proceedings

MDPI

Proceeding Paper

The Argument against (Attempted) Fact Check †

Robin K. Hill

Department of Computer Science, University of Wyoming, Laramie, WY 82071, USA; hill@uwyo.edu

† Presented at Philosophy and Computing Conference, IS4SI Summit 2021, online, 12–19 September 2021.

Abstract: Attempts by social media to conduct vetting and verification of user-posted content, unless fully successful, will do more harm than good. Such attempts cannot be fully successful due to the vagaries of automated assessment of reliability and truth; no algorithm can be immune to the adversarial submission of material that passes all checks yet still contains misinformation. In light of the inevitable failures, public assurances that falsehoods are blocked will render fake news more trusted and more influential.

Keywords: fake news; filtering; philosophy of computing

1. Introduction

The promulgation via social media of the mendacious, selfish, vicious, and destructive facets of human interaction rightly stirs great concern. Calls to rectify these problems usually promote the removal of such offensive material. Objections to this include violations of free speech. This work explores a different objection—an ethical argument against (programmed) removal on the grounds that it will undercut itself to the further detriment of society. Although the chain of reasoning is straightforward, it does not seem to appear in public discourse.

2. Scenario

As illustration, for a fictional example, consider a timestamped stream of fictional headlines from a web publication, BuzzBook.

8:41 Stock market rises then falls
8:49 Delegations meet to discuss border controls
8:53 Greenland declares war on Canada
9:02 No hurricanes predicted for next two weeks
9:03 Female candidates show tattoos
9:08 New diet drops pounds faster than exercise
9:13 Ambulance destroys star's limo
9:14 Edmonton wins match
9:16 Fish County introduces ballotless voting in disenfranchised communities

Some are phony! BuzzBook realizes this, and tries to rectify the situation by flagging dubious items with a red asterisk *.

8:41 Stock market rises then falls
8:49 Delegations meet to discuss border controls
8:53 Greenland declares war on Canada *
9:02 No hurricanes predicted for next two weeks
9:03 Female candidates show tattoos
9:08 New diet drops pounds faster than exercise *
9:13 Ambulance destroys star's limo *
9:14 Edmonton wins match
9:16 Fish County introduces ballotless voting in disenfranchised communities

check for updates

Citation: Hill, R.K. The Argument against (Attempted) Fact Check. *Proceedings* **2022**, *81*, 63. https://doi.org/10.3390/proceedings2022081063

Academic Editor: Peter Boltuc

Published: 22 March 2022

The problem is the last item, which might sound plausible enough to pass whatever detection algorithm is active.

3. The Issue

Item 9:16 is the following:

- False, adapted from the satirical publication "The Onion" [1];
- Chosen because (in accordance with their standards) it sounds reasonable until the surprise hits;
- Is likely acceptable to obvious filtering and fake news detection.

Reader permitting, we can interpret the surprise as a crude approximation of the deficit of a programmed fact checker, which fails in this instance.

3.1. Implied Truth

Research by Pennycook and colleagues describes this as the Implied Truth Effect:

> . . . tagging some false news headlines with warnings will have the unintended side effect of causing untagged headlines to be viewed as more accurate [2]

The example given was humorous, fortunately. However, consider a maleficent piece that escapes the fact-checking process. We claim that the danger is greater in the case of attempted filtering, because that piece is more likely to be believed, and hence, acted upon. If we substitute this item at 9:16 in the following stream:

8:41 Stock market rises then falls

. . . . . .

8:53 Greenland declares war on Canada *

. . . . . .

9:14 Edmonton wins match

9:16 Clintons run child sex trafficking ring in basement of Comet Ping Pong

Then, the point is made more starkly. No one familiar with the Comet Ping Pong pizza shop invasion need be told that false information is dangerous in the actions inspired by it [3]. Note that the focus here is not on free speech, commerce, or the construal of trust as a property of relations directed at action [4], but is rather on plain informational credibility.

3.2. Information Pollution

The concern is information pollution, regardless of motivation, that spreads on the internet: "I information pollution' contaminates public discourse on a range of issues" [5] p. 10. We need not distinguish among disinformation, misinformation, and malinformation, or apply other taxonomies [5,6]. Recent American election cycles, and the novel coronavirus, have inspired such information pollution. The process is generally understood: Sensational stories propagate rapidly on the internet, and social networks feed, to the individual user, items calculated to compel that user's attention.

