Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources

Ziv Epstein
MIT Media Lab
zive@mit.edu

Gordon Pennycook
Hill/Levene Schools of Business
University of Regina
gordon.pennycook@uregina.ca

David Rand
Sloan School and Department of Brain and Cognitive Sciences, MIT
drand@mit.edu

ABSTRACT

How can social media platforms fight the spread of misinformation? One possibility is to use newsfeed algorithms to downrank content from sources that users rate as untrustworthy. But will laypeople unable to identify misinformation sites due to motivated reasoning or lack of expertise? And will they "game" this crowdsourcing mechanism to promote content that aligns with their partisan agendas? We conducted a survey experiment in which $N = 984$ Americans indicated their trust in numerous news sites. Half of the participants were told that their survey responses would inform social media ranking algorithms - creating a potential incentive to misrepresent their beliefs. Participants trusted mainstream sources much more than hyper-partisan or fake news sources, and their ratings were highly correlated with professional fact-checker judgments. Critically, informing participants that their responses would influence ranking algorithms did not diminish this high level of discernment, despite slightly increasing the political polarization of trust ratings.

Author Keywords

Misinformation, crowdsourcing, social media

CCS Concepts

•Information systems → Social networking sites; •Human-centered computing → Social media; Empirical studies in collaborative and social computing; Laboratory experiments; **•Applied computing** → *Law, social and behavioral sciences;*

INTRODUCTION

In recent years, social media has become the primary way that many people consume news [24]. Numerous features of the social media ecosystem, however, make it particularly vulnerable to the spread of "fake news" and other forms of misinformation [21, 45]. Given widespread concerns about the impact of such content, there have been significant efforts by social media platforms, as well as by academics across the computational and social sciences, to develop methods to reduce the proliferation of misinformation on social media.

One such method that has received considerable attention - including by Facebook [28, 43, 50] - is to use crowdsourcing to identify misinformation.ratings as inputs into the ranking algorithm. Here, we focus on one such system in which users judge the trustworthiness of domains that produce (mis)information (as opposed to evaluating individual pieces of content). The newsfeed algorithm would then use these trust ratings to weight content, such that content from domains that are distrusted by the crowd would be less likely to be displayed. In the current paper, we empirically investigate the feasibility of this approach by asking whether laypeople can, in fact, accurately identify misinformation sites.

There are three reasons to expect that layperson ratings may *not* successfully identify misinformation sites. First, and perhaps most notably, layperson trust judgments may be unduly swayed by partisan bias - such that people will preferentially trust news sources that produce content that they find ideologically reinforcing. That is, laypeople's trust judgments may be distorted by politically motivated reasoning [19]. If so, the actual veracity of the content produced by a given site may not be a meaningful predictor of the trust laypeople place in it. In other words, if trust judgments are dominated by partisanship rather than veracity, misinformation sites may not receive lower trust scores than non-misinformation sites.

Contrary to this view, however, there is a growing body of evidence that suggests that reasoning is not, in fact, held captive by ideology when evaluating the accuracy of news. Survey studies find that people who are more likely to engage in reasoning are less likely - not more likely - to believe and share false political headlines, regardless of ideological alignment [27, 34, 39]. Experiments show that - regardless of ideological alignment - engaging in reasoning *causes* decreased belief in false political headlines [3], whereas reliance on emotion causes increased belief in false headlines [23]. Furthermore, putting people into an accuracy mindset makes them less likely to share misinformation online [31]. Taken together, these results suggest that if laypeople are asked to think about the trustworthiness of news sources, their judgments may not be unduly swayed by partisanship.

Second, even if their judgments are not impaired by partisan bias, many laypeople may simply be unequipped to identify misinformation sites due to a lack of media literacy. For example, a 2018 Pew poll found that Americans had difficulty distinguishing factual news content from opinion [25]. Similarly, many laypeople may be unfamiliar with most news sources, especially since many people get their news from social media and not directly from the source [26]. Therefore, rather than (or in addition to) being heavily biased in a particular direction, layperson ratings might be characterized by a high level of noise or randomness - making them ineffective. In counterpoint to this concern, however, is the large literature on the "wisdom of crowds," which shows that aggregating responses can dramatically reduce noise [11, 13, 47]. Thus, even if the ratings of *individual* laypeople are noisy and ineffective, in the *aggregate* layperson judgments may be highly accurate.

Third, even if laypeople are *able* to produce effective ratings, they may *choose* not to, in an effort to "game," "astroturf" or otherwise manipulate the crowdsourcing system to achieve partisan ends [2, 22, 37, 50]. One approach to manipulation involves flooding the rating system with misleading responses. For example, one might have large numbers of accounts - potentially including bots - indicate that they trust a site that posts misinformation, leading to content from that site being promoted rather than demoted. Crowdsourcing approaches in which any user can indicate their opinion (e.g. up/down voting on Reddit) are vulnerable to these kinds of coordinated attacks. However, this danger is largely eliminated by using a rating system in which a subset of users are invited to provide their opinions (as in, for example, election polling). When the crowd is recruited in this manner, it is much more difficult for the mechanism to be infiltrated by a coordinated attack, as the attackers would have to be invited in large numbers to participate. Furthermore, rather than inviting random users, social media platforms could screen out users with suspicious activity profiles, further reducing the likelihood that malicious accounts substantially influence the crowd ratings.

Even if the platforms were able to screen out bad actors such as bots, trolls and spammers, there is another way in which the crowdsourcing mechanism could be manipulated. If people know that their opinion will be used to inform the newsfeed algorithm, they may try to game the system by giving higher trust ratings to websites that align with their political ideology - irrespective of how much they actually trust the information from those websites (e.g. their opinion about relative journalistic standards). That is, even if they do not in fact trust hyper-partisan sites, they may report trusting them in order to promote their partisan agenda (or to counteract what they imagine members of the other party may be doing to game the system). However, research from political science suggests that in fact most Americans do not care very much about politics [7] - such that they would not have a strong motivation to misrepresent their trust ratings for partisan ends. Furthermore, a large body of evidence suggests that most people are averse to lying for personal gain [12, 10], again suggesting that the incentive to game the system may not actually result in substantial changes in ratings.

