are predictions about cognitive biases that may affect crowd workers (RQ2).

- **Hypothesis 2a (H2a):** Workers generally overestimate truthfulness.
  - Rationale: We found that workers overestimated truthfulness in our exploratory study, so we expect to find the same in novel data.
- **Hypothesis 2b (H2b):** Workers' tendency to over- or underestimate the Overall Truthfulness of a statement is related to the degree to which they like the statement claimant.
  - Rationale: Our exploratory study revealed several relationships that hint at a potential *affect heuristic*. As detailed by Draws et al. [9], this bias occurs when workers' judgments are affected by the degree to which they like the document they annotate.
- **Hypothesis 2c (H2c):** Workers' tendency to overestimate or underestimate the Overall Truthfulness of a statement is related to the degree to which they personally support the goal of the statement.

  - Rationale: Some relationships we found as part of our exploratory study hint at a potential *confirmation bias*, which occurs when workers' judgments are affected by their pre-existing opinions [9].
- **Hypothesis 2d (H2d):** Workers with higher confidence in their ability to correctly judge the truthfulness of items exhibit more bias compared to other workers.
  - Rationale: We found that workers' confidence in their judgments are directly related to their degree of bias in our exploratory study. We thus expect to find similar *overconfidence* [9] in novel data that we collect.

4 METHODS

To test the hypotheses detailed in Section 3.4, we conducted a further crowdsourcing study. Note that we preregistered our hypotheses, research design, and data analysis plan before data collection.[7]

4.1 Procedure

For the data collection, we relied on the same experimental design as Soprano et al. [43]. Specifically, we used the same interface and the same HITs, to keep the new task as similar as possible. We also relied on the same code and framework used in Soprano et al. [43], discussed in Soprano et al. [42].

To investigate our hypotheses (see Section 3.4), we identified three additional variables (i.e., *trust in politics*, *belief in science*, and *affect for statement claimant*; see Section 3.4) that required modifications to the original task. We used a generalized version of the *Citizen Trust in Government Organizations* (CTGO) questionnaire [18] to measure workers' trust in politics and the *Belief in Science Scale* (BISS) [6] to record workers' belief in science.[8] These two surveys were placed in the task right after the original initial questionnaire. Finally, we added a single, five-point Likert scale item to capture the degree to which the workers like the claimant of the statement. This item also included an additional answer option that allowed the worker to state that they do not know the claimant.

4.2 Variables

Our task recorded the following *Independent Variables*:

- *Trust in politics* (continuous; $[-2, 2]$): the degree to which workers trust in media and politics as measured by the CTGO questionnaire (i.e., averaging all responses). Higher scores mean greater trust in politics.
- *Belief in science* (continuous; $[-2, 2]$): the degree to which workers believe in science as measured by the BISS questionnaire (i.e., averaging all responses). Higher scores mean greater belief in science.
- *Cognitive reasoning* (ordinal; $[0, 4]$): worker's cognitive reasoning abilities as measured by the CRT; we also measure the time spent on CRT as a proxy for cognitive effort. Higher scores mean greater cognitive reflection.
- *Political party affiliation* (categorical): whether workers consider themselves as Republican, Democrat, independent or something else (i.e., not represented by any of the three previous political parties). We here relied on workers' responses to Q5 of the initial questionnaire (see our repository for the full questionnaire).
- *Affect for the statement claimant* (ordinal; $[-3, 3]$): each worker rated on a five-point Likert scale the degree to which they like each statement claimant; we also included the option "I don't know the claimant".
- *Mean confidence* (ordinal; $[-2, 2]$): workers' average self-reported confidence regarding the accuracy across their truthfulness judgments (on a five-point Likert scale).
- *Statement support* (categorical): we approximate the degree to which workers support the cause of the statement (whether true or false) with their personal political orientation.

We considered the `eE`, `eME`, and `eMAE` as *Dependent Variables* (see Sections 3.2.2 and 3.2.3). Finally, we considered `iE`, `iME`, age group, gender, level of education, income, political views, opinion on US southern border and about US environmental regulation of the workers as *descriptive and exploratory variables* (i.e., we do not conduct any conclusive hypothesis tests using those variables). We collected data on these variables using a survey (see our repository).

