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

Ziv Epstein, MIT Media Lab
Gordon Pennycook, University of Regina
David Rand, MIT Sloan School and Brain and Cognitive Science

1. INTRODUCTION

In recent years, social media has become the primary way that many people consume news [Matsa and Shearer 2018]. Numerous features of the social media ecosystem, however, make it particularly vulnerable to the spread of "fake news" and other forms of misinformation [Lazer et al. 2018; Vosoughi et al. 2018]. 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 possibility that has received considerable attention - including by Facebook [Mosseri 2018; Silverman 2019; Zittrain and Zuckerberg 2019] - 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. But will laypeople be 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? To investigate these questions, we empirically investigate the feasibility of this approach by asking whether laypeople can, in fact, accurately identify misinformation sites.

2. 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 [Coppock and McClellan 2019]. 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), web- sites 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. world- newsdailyreport.com, dailybuzzlive.com, dailyheadlines.net). Our list of domains was taken from a previously published paper [Pennycook and Rand 2019], which arrived at their list by combining several lists published by others of fake news sites, and of hyper-partisan sites. Each participant in our experiment was shown 10 mainstream sources, 10 hyper-partisan sources, and 10 fake news sources (sample from a set of 89 sources) and asked about their levels of trust for those sources. Critically, half of the participants were told that their survey responses would inform social media ranking algorithms - creating a potential

	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 $\times$ Source Type	-0.024	0.0542	-0.446	0.655
Intercept	2.147	0.0277	77.385	0.001
$r^2 = 0.083$				

Table I.

incentive to misrepresent their beliefs. 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.

3. RESULTS

We begin by comparing trust across mainstream, hyper-partisan, and fake news sites. 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 I. 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.

Participant ratings were highly correlated with professional fact-checker judgments in both conditions: r = 0.868 and r = 0.877 for control and treatment, respectively. Additional analyses find the same pattern of a significant effect of source type and no interaction with condition when restricting to Democrats or Republicans; 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 effect of source type (more trust of mainstream sources) remains significant (albeit with a smaller effect size) when controlling for familiarity with the news sources.

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, 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). Polarization is higher in the knowledge treatment than the control ($p = 0.035$).

4. 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.

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.

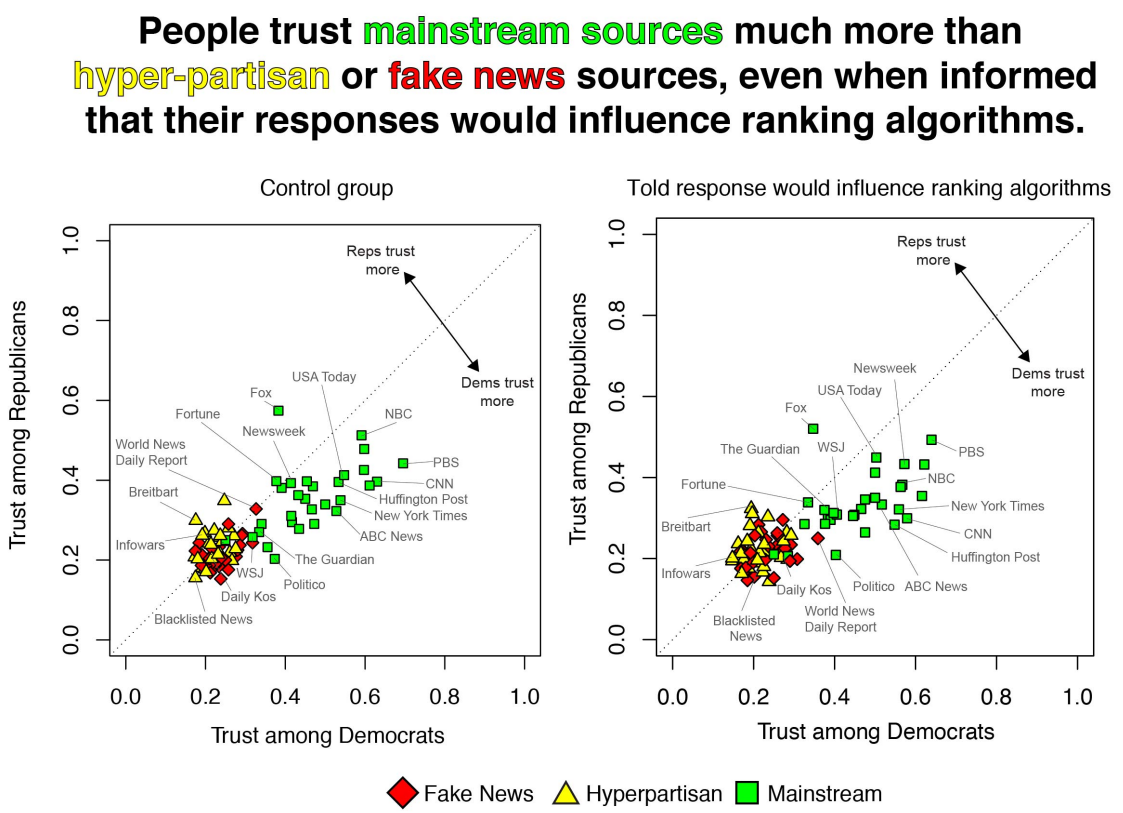

Fig. 1. Trust among Democrats and Republicans for the 89 newsources in control (A) and treatment (B).

REFERENCES

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.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, and others. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

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/

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

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* (2019), 201806781.

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

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

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