These arguments and counterarguments underscore the fact that although this approach is fundamentally algorithmic, the challenges that must be overcome in order to implement it successfully are social in nature rather than technical, and thus involve empirical questions about how people would interact with such a system. Here, we shed light on these empirical social science questions in two ways. First, we assess the replicability of a recent study that suggested that layperson trust ratings do in fact effectively identify misinformation outlets [33]. Second, we investigate the extent to which participants change their responses when they are informed that the results will be used to inform social media ranking algorithms.

Consistent with prior work [33], we find that laypeople across the political spectrum distrust misinformation sites. Furthermore, we find no evidence that "gaming the system" to advance political agendas undermines the crowd's ability to identify such sites. Thus, our results suggest that using crowdsourcing to identify sources of misinformation is a promising approach for social media platforms.

RELATED WORK

One approach to the misinformation problem involves using computational methods to detect misinformation content. Many purely algorithmic detection methodologies have been proposed that leverage statistical markers of misinformation [5, 6, 14, 18, 38, 41, 42, 46, 49]. Some are text-based methods that rely on linguistic and stylistic regularities [9, 18, 36]. Others leverage existing knowledge ontologies to attempt to detect low-quality content [6, 14, 41, 46, 49]. While important progress is being made on this front, there are numerous practical challenges, including lack of a clear definition of what content should be included in training sets and what relevant features to include, as well as the non-stationarity of misinformation content (which tends to evolve rapidly). The crowdsourcing approach we study does not suffer from these challenges because a strict definition of "misinformation" is not required. Instead, sites are given graded (and thus more nuanced than just true/false) trust ratings based on humans' more contextualized (and constantly updating) understanding of the news ecosystem. Furthermore, non-stationarity is less of a problem because source-level trustworthiness is likely to change much less quickly than particular story-level signatures of misinformation.

A second approach to the misinformation problem is to have professional fact-checkers evaluate content as it appears and determine its veracity [1]. Content deemed to be false may then be downranked as well as labeled with a warning. This approach, however, is not scalable because bad actors can create false content at a much faster rate than fact-checkers can evaluate it, and the evaluation process itself is comparatively slow. Thus, most problematic content never winds up getting identified, and even the content that does eventually get flagged will likely be unflagged during its period of peak virality. In addition to limiting the effectiveness of the fact-checks, this scalability problem may actually promote the acceptance of misinformation via the "implied truth effect", whereby people interpret the absence of a warning as evidence that a story may have been fact-checked and validated [30]. The crowdsourcing approach we study here, conversely, is scalable because recruiting large numbers of laypeople is trivial on social media platforms. Furthermore, performing ratings at the level of the source, rather than the article, requires a much lower volume of ratings.

A third approach to the misinformation problem involves identifying and emphasizing the publishers of news content. For example, Facebook's "Article Context" feature provides information about the sources of articles linked in posts [16] and YouTube "notices" inform users when they are consuming content from government-funded organizations [40]. However, it is unclear to what extent these approaches actually improve truth discernment. For example, Jakesch et al. find only a weak effect of source label [17]; and Dias et al. find no impact of hiding versus emphasizing the source on most articles, because trusted sources typically publish stories that seem accurate even without source information, whereas distrusted sources typically publish stories that seem inaccurate even without source information [35].

A fourth approach to the misinformation problem involves the development of tools to help users detect misinformation themselves. For example, FeedReflect is a Chrome extension that nudges users to be more reflective and thus discerning in their news consumption [4]. UnbiasedCrowd is a automated assistant to help identify biases and prompt action in visual news media [29]. NewsR is a mobile app that allows users to annotate news articles to facilitate more critical interaction with news media [48]. A major limitation of such tools, however, is that they require people to opt in to using them. This is critical, because it seems likely that the people who are most susceptible to misinformation (e.g. who engage in less analytic thinking [32]) may be less likely to choose to use such tools. The crowdsourcing approach we study, conversely, does not have this problem because it is not opt-in: with the crowd ratings incorporated directly into the ranking algorithm, the ratings impact the content seen by everyone on the platform.

Finally, there are crowdsourcing approaches, one of which is the approach we study. Most prior work on crowdsourcing has focused on the evaluation of articles, for example by allowing users to flag content as misinformation. Kim et al propose CURB, a marked temporal points process framework that selects news to be fact-checked by solving a stochastic optimal control problem [20]. Tschiatschek et al propose DETECTIVE, an online algorithm that performs Bayesian inference to jointly learn user flagging activity and detect misinformation [44]. The approach we study differs from these approaches by focusing on evaluating news *sources*, rather than individual articles. This has the advantage of requiring a much lower volume of ratings (as there are many fewer sources than articles), allowing for greater scalability. Source-level ratings are also less susceptible to variation based on the idiosyncrasies of specific headlines.

The piece of prior work which is most relevant to the current paper is that of Pennycook & Rand [33], as we use the same trust/familiarity measures and list of news sources. We build on this prior work by adding the knowledge treatment, which allows us to test whether informing subjects that their ratings will be used to inform ranking algorithms (rather than just being part of an academic survey) increases partisan bias and reduces the performance of the crowd. This is a critical question, as any real application of crowdsourcing would entail such knowledge of the part of respondents. Furthermore, we assess the replicability of the previous findings. This is also critical, given the widespread "replication crisis" in the experimental social sciences, wherein many published findings turn out to be flukes rather than true results. If policy is going to be informed by this work, it is essential to know if it is replicable. Finally, we hope that the current paper will help to bring these findings to the attention of those working on platform design, who are best positioned to apply them in a useful way.