4.3 Crowd Workers

We planned to collect data from at least 255 crowd workers. We computed this required sample size in a power analysis for a Between-Subjects ANOVA (Fixed effects, special, main effects, and interactions; see Section "Analysis Plan") using the software *G*Power* [14]. Here, based on our findings in the exploratory study, we specified a small effect size of $f = 0.10$, a significance threshold $\alpha = 0.05/7 = 0.007$ (due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and that we have three between-subjects groups (i.e., Republican, Democrat, independent/else) and four within-subjects groups (i.e., Republican, Democrat, Liberal, Labor). We computed the required sample size for each of our hypotheses using their respective degrees of freedom.

We deployed 200 MTurk HITs to evaluate the set of 180 statements outlined in Section 3.1. We collected 2200 judgments in total. We recruited crowd workers who were based in the United States. Each crowd worker was rewarded \$2 for completing the task. This amount was based on the minimum time required to complete the task and the United States minimum wage of \$7.25 per hour.

4.4 Statistical Analyses

To test our hypotheses, we conducted several statistical analyses. We performed a multiple linear regression to predict `eMAE` from *trust in politics* (**H1a**), *belief in science* (**H1b**), and *cognitive reflection* (**H1c**), and *mean confidence* (**H2d**). We conducted a one-sample *t*-test to assess **H2a** (i.e., comparing `eME` to a test value of 0) and a Spearman correlation analysis to test **H2b** (i.e., computing a correlation between affect for the statement claimant and `eE`). Finally, we tested **H2c** by conducting a factorial mixed ANOVA with `eE` as dependent variable, workers' political party affiliation as between-subjects factor, and statement's political affiliation as within-subjects factor (i.e., **H2c** describes an interaction effect between these two variables).

[7] The preregistration is available at https://osf.io/5jyu4.

[8] All questionnaires can be found on our repository (see Section 1 for a link).

5 RESULTS

In this section, we describe the results of the crowdsourcing study outlined in Section 4.

5.1 Descriptive Statistics

Abandonment. We measured the abandonment rate of the crowdsourcing task using the definition provided by Han et al. [20] (i.e., how many workers voluntarily terminated the task before completion). Overall, 2742 workers participated. About 302 (11%) workers completed the task, while 2065 workers (75%) voluntarily abandoned it. Furthermore, 375 workers (14%) failed at least one quality check at the end of the task. Each worker had up to 10 tries to complete the task. We compared abandonment and failure distributions with those of Soprano et al. [43] (see Figure 2).

The left-hand panel of Figure 2 shows how many workers abandoned the task per number of statements annotated. The vast majority of workers (98%) abandoned the task when reaching the first statement. The number of workers who abandoned the task after the first statement is negligible. There is an 18% increase in abandonment rate when comparing our values with those of Soprano et al. [43], compared to which our task adds two additional questionnaires and an evaluation dimension. Thus, our task required somewhat more effort from workers. A higher number of workers may have become bored or frustrated sooner. Indeed, when considering the task described by Soprano et al. [43], it can be seen that a fraction of workers abandoned the task even after reaching the fourth statement. Despite this difference, the general trend was that workers abandoned the task when reaching the first statement.

The right-hand panel of Figure 2 shows how many workers failed at least one quality check after submitting their work within their current try. The majority of workers who failed the task performed it only once (216, 58%), with 103 (27%) workers doing it a second time. The remaining 15% of workers who failed the task performed it from three up to 10 times. The failure rate drops from 18% to 14% compared to the task by Soprano et al. [43], meaning that those who submitted their work were less likely to fail. However, the failure distribution of our task is in line with the one of Soprano et al. [43].

Demographics. We derived the following demographic statistics considering the 302 workers who completed the crowdsourcing task. Nearly 36% of workers were between 26 and 35 years old, while the 34% were between 35 and 50 years old. The majority of workers (52%) had a college/bachelor's degree. Concerning the total income before taxes, 25% of workers earned $50k to less than $75k, while 19% earned $75k to less than $100k. When considering workers' political views, 27% identified as moderate, 27% as conservative, and 26% as liberal. The majority of workers (53%) considered themselves Democrats, while the 27% as Republicans and the 17% as independent. The majority of workers (53%) agreed with building a wall at US southern border, with 25% of them disagreeing. Finally, the vast majority of workers (84%) thought that the government should increase environmental regulations to prevent climate change, while only 9% disagreed. In general, our sample was well balanced apart from a few categories and similar to the one of Soprano et al. [43], except that most workers in that study disagreed with building a wall at the US southern border.