Because the effects directly follow from the upholding of free speech, platforms at first strenuously resisted what they viewed as self-censorship. The threat has become so palpable that many concerned citizens, including congresspeople, call on social media companies to fix it. Yet, social media platforms, such as Facebook and Twitter, do not practice journalism. They do not have sufficient staff, organization, or skills to review and verify each post; at the scale of social media activity, such staffing would be cost-prohibitive, or impossible.

4. Attempted Alleviation

4.1. Good Intentions

Social media platforms turn to program solutions, algorithms intended to perform the review and marking task. These algorithms are proprietary, but they presumably check new posts for inflammatory content expressed in the text, perhaps starting with single words, such as "Nazi" and "torture", and specifics related to breaking news, such as "Comet Ping Pong", and also perhaps combinations such as "kill" with "Obama" or "Trump" [7]. These

are the simplest of suggestions. Human content moderators are also employed, with mixed success [8].

Sophisticated artificial intelligence programs will be able to detect a great deal of viral information pollution, using much more acute criteria than those outlined above. Yet, a malicious agent who wants to spread a rumor that the novel coronavirus helps with weight loss can write a story with locutions, such as "the new condition from China gives you a slim and trim figure", escaping obvious word flagging. These criteria can be based not only on the text, but on any other data available—context, source, etc.

4.2. Incapacity

However, no program can successfully anticipate all such situations and thereby capture all the conditions of falsehood in advance. Any such algorithm can and will be gamed. It can be gamed because it is an algorithm, the implementation of a finite decision tree, where the conditions can be satisfied without the intention being satisfied [9]. No matter what finite list of algorithmic criteria an organization specifies, someone else can come up with a piece that (1) meets those criteria, yet (2) is recognized as suspect by an educated layperson. It will be gamed because such engagement is tempting to human nature, and because the sensational is a source of revenue on the web.

Filters can be improved, tuned, and refined, but no program can successfully anticipate all such situations and thereby capture all the conditions of falsehood in advance [10]. "Code will inevitably make mistakes, classifying real news as fake, and vice versa" [11].

Accounts such as [8] criticize Facebook for poor content moderation, but they also serve to show how difficult such moderation is for a company with no pretense to journalistic expertise. The traditional extension of the term "media" is the major newspapers and broadcasters of respectability, those who make a public commitment to verification. Certainly, errors occur, with mistakes rectified in print according to the standards of professionalism [12]. Patrons of print journalism understand this pledge, continually upheld by human monitoring.

5. Preventive Medicine

A reader's vulnerability to insidious falsehood is that person's likelihood of buying into fake news (whether the source is communicating through malice or mistake). Vulnerability increases under fact checking, as shown in Figure 1.

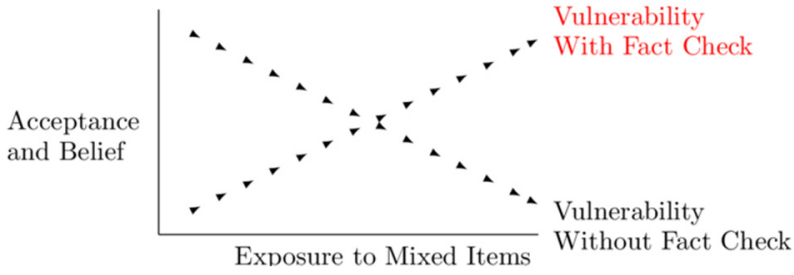

Figure 1. As blind belief falls on exposure, vulnerability decreases; vice versa with reliance on fact checking.

We can invoke a preventive care medical metaphor, where the goal is immunization (collective), and the means is inoculation (individual). In fact, the online game Bad News offers not only a "vaccine", but a test, showing success [13].

As shown in Table 1, the 9:16 claim shown, on BuzzBook unflagged, is *less* dangerous, as shown in the "Assessed" column, because its readers will have realized, via earlier processing, that its credibility is weak. That is the "inoculation".

Table 1. Headlines as understood by reader exercising judgment.

Time	BuzzBook Unflagged	Assessed by Alert Reader
8:41	Stock market rises then falls	Stock market rises then falls
8.49	Delegations meet to discuss border controls	Delegations meet to discuss border controls
8:53	Greenland declares war on Canada	Oh, pshaw *.
9:02	No hurricanes predicted for next two weeks	No hurricanes predicted for next two weeks
9:03	Female candidates show tattoos	Female candidates show tattoos
9:08	New diet drops pounds faster than exercise	As if . . . *
9:13	Ambulance destroys star's limo	Sensationalism *
9:14	Edmonton wins match	Edmonton wins match
9:16	Fish County introduces ballotless voting in disenfranchised communities	Wait a minute *

* Reader's thoughts.