METHODS

We recruited $N = 1130$ Americans, of which $N = 984$ completed the survey, using Lucid, an online recruiting source that aggregates survey respondents from many respondent providers [8]. For a roughly 10-minute long survey such as ours, Lucid charges researchers $1 per participant. The participants are then compensated by the providers in a variety of ways, including cash and various points programs. Lucid mostly provides data to market research firms, and uses quota sampling to provide a sample which is nationally representative on age, gender, ethnicity and geographic region. Our sample had mean age = 45.47, 48.3% female, and 73% white. As a result of this representativeness, our sample also had good representativeness in terms of partisanship. For example, in a forced choice, 56% preferred the Democratic party and 44% preferred the Republican party.

Each participant was shown a list of website domains, and was asked: "Do you recognize the following websites?" (Yes/ No) and "How much do you trust each of these domains?" (Not at all/ barely/ somewhat/ a lot/ entirely). The domains were randomly sampled from a set of 89 news website domains across the right-left political spectrum that fall into the categories of mainstream media outlets (e.g. cnn.com, foxnews.com), websites with strong partisan biases that produce misleading coverage of events that did actually occur ("hyper-partisan" sites, e.g. breitbart.com, dailykos.com), and websites that generate mostly blatant false content ("fake news" sites, e.g. worldnewsdailyreport.com, dailybuzzlive.com, dailyheadlines.net). Our list of domains was taken from a previously published paper [33], which arrived at their list by combining several lists published by others of fake news sites, and of hyper-partisan sites. A website qualified as being fake news if it appeared on least two lists of fake news sites; and hyper-partisan if it appeared on at least two lists of hyper-partisan sites. The selection of which specific qualifying sites to include was biased towards sites with the greatest number of unique URLs on Twitter between January 1, 2018, and July 20, 2018.

Each participant in our experiment was shown 10 mainstream sources, 10 hyper-partisan sources, and 10 fake news sources (30 domains total). Thus, we can compare their trust ratings of mainstream sources to their trust ratings of hyper-partisan and fake news sources to construct a measure of how "discerning" their ratings are. This allows us to distinguish between two alternative hypotheses regarding the ability of laypeople to identify misinformation sites. The hypothesis that laypeople are unable to effectively identify misinformation sites (due either to motivated reasoning or lack of knowledge) predicts that average trust scores for mainstream sites will not be higher (and might even be lower) than average trust scores for fake news and hyper-partisan sites. Conversely, the hypothesis that laypeople will in fact be able to effectively identify misinformation sites predicts that average trust scores for mainstream sites will be substantially higher than average trust scores for fake news and hyper-partisan sites.

Furthermore, to provide some firmer ground-truth (rather than relying only on a classification of fake news / hyper-partisan versus mainstream), for a subset of 60 of the domains we use trust ratings collected from professional fact-checkers [33]. For these sites, we can assess the effectiveness of the layperson ratings by comparing them with ratings of the professional fact-checkers.

Critically, the study had a between-subject experimental design in which participants were randomly assigned to a control condition or a "knowledge" treatment. In the knowledge treatment, participants were informed at the outset of the study that their responses would be used to inform the ranking algorithms. Specifically, they were told

> The overall results of this study (but not any individual's responses) will be used to determine which news sources are relatively trustworthy and will be shared with Facebook with the goal of improving their platform. In particular, the goal is for content from sites which receive high trust ratings to be shown to more Facebook users than content from sites which receive low trust ratings.

Thus, by comparing ratings between the control and the knowledge treatment, we gain insight into how responses are affected by knowing that one's responses could influence the content that appears on social media. In particular, we can test the "gaming" hypothesis that participant trust ratings will be less discerning (i.e. there will be a smaller difference between trust ratings for mainstream versus fake news/hyper-partisan sites) in the knowledge condition.

Sample size and primary analyses for this study were pre-registered, and available at: http://aspredicted.org/blind.php?x=tf7y5r. Any analyses that were not pre-registered are labeled as post hoc.

RESULTS

We begin by comparing trust across mainstream, hyper-partisan, and fake news sites. The average trust ratings by condition for each source type among Democrats and Republicans are shown in Figure 1, and the distribution of trust scores by condition for each source type are shown in Figure 2. We see that there is an extremely similar pattern across both conditions: despite some partisan differences (e.g. foxnews.com was trusted much more by Republicans than Democrats), mainstream sites received much higher overall scores than either hyper-partisan or fake news sites.

This visual impression is confirmed by entering trust ratings into a regression (one observation per rating, standard errors clustered on participant) with the following independent variables: source type (hyper-partisan/fake news versus mainstream), condition (control versus knowledge treatment), and the interaction between the two. To make the regression coefficients for source type and condition directly interpretable in the presence of the interaction term, we zeroed the dummy variables. Source type was coded as mainstream = 2/3, hyper-partisan or fake news = -1/3, such that 0 corresponds to equal likelihood of non-misinformation vs misinformation source. Condition was coded as control = -0.5, knowledge treatment = 0.5, such that 0 corresponds to equal likelihood of either condition.

The results of this regression are shown in Table 1. We see a significant positive effect of source type (p<0.001), such that mainstream sources received higher trust ratings than non-mainstream sources; and no significant main effect of condition (p=.399) nor a significant interaction between source type and condition (p=.655), such that knowing that the ratings will inform ranking algorithms had no significant impact on average trust ratings.