Agreement. We measured the internal agreement among workers using Krippendorff's α [22] on the unit level. The use of this metric is motivated by earlier work [37, 43] and theoretical reasons [43]. We found a low level of agreement overall between the workers for each considered truthfulness dimension, which is in line with previous research [37, 43].

We also measured the external agreement between workers' aggregated scores for the Overall Truthfulness and corresponding experts' values. We recall that the judgment scales used by the experts and the workers are different. Whereas the experts used six- (*Politifact*) or three-level scales (*ABC*), the workers evaluated the statements using a five-level scale. Results for Overall Truthfulness and a sample of dimensions (Precision and Correctness) are summarized in Figure 3: on the x-axis of each plot, we show the ground truth, while on the y-axis, worker judgments are aggregated by taking the mean annotation for each document. Figure 3 thus shows a box plot for each ground truth label (i.e., from both *Politifact*; left of the dotted line; and *ABC*; right of the dotted line), where each dot represents one evaluated document.[9] It illustrates that workers tended to provide judgments with higher mean value when moving from left to right (i.e., when considering ground truth values with a higher value) in agreement with the experts. Although Figure 3 only shows the truthfulness dimensions Precision and Correctness as examples, we observed similar patterns for all dimensions. We recall that, although Overall Truthfulness is directly correlated with the ground truth, all the other dimensions capture orthogonal and independent information not directly measured by the experts. Moreover, the inter-quartile range is lower for *ABC* statements when compared to the analysis of Soprano et al. [43].

5.2 Hypothesis Tests

Our multiple linear regression analysis revealed no evidence for a relationship between eMAE and *trust in politics* (**H1a**; $\beta = -0.04, p = 0.020$) or *cognitive reflection* (**H1c**; $\beta = 0.02, p = 0.152$). However, *belief in science* (**H1b**; $\beta = 0.07, p = 0.003$) and *mean confidence* (**H2d**; $\beta = 0.06, p =< 0.001$) were both significant predictors of eMAE. Partly in contrast to what we expected, workers with stronger *belief in science* and those with greater mean confidence were *more* biased in their truthfulness judgments compared to others. We also found that workers generally overestimated truthfulness, as their mean eME (i.e., 0.33, sd = 0.46) lay significantly above 0 in the one-sample t-test we performed (**H2a**; $t = 12.18, p < 0.001$). Our Pearson correlation analysis revealed a significant positive relationship between affect for the statement claimant and eE (**H2b**; $r = 0.25, p < 0.001$). Thus, the more the workers liked the statement claimant, the more they overestimated truthfulness; and the more workers disliked the statement claimant, the more they underestimated truthfulness. Our final analysis was an ANOVA with the statement's affiliated party and worker's affiliated party as independent variables and eE as dependent variable. This analysis revealed no evidence in favor of an interaction effect between the two independent variables (**H2c**; $F = 1.59, p = 0.112$), which means that we can make no conclusion about whether workers had different degrees of over- or underestimating truthfulness depending

[9] Figure 3 can be directly compared with Soprano et al. [43, Section 5.3.2, Figure 3].