Next, we test whether there was a differential effect of condition based on participant partisanship. To do so, we conduct the same analysis but also include a dummy for participant partisanship (-0.5=Prefers the Democratic party, 0.5=Prefers the Republican party) and all interactions (see Table 3). We find no significant 3-way interaction between participant partisanship, source type, and condition (p =0.399). This indicates that knowing that the ratings will impact the newsfeed does not affect how discerning people's trust ratings are (i.e. how effectively they differentiate mainstream versus fake news/hyper-partisan sources) for supporters of either party. Consistent with prior work, we do observe a significant 2-way interaction between participant partisanship and source type (p<0.001), such that Republicans trust mainstream sources less than Democrats. Additional analyses find the same pattern of a significant effect of source type and no interaction with condition when restricting to participants above versus below 45 years of age; men versus women; and participants with less than a college degree versus a college degree or higher. Furthermore, the significant effect of source in all regressions is robust ($p < .001$) when applying a Bonferroni correction for multiple comparisons.

That is not to say, however, that the knowledge treatment had no effects whatsoever. Although the treatment did not affect the crowd's ability to effectively discern between mainstream and hyper-partisan/fake sources, we did observe an increase in political polarization in the knowledge treatment. Specifically, in a post hoc analysis, we define the polarization in ratings for a given source as the absolute value of the difference in trust ratings between Democrats and Republicans (which presents visually as degree of dispersion from the 45 degree line in Figure 1). The distribution of polarization scores in the control versus treatment are shown in Figure 3. Visual inspection shows an increase in polarization in the treatment, as expected if (at least some) participants were strategically responding in the treatment. Consistent with this visual impression, a paired-sample t-test at the level of the source (i.e. two observations per source, control versus knowledge treatment) suggests that polarization was higher in the knowledge treatment than the control, t(88) =2.1364, p = 0.035. Thus, there is evidence that our treatment successfully induced participants to respond in a more partisan fashion. Critically, however, because these (small) polarization effects were essentially symmetric across party lines, they cancel out when computing overall discern-

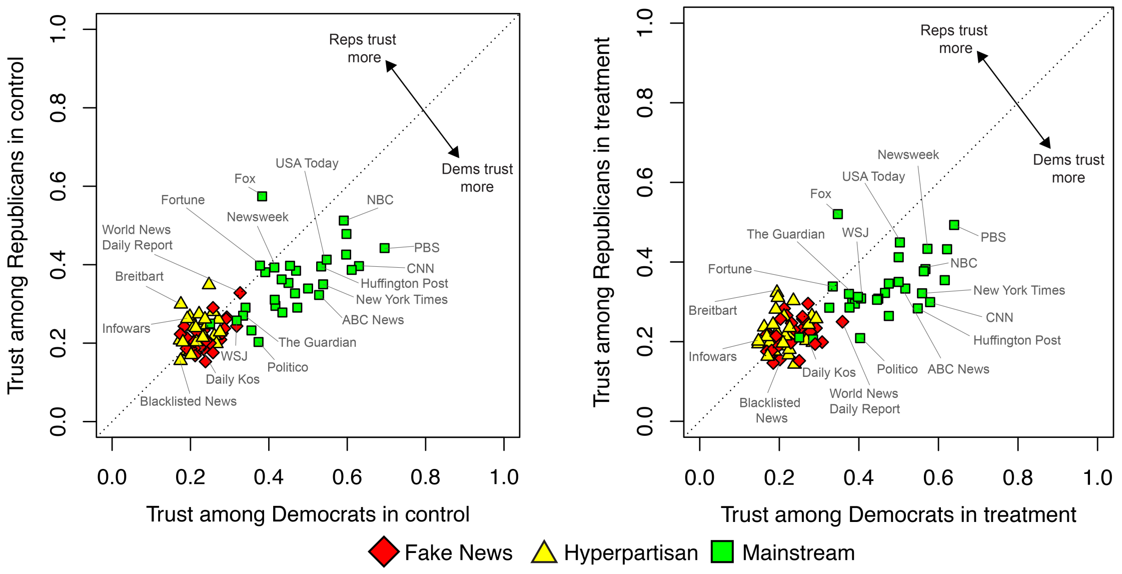

Figure 1. Trust among Democrats and Republicans for the 89 newsources in control (left) and treatment (right).

Table 1. Linear regression predicting trust ratings, with robust standard errors clustered on participant.

	Estimate	Standard Error	t value	p value
Condition (Knowledge Treatment)	-0.046	0.0554	-0.843	0.399
Source Type (Mainstream)	0.743	0.0271	27.406	<0.001
Condition × Source Type	-0.024	0.0542	-0.446	0.655
Intercept	2.147	0.0277	77.385	<0.001
$r^2 = 0.083$				

Table 2. Linear regression predicting trust ratings including participant partisanship as a covarite, with robust standard errors clustered on participant.

	Estimate	Standard Error	t value	p value
Condition (Knowledge Treatment)	-0.0297	0.0570	-0.5219	0.601
Source Type (Mainstream)	0.723	0.0258	28.032	<0.001
Partisanship (Republican)	-0.1693	0.0570	-2.968	0.003
Condition × Source Type	-0.0252	0.0516	-0.4888	0.625
Condition × Partisanship	-0.0100	0.1140	-0.088	0.929
Source Type × Partisanship	-0.4607	0.0516	-8.923	<0.001
Condition × Source Type × Partisanship	-0.0944	0.10327	-0.914	0.360
Intercept	2.135	0.0285	74.884	<0.001
$r^2 = 0.096$				