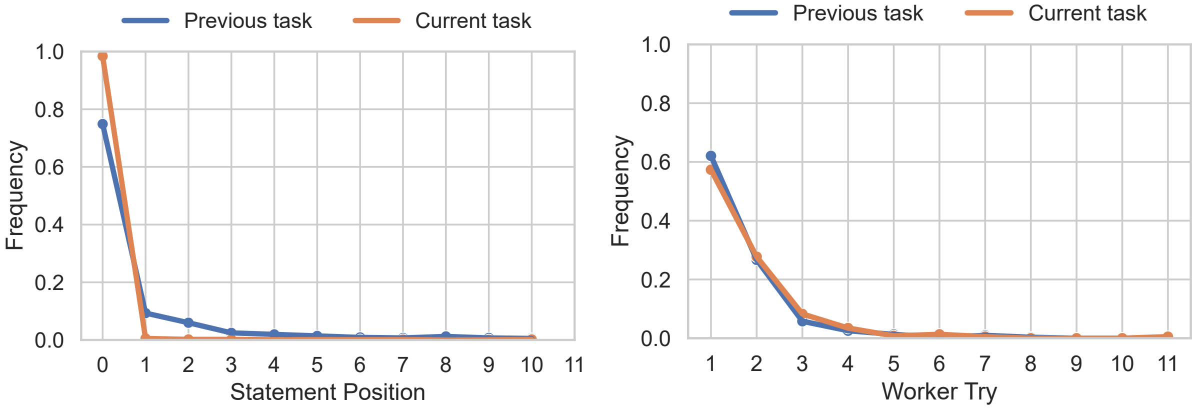

Figure 2: Comparison of workers' abandonment distribution (left) and workers' failure distribution (right). The orange line represents our task. The blue lines represent Soprano et al. [43] task.

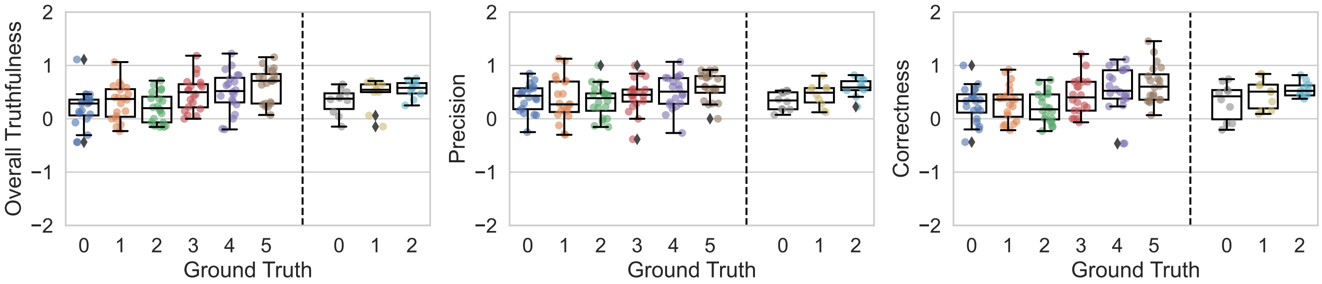

Figure 3: Summary of workers' judgments for the Overall Truthfulness, Precision, and Correctness dimensions, split by statements' ground truth labels. Whereas *Politifact* labels (left side of the dotted line) range from 0 (*pants-on-fire*) to 5 (*true*), ABC labels (right side of the dotted line) range only from 0 (negative) to 2 (positive).

on whether the statement party matched their personally favored party or political direction.

In sum, we found evidence in favor of some of our hypotheses (i.e., **H2a**, **H2b**, and **H2d**), suggesting that workers with greater confidence were more biased in their truthfulness judgments, workers generally overestimated truthfulness, and workers' truthfulness judgments were affected by the degree to which they liked the statement claimant. We also found evidence for a relationship between *belief in science* and bias in truthfulness judgments; however, in contrast to **H1b**, our results show that workers with a stronger belief in science were more biased than others.

5.3 Exploratory Analyses

Next to the descriptive analyses and hypothesis tests detailed above, we also performed several exploratory analyses on the data we collected. In doing so, we aimed to explain some of the outcomes from the hypothesis tests and identify interesting trends that had not been covered by those planned analyses. Note that the results we report in this subsection are indeed of exploratory nature, as we did not preregister these analyses.

5.3.1 Predicting eMAE. Our multiple linear regression identified workers' *belief in science* and *mean confidence* as significant predictors of eMAE. Interestingly, we found that, when conducting individual Pearson correlation analyses, only *mean confidence* correlates considerably with eMAE ($r = 0.20, p < 0.001$), whereas *belief in science* does not (see also Figure 4). This suggests that *belief in science* only becomes a relevant predictor of eMAE when also taking *trust in politics* and/or *cognitive reasoning* into account, as we did in our multiple linear regression analysis. These two variables might thus still play an important role in predicting workers' eMAE, although we did not find such evidence.