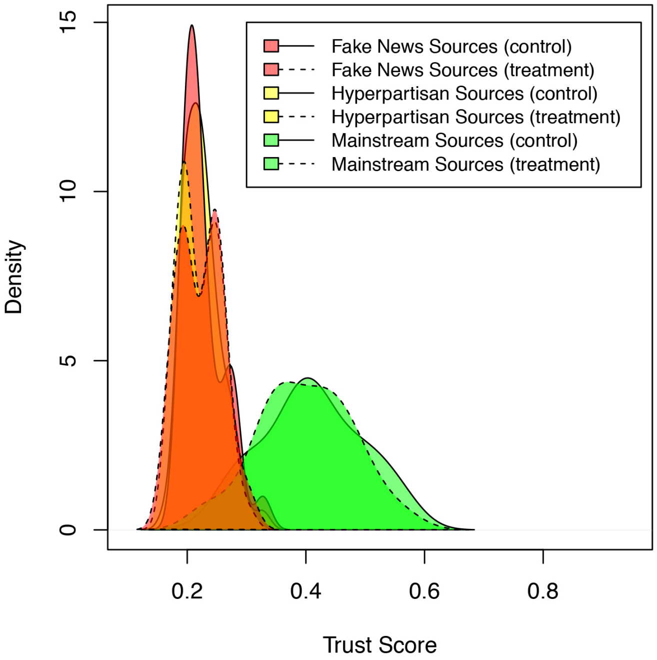

Figure 2. Kernel density plot showing trust scores by source type and experimental condition.

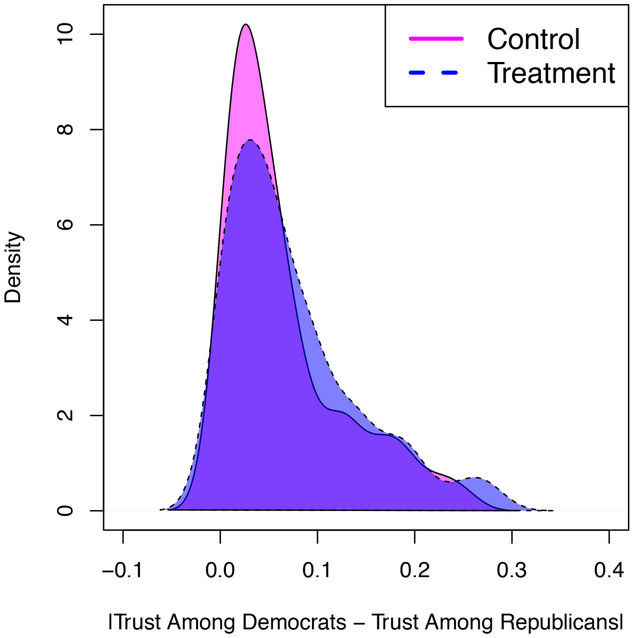

Figure 3. Kernel density plots for polarization, defined as the absolute value of the difference in trust ratings between Democrats and Republicans, by cn.

ment scores - and as a result, the crowd ratings still effectively identify misinformation sources (as shown in Figures 1 and 2).

Thus far, our analyses have implicitly considered all mainstream sources to be non-misinformation sites, and all hyperpartisan and fake news sites to be (at least in relative terms) misinformation sites. We now provide a more nuanced analysis by comparing our participants' ratings to the ratings of professional fact-checkers (examining the subset of 60 sites for which professional fact-checker trust ratings were available). For each condition, we calculate a politically-balanced layperson trust rating (weighting Democrats and Republicans equally) for each source. We then calculate the correlation between the politically balanced layperson ratings and the fact-checker ratings, which are very high in both conditions: r = 0.868 and r = 0.877 for control and treatment, respectively (see Figure 4a). Using a Fisher r-to-z transformation, we find that these two correlation coefficients are not significantly different from each other (z = -0.2, p = 0.84). Thus, the judgements of the laypeople in the treatment are just as highly aligned with those of the professional fact-checkers as the laypeople in the control - and our results above are not an artifact of our classification scheme of mainstream versus hyper-partisan or fake news. Given the lack of treatment effect, we collapse across conditions and calculate the politically-balanced layperson ratings for each of the 89 sources in our sample. The results are shown in Figure 4.

For completeness, we then repeat the same comparison with the fact-checkers considering Democrats and Republicans separately. Democrats had correlation coefficients with fact-checkers of r = 0.884 and r = 0.887 for control and treatment, respectively (see Figure 4b; no significant difference between conditions using a Fisher r-to-z transformation, z = -0.07, p = 0.94). Republicans had somewhat lower but still quite high correlation coefficients with fact-checkers of r = 0.726 and r = 0.686 for control and treatment, respectively (see Figure 4c; again no significant difference between conditions using a Fisher r-to-z transformation, z = -0.42, p = 0.67). A post hoc test indicates that the correlation with the fact-checkers was significantly lower for Republicans compared to Democrats in both conditions (control: z = 2.55, p = 0.0108; treatment: z = 3.04, p = 0.0024).

We now consider the role of familiarity in trust judgments. First, we conduct a post hoc analysis in which we re-run our main regression analysis from Table 1 with the addition of a z-scored dummy for familiarity and all interactions, shown in Table 3. Most importantly, we continue to observe the key findings from Table 1: there is a significant effect of source type (p<0.001), such that mainstream sources are trusted more than fake news or hyper-partisan sites, even when accounting for familiarity; and there continues to be no significant effect of condition (p=0.637). [We also note that removing source type from the model shown in Table 3 increases the AIC from 87983.5 to 88539.42, such that model selection supports inclusion of source type.] Turning to familiarity itself, we find a significant positive effect (p<0.001), such that familiar sources were trusted more than unfamiliar sources. We also found a significant positive interaction between familiarity and source type (p=0.0158), such that familiarity mattered more for mainstream sources than it did for fake news or hyper-partisan sites.

Finally, we consider the role of familiarity in more detail in Figure 6 by examining the distribution of trust scores across un-

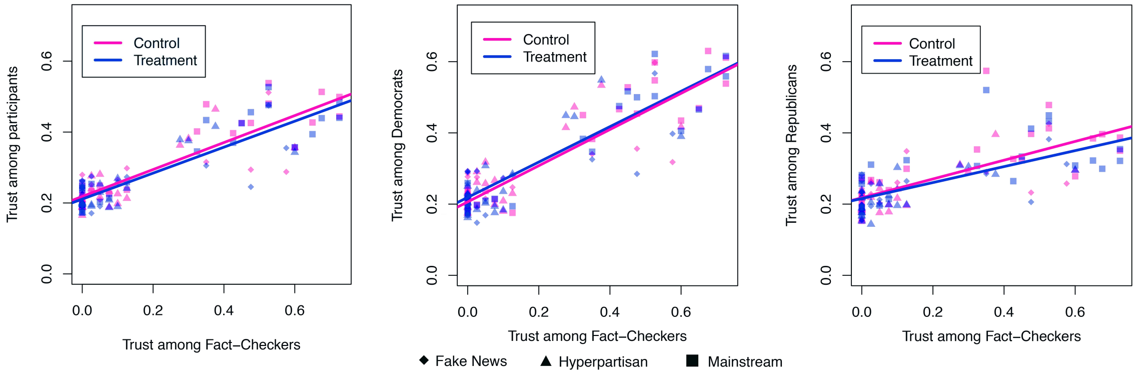

Figure 4. Trust among professional fact-checkers versus all participants (left), Democrats (middle) and Republicans (right). Included are the 60 news sources for which professional fact-checker ratings were available.

Table 3. Linear regression including source familiarity ratings, with robust standard errors clustered on participant.

	Estimate	Standard Error	t value	p value
Condition (Knowledge Treatment)	-0.065	0.058	-1.124	0.261
Source Type (Mainstream)	0.335	0.028	11.773	<0.001
Familiarity (Familiar)	0.424	0.026	16.278	<0.001
Condition × Source Type	0.033	0.057	0.577	0.563
Condition × Familiarity	-0.069	0.052	-1.325	0.186
Source Type × Familiarity	0.083	0.034	2.417	0.016
Condition × Source Type × Familiarity	0.020	0.068	0.293	0.769
Intercept	2.131	0.029	73.230	<0.001
$r^2 = 0.188$				

familiar versus familiar sources. As can be seen, participants in both conditions overwhelmingly distrusted news sources with which they were unfamiliar. Familiar sources, conversely, were not overwhelmingly trusted - instead, participants exhibited a wide range of trust levels for familiar sources. This asymmetry suggests that familiarity is necessary but not sufficient for trust.

DISCUSSION

The results we have presented here suggest that using crowdsourcing to identify outlets that produce misinformation, and then using those ratings as an input to social media ranking algorithms has promise for reducing the amount of misinformation on social media platforms. Specifically, we find that layperson trust ratings are quite effective in discerning between high and low quality news outlets. Rather than being blinded by partisanship, our participants tended to trust mainstream sources much more than hyper-partisan or fake news sources. Critically, in this work we find that layperson discernment is unaffected by informing participants that their responses will influence ranking algorithms: While this knowledge does indeed increase polarization of responses, these increases cancel out when calculating overall trust ratings. This observation helps to address concerns about individuals “gaming the system”, suggesting that strategic behavior by respondents aimed at affecting what content appears on social media may not pose such a serious problem for interventions that use crowdsourced ratings of trust in news sources to inform ranking algorithms.

An important issue with this approach, however, involves the role of familiarity in trust judgments. In our study, most participants were not familiar with most sources - there was an overall 30% familiarity rate. On the one hand, our results therefore show that a high level of familiarity with the relevant sources is not required for the crowdsourcing approach to be successful. On the other hand, however, familiarity does play an important role: Our results (as well as prior work [33]) suggest that familiarity is necessary but not sufficient for trust, such that unfamiliar outlets were overwhelmingly distrusted whereas trust ratings for familiar outlets were distributed across the full range of trust values. This observation (along with the regression results in Table 3) shows that trust ratings capture more than just familiarity. Yet this observation also suggests that sources that are reputable but not well-known are likely to receive low trust scores, and thus to be unfairly downranked (since people are not familiar with them).

This observation has important implications for platform design. How can this familiarity problem by addressed? It is not advisable to address it by only considering judgments of people who are familiar with a given source [33], as there are large selection effects: for example, people who tend to believe fake news are much more likely to visit - and therefore be familiar with - fake news sources. Instead, potential solutions include (i) showing raters sample content from each website before asking for their trust ratings, and (ii) having raters rate the accuracy of individual articles (without knowing the sources from which the articles come), and then creating site-level ratings by aggregating the accuracy scores of the articles from

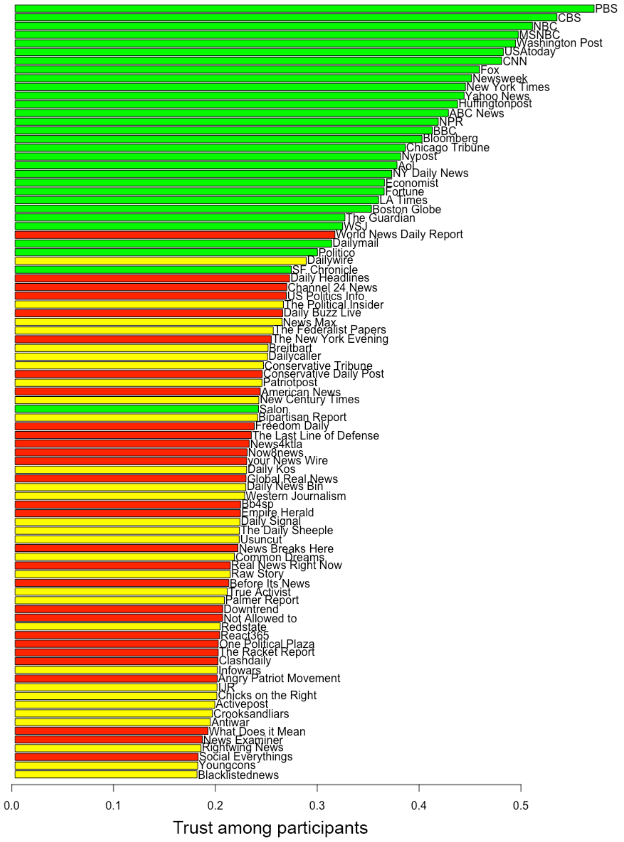

Figure 5. Average trust rating among of each of the 89 newsources, balanced by political partisanship (i.e. equally weighting Democrats and Republicans).