5.3.2 The Role of Workers' and Statements' Political Affiliations. The ANOVA we conducted shows no evidence for an interaction effect between workers' and statements' political affiliations in predicting eE (**H2c**). This suggests that workers may not overestimate or underestimate truthfulness systematically based on whether they support the political party that the statement is affiliated with. The same model also contains no evidence for the main effect of workers' political affiliation on eE ($F = 1.43, p = 0.232$), thus suggesting that workers' political affiliation may not matter at all here. However, there is a significant main effect for statements' political

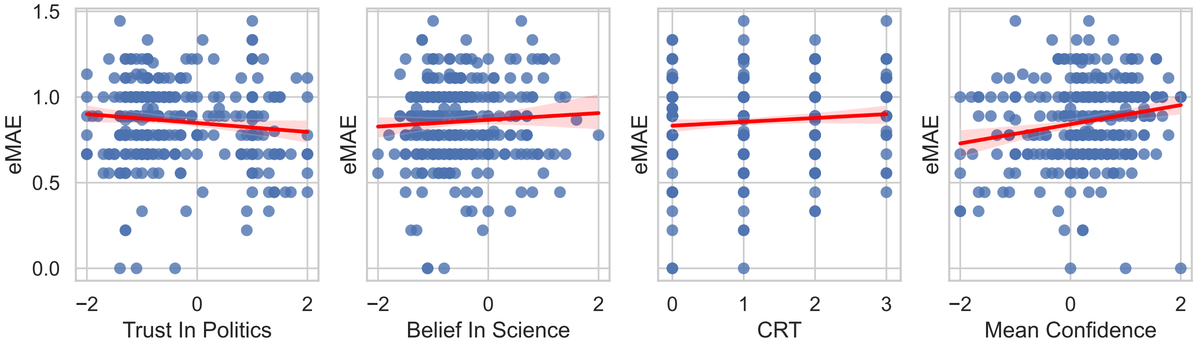

Figure 4: Scatter plots showing the relationships between workers' eMAE and their *trust in politics* (H1a, left-hand plot), *belief in science* (H1b, center-left plot), CRT (H1c, center-right plot), and mean confidence (H2d, right-hand plot). Our multiple linear regression analysis identified *belief in science* as well as mean confidence as significant predictors of eMAE (see Section 3.4).

affiliation ($F = 10.55, p < 0.001$). Comparing the different statement affiliations shows that workers overestimated the truthfulness of statements relevant to the Australian Labor party significantly more than those relevant to other parties (mean eE = 0.51, Tukey-adjusted $p = [< 0.001, 0.018]$). Workers also judged the truthfulness of statements affiliated with the Australian Liberal party significantly lower than those affiliated with other parties (mean eE = 0.08, Tukey-adjusted $p = [< 0.001, 0.014]$). Republican and Democrat statements were rated roughly equally on average. This suggests that the political parties connected to the statements may matter for predicting bias in crowd workers' truthfulness judgments, even –or perhaps especially– when those parties are not well-known among the crowd worker population (i.e., the crowd workers in our study were all US-based).

5.3.3 Looking at Individual Truthfulness Dimensions. **RQ3** concerns whether different truthfulness dimensions are affected by different biases. Next to an overall tendency towards overestimation of truthfulness, our hypothesis tests revealed that workers' *belief in science*, mean confidence, and the degree to which they like the statement claimant may be related to bias in their truthfulness judgments. We thus looked at which specific truthfulness dimensions were particularly affected by these biases to get some more insight into the nature of these biases.

We found that the best iME predictors of eMAE were Neutrality and Comprehensibility. Workers thus exhibited more bias when they judged Neutrality higher ($\beta = 0.10, p = 0.001$) or Comprehensibility lower than others ($\beta = -0.08, p = 0.013$). Moreover, we found that workers' *belief in science* affected no other truthfulness dimensions except Overall Truthfulness, while the mean confidence of a worker was a significant predictor for all iME measures. We also found other interesting relationships, i.e., between workers' *trust in politics* and lower scores on Neutrality ($\beta = -0.09, p = 0.028$), and between cognitive reasoning and higher scores on Comprehensibility ($\beta = 0.08, p = 0.027$). Finally, affect for the statement claimant was positively related to all considered truthfulness dimensions.