each site (this form of crowdsourcing would also have the added benefit of inducing an accuracy mindset in users, potentially leading them to share less misinformation themselves [31]). Investigating the effectiveness of these approaches to addressing the familiarity issue is an important direction for future research. In would also be fruitful for future work to investigate optimal criteria for which users to invite to provide ratings, and how to weight such ratings. For example, Hube et al find that crowd workers with strong opinions tend to produce more bias subjective evaluations [15].

In addition to these implications for fighting misinformation, our results are also of interest for more basic social science research. For example, we found that Republicans are less discerning than Democrats in their trust judgments (i.e. are worse at differentiating mainstream versus fake news/hyper-partisan sources). This was not because Republicans trust fake news or hyper-partisan sources more, but rather because Republicans trust mainstream sources *less*. This adds another piece of evidence to debates about ideological asymmetries in judgment. Future work should investigate *why* it is that conservatives are more likely to distrust reliable political information. Furthermore, our findings from the control condition provide a successful direct replication of prior findings [33], which is important given the surprising nature of the previous results and existing replication crisis in the social sciences.

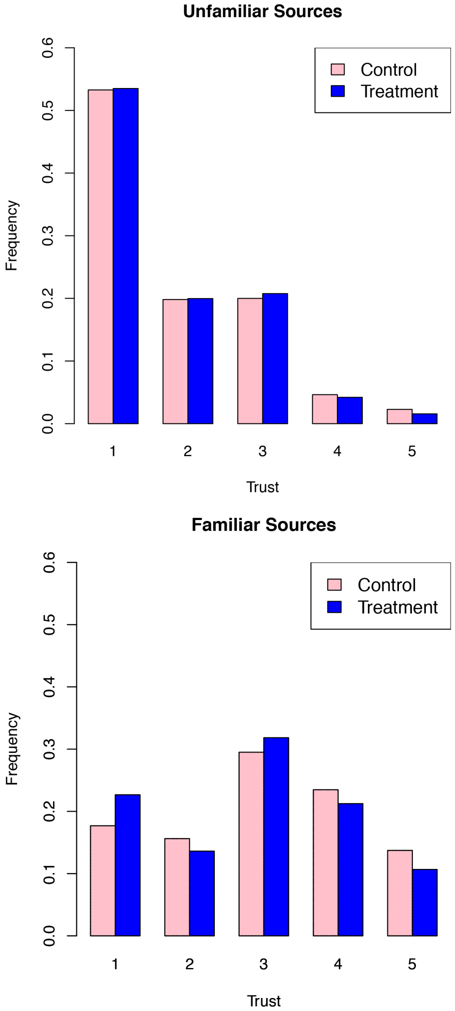

Figure 6. Histogram showing trust score for unfamiliar and unfamiliar sources, for control and treatment.

The are several limitations to the current study which are important to acknowledge. First, while the participants of the study were representative of the U.S. overall in age, gender, ethnicity, and geographic region, they may not match the users that a given platform recruits for eliciting trust scores. This is particularly important when considering applications of the crowdsourcing approach outside of the U.S. Assessing the cross-cultural generalizability of our findings is an extremely important direction for future research, and is necessary before platform designers implement such an approach elsewhere. Second, users might take the influence of their responses more seriously - and thus be more inclined to inflate the trust levels of ideologically consistent sources - if the platform was actually administering the survey, instead of our experimental survey. However, so long as that inflation is symmetric across parties and sources, it will cancel out when creating average trust scores. Also, we only consider 89 outlets, and it would be important to see how this generalizes to outlets more broadly. Finally, our framework utilizes a website level trust score, which does not take into account the variance in content quality that each website publishes. Thus, future work might explore how effective a site-level trust score is at predicting content-level quality. Such an approach will be unable, for example, to detect misinformation published by typically trusted sources (although such content may be quite rare [35]). Future work might also look at how a source can regain trust after its reputation has been damaged, how the crowd could score sources that contain aggregated news from various sources, or how to effectively design systems to increase the efficacy of crowd scoring.

Here we have provided experimental evidence that we hope will help to guide the development of platforms grappling with the challenge of misinformation. Our results suggest that the crowdsourcing approach described here is successful in identifying misinformation, and thus may be a useful addition to the social media platform designer's toolkit.

ACKNOWLEDGEMENTS

The authors would like to thank Erez Yoeli, Neil Gaikwad, Mor Naaman, Matt Groh, Adam Bear, Anna Chung, and Sandy Pentland for invaluable comments, as well as the Salt Lake City public library for crucial computational resources. The authors acknowledge funding from the Ethics and Governance of Artificial Intelligence Initiative of the Miami Foundation, the William and Flora Hewlett Foundation, the John Templeton Foundation, and the Social Sciences and Humanities Research Council of Canada.

REFERENCES

[1] 2018. Fact-Checking on Facebook: What Publishers Should Know. (Jun 2018). `https://www.facebook.com/help/publisher/182222309230722`

[2] Dennis Alann, Atino Kim, and Tricia Moravec. 2018. Facebook's Bad Idea: Crowdsourced Ratings Work For Toasters, But Not News. `https://www.buzzfeednews.com/article/alandennis/facebooks-bad-idea-crowsourced-ratings-work-for-toasters`. (Jan 2018).

[3] Bence Bago, David Rand, and Gordon Pennycook. 2019. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. (2019).

[4] Md Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael A Horning, and Tanushree Mitra. 2018. FeedReflect: A Tool for Nudging Users to Assess News Credibility on Twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 205–208.

[5] Xunru Che, Danaë Metaxa-Kakavouli, and Jeffrey T Hancock. 2018. Fake News in the News: An Analysis of Partisan Coverage of the Fake News Phenomenon. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 289–292.

[6] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015), e0128193.

[7] Philip E Converse. 2000. Assessing the capacity of mass electorates. *Annual review of political science* 3, 1 (2000), 331–353.

[8] Alexander Coppock and Oliver A McClellan. 2019. Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics* 6, 1 (2019), 2053168018822174.

[9] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 171–175.

[10] Simon Gächter and Jonathan F Schulz. 2016. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* 531, 7595 (2016), 496.

[11] Francis Galton. 1907. Vox populi (the wisdom of crowds). *Nature* 75, 7 (1907), 450–451.

[12] Uri Gneezy, Bettina Rockenbach, and Marta Serra-Garcia. 2013. Measuring lying aversion. *Journal of Economic Behavior & Organization* 93 (2013), 293–300.

[13] Benjamin Golub and Matthew O Jackson. 2010. Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics* 2, 1 (2010), 112–49.

[14] Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. In *Proceedings of the 2015 Computation+ Journalism Symposium*.

[15] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 407.

[16] Taylor Hughes, Jeff Smith, and Alex Leavitt. 2018. Helping People Better Assess the Stories They See in News Feed with the Context Button. (Jun 2018). `https://about.fb.com/news/2018/04/news-feed-fyi-more-context/`

[17] Maurice Jakesch, Moran Koren, Anna Evtushenko, and Mor Naaman. 2018. The Role of Source, Headline and Expressive Responding in Political News Evaluation. *Headline and Expressive Responding in Political News Evaluation (December 5, 2018)* (2018).

[18] Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 82.

[19] Dan M Kahan. 2017. Misconceptions, misinformation, and the logic of identity-protective cognition. (2017).

[20] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 324–332.

[21] David MJ Lazer and others. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[22] Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science* 62, 12 (2016), 3412–3427.

[23] Cameron Martel, Gordon Pennycook, and David Rand. 2019. Reliance on emotion promotes belief in fake news. (2019).

[24] Katerina Eva Matsa and Elisa Shearer. 2018. News Use Across Social Media Platforms 2018. (Sep 2018). `https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/`

[25] A Mitchell, J Gottfried, M Barthel, and N Sumida. 2018. Distinguishing between factual and opinion statements in the news. (2018).

[26] Amy Mitchell, Jeffrey Gottfried, and Katerina Eva Matsa. 2015. Millennials and political news. *Pew research center* 1 (2015).

[27] Mohsen Mosleh, Gordon Pennycook, Antonio Alonso Arechar, and David Rand. 2019. Digital fingerprints of cognitive reflection. (2019).

[28] Adam Mosseri. 2018. Helping Ensure News on Facebook Is From Trusted Sources. (2018). `https://about.fb.com/news/2018/01/trusted-sources/`

[29] Vishwajeet Narwal, Mohamed Hashim Salih, Jose Angel Lopez, Angel Ortega, John O'Donovan, Tobias Höllerer, and Saiph Savage. 2017. Automated assistants to identify and prompt action on visual news bias. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2796–2801.

[30] Gordon Pennycook, Adam Bear, Evan Collins, and David G. Rand. 2019a. The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384`, (2019).

[31] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2019b. Understanding and reducing the spread of misinformation online. (2019).

[32] Gordon Pennycook and David G Rand. 2018. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* (2018).

[33] Gordon Pennycook and David G Rand. 2019a. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* (2019), 201806781.

[34] Gordon Pennycook and David G Rand. 2019b. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.

[35] Gordon Pennycook, David G Rand, and Nic Dias. 2019c. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. (Dec 2019). `osf.io/m74v2`

[36] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638* (2017).

[37] Rob Price. 2019. Mark Zuckerberg has started his 2019 challenge of doing public debates — here are the highlights from the first one. `https://www.businessinsider.com/facebook-mark-zuckerberg-first-2019-public-discussion-2019-2`. (Feb 2019).

[38] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2931–2937.

[39] Robert M Ross, David Rand, and Gordon Pennycook. 2019. Beyond "fake news": The role of analytic thinking in the detection of inaccuracy and partisan bias in news headlines. (2019).

[40] Geoff Samek. 2018. Greater transparency for users around news broadcasters. (2018). `https://youtube.googleblog.com/2018/02/greater-transparency-for-users-around.html`

[41] Baoxu Shi and Tim Weninger. 2016. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 101–102.

[42] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[43] Henry Silverman. 2019. Helping Fact-Checkers Identify False Claims Faster. (Dec 2019). `https://about.fb.com/news/2019/12/helping-fact-checkers/`

[44] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 517–524.

[45] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[46] Xuezhi Wang, Cong Yu, Simon Baumgartner, and Flip Korn. 2018. Relevant document discovery for fact-checking articles. In *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee, 525–533.

[47] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. 2010. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*. 2424–2432.

[48] Gavin Wood, Kiel Long, Tom Feltwell, Scarlett Rowland, Phillip Brooker, Jamie Mahoney, John Vines, Julie Barnett, and Shaun Lawson. 2018. Rethinking Engagement with Online News through Social and Visual Co-Annotation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 576.

[49] You Wu, Pankaj K Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. *Proceedings of the VLDB Endowment* 7, 7 (2014), 589–600.

[50] Jonathan Zittrain and Mark Zuckerberg. 2019. Mark Zuckerberg discussion with Jonathan Zittrain. `https://www.youtube.com/watch?v=WGchhsKhG-A`, (2019).