6 DISCUSSION AND CONCLUSION

In this section, we report a summary of the key findings derived in this work, list their practical implications, and sketch possible directions for future research.

6.1 Key Findings

We have presented a study on the impact of worker biases in crowdsourced fact-checking. To perform our analyses, we conducted an exploratory study using a publicly available data set from which we derived several hypotheses. We then tested these hypotheses in a novel crowdsourcing study. Below, we summarize our findings.

RQ1. Our first research question concerned what *individual characteristics* of crowd workers may lead to systematic biases in crowd workers' truthfulness judgments. In this context, we found no evidence for any influence of workers' *trust in politics* (**H1a**) or cognitive reasoning abilities (**H1c**). Our results do indicate a relationship between workers' degree of *belief in science* (**H1b**). However, in contrast to what we expected, we found that workers who reported a stronger belief in science were *less accurate* in their truthfulness judgments.

RQ2. The second research question that guided this paper concerned what *cognitive biases* can affect crowd workers' truthfulness judgments. Our results indicate that several cognitive biases can affect crowd workers' truthfulness judgments. Although we found no evidence for a *confirmation bias* [9] in this context (i.e., there was no interaction effect between workers' and statement's party affiliation on truthfulness judgments; **H2c**), we found that workers generally overestimate truthfulness (**H2a**). Our findings also suggest an influence of the *affect heuristic* [9]: the more workers like the claimant of a statement, the more they overestimate the statement's truthfulness (and vice versa; **H2b**). Finally, we found evidence for *overconfidence* in crowd workers: the higher workers' self-reported confidence in their ability to judge the truthfulness of statements, the less accurate their judgments generally were (**H2d**).

RQ3. Our final research question concerned whether different truthfulness dimensions are affected by different biases. Our study returned exploratory evidence that more biased workers judged the Neutrality of statements higher, and the Comprehensibility of statements lower than others. Moreover, workers' trust in politics was negatively correlated with their Neutrality judgments.

6.2 Practical Implications

Following the results of our study, we note several practical implications for crowdsourcing truthfulness judgments as well as adjacent domains such as the collection of document viewpoint annotations [8, 9, 27]:

- Although crowd workers generally seem to be reliable when judging the truthfulness of statements, individual characteristics (e.g., their belief in science) or cognitive biases (e.g., the affect heuristic or overconfidence) can negatively affect the accuracy of their judgments. We therefore recommend assessing, documenting, and –where possible– mitigating these biases [9, 11, 21]; either by adapting the task design or corrective post-processing of the collected data.
- Where applicable, we recommend that requesters measure relevant concepts such as workers' belief in science [6] to enable effective assessment of systematic biases. Requesters could also consider prioritizing workers with moderate political affiliation, *belief in science*, and confidence in their judgment abilities, as our study suggests that overly strong convictions in these contexts can lead to worse quality in truthfulness judgments.
- Related to the above point, we also recommend avoiding the employment of instruments that may *not* be strictly necessary, e.g., the cognitive reasoning test (CRT) for which we found no relationship to the quality of truthfulness judgments. Requesters should be aware that each such test may reduce the cognitive capacity of crowd workers to eventually perform the actual task. Thus, although we recommend assessment and mitigation of systematic biases, we note that requesters should also not overdo it in this respect.
- Where possible, we recommend that requesters hide unnecessary information (e.g., statement claimant identities or political affiliations) to mitigate the influence of cognitive biases such as the affect heuristic.
- Judgments coming from workers with high self-reported confidence in their ability to identify misinformation should be carefully adjusted, as we found that such workers tend to be more biased than others.

6.3 Future Work and Conclusion

This experiment lays the foundation for several future research directions. Since the data collected is consistent with previous studies dealing with fake news detection [39, 43], we plan to conduct additional analyses comparing those data sets, for example by leveraging worker behavior, interactions, URLs, or retrieved evidence. Another possibility is to perform a longitudinal study using the aforementioned data sets and the one collected in this work to compare them.

We also plan to conduct several experiments using the multidimensional truthfulness judgments collected. For example, since we found evidence that some dimensions are more important for the bias, we plan to test strategies to correct and de-bias the different truthfulness dimensions. By implementing such strategies, we aim to produce a set of non-biased data sets containing truthfulness labels collected using crowdsourcing; such resources can then be used as training data for state-of-the-art deep learning algorithms that target the automatic assessment of misinformation. Then, we plan to derive confidence scores from those algorithms and compare them with the self-reported workers' confidence scores. These de-biasing approaches could further be enhanced by analyzing the different ways disinformation targets subgroups [28]. Moreover, modern computational propaganda and social media platforms configurations are characterized by communication techniques that do not only misinform but also exhaust critical thinking, degrading the public's ability to share a system of interpretation of the social reality [46]. It is thus important to characterize the socio-technical features, platform metrics, and algorithmic configurations that affect the content production pipeline, to improve communities' resilience to the degradation of the public sphere.

We hope that the findings presented in the current paper, together with the future work detailed above, contribute toward a more sound, robust, and bias-free pipeline to effectively crowdsource reliable truthfulness assessments.

ACKNOWLEDGMENTS

Partly supported by the Credibility Coalition.

REFERENCES

[1] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science advances* 7, 36 (2021), eabf4393.

[2] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (2018), 54–61.

[3] Momen Bhuiyan, Amy Zhang, Connie Sehat, and Tanushree Mitra. 2020. Investigating "Who" in the Crowdsourcing of News Credibility. In *Computational Journalism Symposium.*

[4] Praveen Chandar and Ben Carterette. 2018. Estimating Clickthrough Bias in the Cascade Model. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. ACM, New York, NY, USA, 1587–1590.

[5] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and Debias in Recommender System: A Survey and Future Directions. *arXiv preprint arXiv:2010.03240* (2020).

[6] Neil Dagnall, Andrew Denovan, Kenneth Graham Drinkwater, and Andrew Parker. 2019. An Evaluation of the Belief in Science Scale. *Frontiers in Psychology* 10 (2019), 861.

[7] Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *Bulletin of IEEE Computer Society* 43, 3 (2020), 65–74.

[8] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 2022. Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics. In *ACM SIGIR Conference on Human Information Interaction and Retrieval* (Regensburg, Germany) *(CHIIR '22)*. Association for Computing Machinery, New York, NY, USA, 135–145. https://doi.org/10.1145/3498366.3505812

[9] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (Oct. 2021), 48–59. https://ojs.aaai.org/index.php/HCOMP/article/view/18939

[10] Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021. *This Is Not What We Ordered: Exploring Why Biased Search Result Rankings Affect User Attitudes on Debated Topics.* ACM, New York, NY, USA, 295–305.

[11] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM).* ACM, 162–170.

[12] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.

[13] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. ACM, New York, NY, USA, 1–11.

[14] Edgar Erdfelder, Franz Faul, and Axel Buchner. 1996. GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers* 28, 1 (01 Mar 1996), 1–11.

[15] Shane Frederick. 2005. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19, 4 (12 2005), 25–42.

[16] Amira Ghenai and Yelena Mejova. 2017. Catching Zika Fever: Application of Crowdsourcing and Machine Learning for Tracking Health Misinformation on Twitter. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 518–518.

[17] Anastasia Giachanou and Paolo Rosso. 2020. *The Battle Against Online Harmful Information: The Cases of Fake News and Hate Speech*. Association for Computing Machinery, New York, NY, USA, 3503–3504.

[18] Stephan Grimmelikhuijsen and Eva Knies. 2017. Validating a scale for citizen trust in government organizations. *International Review of Administrative Sciences* 83, 3 (2017), 583–601.

[19] Matthew Haigh. 2016. Has the Standard Cognitive Reflection Test Become a Victim of Its Own Success? *Advances in cognitive psychology* 12, 3 (30 Sep 2016), 145–149.

[20] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The Impact of Task Abandonment in Crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* (2019), 1–1.

[21] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of CHI*. 12 pages.

[22] Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. (2011).

[23] David La Barbera, Kevin Roitero, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In *Proceedings of ECIR*. Springer, New York, NY, USA, 207–214.

[24] Eun-Ju Lee. 2012. That's Not the Way It Is: How User-Generated Comments on the News Affect Perceived Media Bias. *Journal of Computer-Mediated Communication* 18, 1 (2012), 32–45.

[25] Kevin G. Love. 1981. Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology* 66, 4 (1981), 451.

[26] Eddy Maddalena, Davide Ceolin, and Stefano Mizzaro. 2018. Multidimensional News Quality: A Comparison of Crowdsourcing and Nichesourcing.. In *CIKM Workshops*. ACM, New York, NY, USA.

[27] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 31–41.

[28] Susan Morgan. 2018. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy* 3, 1 (2018), 39–43. https://doi.org/10.1080/23738871.2018.1462395 arXiv:https://doi.org/10.1080/23738871.2018.1462395

[29] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013), 647–651.

[30] Gordon Pennycook and David G Rand. 2018. Crowdsourcing judgments of news source quality. *SSRN. com* (2018).

[31] Gordon Pennycook and David G Rand. 2019. Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[32] Marcos Rodrigues Pinto, Yuri Oliveira de Lima, Carlos Eduardo Barbosa, and Jano Moreira de Souza. 2019. Towards fact-checking through crowdsourcing. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 494–499.

[33] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 209–216.

[34] Nicolas Pröllochs. 2021. Community-Based Fact-Checking on Twitter's Birdwatch Platform. *arXiv preprint arXiv:2104.07175* (2021).

[35] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. 2021. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. 189–199.

[36] Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2018. How Many Truth Levels? Six? One Hundred? Even More? Validating Truthfulness of Statements via Crowdsourcing. In *Proceedings of the CIKM 2018 Workshops co-located with 27th ACM International Conference on Information and Knowledge Management (CIKM 2018), Torino, Italy, October 22, 2018*. ACM, New York, NY, USA, 6 pages.

[37] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 439–448.

[38] Kevin Roitero, Michael Soprano, Beatrice Portelli, Massimiliano De Luise, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2021. Can the crowd judge truthfulness? A longitudinal study on recent misinformation about COVID-19. *Personal and Ubiquitous Computing* 1, 1 (2021), 1–31.

[39] Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020. The COVID-19 infodemic: Can the crowd judge recent misinformation objectively?. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1305–1314.

[40] Ricky J Sethi. 2017. Crowdsourcing the verification of fake news and alternative facts. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. 315–316.

[41] Shaban Shabani and Maria Sokhn. 2018. Hybrid machine-crowd approach for fake news detection. In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 299–306.

[42] Michael Soprano, Kevin Roitero, Francesco Bombassei De Bona, and Stefano Mizzaro. 2022. Crowd_Frame: A Simple and Complete Framework to Deploy Complex Crowdsourcing Tasks Off-the-Shelf. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 1605–1608. https://doi.org/10.1145/3488560.3502182

[43] Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2021. The many dimensions of truthfulness: Crowdsourcing misinformation assessments on a multidimensional scale. *Information Processing & Management* 58, 6 (2021), 102710.

[44] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2017. Detecting fake news in social networks via crowdsourcing. *arXiv preprint arXiv:1711.09025* (2017).

[45] Jacky Visser, John Lawrence, and Chris Reed. 2020. Reason-Checking Fake News. *Commun. ACM* 63, 11 (oct 2020), 38–40.

[46] Silvio Waisbord. 2018. Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism studies* 19, 13 (2018), 1866–1878.

[47] William Yang Wang. 2017. " Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL, Vancouver, Canada, 422–426.

[48] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of ACL*. 422–426.

[49] Ryen W White. 2014. Belief dynamics in Web search. *Journal of the Association for Information Science and Technology* 65, 11 (2014), 2165–2178.

[50] Pinar Yildirim, Esther Gal-Or, and Tansev Geylani. 2013. User-Generated Content and Bias in News Media. *Management Science* 59, 12 (2013), 2655–2666.

[51] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond Position Bias: Examining Result Attractiveness as a Source of Presentation Bias in Clickthrough Data. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, North Carolina, USA) *(WWW '10)*. Association for Computing Machinery, New York, NY, USA, 1011–1018.