As social media serves up more and more content of less and less credibility, subscribers will learn the proper degree of skepticism: " . . . individuals learn on the basis of their experience to correctly assess the trustworthiness of the others" [4]. If they know not to rely on the platform to filter out misinformation, users will realize quite rapidly that certain outlets have no credibility. It is everyday practical judgment applied to traditional outlets that enables us to dismiss sleazy print publications. Strong assessment skills will direct readers toward serious news sources other than social media, quite appropriately.

6. Exercising our Discrimination

Under this assurance, falsehood that escapes the filters and makes it to publication on the platform carries the warrant of veracity. If an outlet claims to verify content, and fails, then the reader suffers the same degree of harm whether that failure is due to neglect or incapacity. Furthermore, the reader's vulnerability to insidious falsehood is inversely proportional to its frequency. A platform that blocks the bulk of misinformation inflicts an increased likelihood that rare misinformation will be trusted. A hack of Snopes or Politifact that disseminated misinformation would do more harm than the same misfortune from other sites. Other professions face similar problems: auditors of high reputation convey that quality to their clients—an arrangement that can be corrupted [14].

The ethics of public engagement require recognition of this caution. It is not the accidental appearance of falsehood, nor the moderation of content, that forms an ethical violation, nor is it the attempt to moderate content and verify claims, but it is the failure to state the limitations in public—failure to identify and place the danger—that constitutes an ethical violation. Social media should explicitly disavow any claim to veracity in the posts that they carry. (They should also refer users to sources of serious journalism.) We see some movement toward this tempered stance in Facebook's cautious labeling of dubious posts, but we believe that such assessment is better carried out by the alert reader's own critical thinking skills, honed by exposure.

In short:

1. Misinformation inflicts danger when people believe and act on it.
2. Social media contributors circulate voluminous misinformation.
3. At scale, social media must use programs to verify content.
4. Programs cannot adequately detect falsehood.
5. Ergo, social media verification is inadequate in blocking all misinformation.
6. Claims by outlets to verify content fosters trust in the content.
7. Ergo, fact-checking increases the damage done by those falsehoods that escape verification.

Therefore, in social media, public commitments to verify online content are more 142 dangerous than forgoing such a public commitment.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. The Onion. Florida GOP Introduces Ballotless Voting in Disenfranchised Communities. *The Onion*, 25 February 2021.
2. Pennycook, G.; Bear, A.; Collins, E.T.; Rand, D.G. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Manag. Sci.* **2020**, *66*, 4944–4957. [CrossRef]
3. Wikipedia Contributors. Pizzagate Conspiracy Theory—Wikipedia, The Free Encyclopedia. 2020. Available online: https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory (accessed on 15 November 2020).
4. Turilli, M.; Vaccaro, A.; Taddeo, M. The Case of Online Trust. *Knowl. Technol. Policy* **2010**, *23*, 333–345. [CrossRef]
5. Wardle, C.; Derakhshan, H. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Available online: https://rm.coe.int/information-disorder-report-november-2017/1680764666 (accessed on 23 October 2021).
6. Molina, M.D.; Sundar, S.S.; Le, T.; Lee, D. "Fake news" is not simply false information: A concept explication and taxonomy of online content. *Am. Behav. Sci.* **2021**, *65*, 180–212. [CrossRef]
7. Wiesemann, S. Software That Can Automatically Detect Fake News. Available online: https://www.fraunhofer.de/en/press/research-news/2019/february/software-that-can-automatically-detect-fake-news.html (accessed on 23 October 2021).
8. Marantz, A. Explicit Content: Why Facebook Can't Fix Itself. *The New Yorker*, 12 October 2020.
9. Hill, R.K. Gaming the System: Definition. Available online: https://cacm.acm.org/blogs/blog-cacm/254472-gaming-the-system-definition/fulltext# (accessed on 23 October 2021).
10. Mrcus, G.; Davis, E.A.I. Won't Fix Fake News. *The New York Times*, 20 October 2018.
11. Verstraete, M.; Bambauer, D.E.; Bambauer, J.R.Y. Identifying and Countering Fake News. *Hastings Law J.* **2021**, *73*. [CrossRef]
12. Society of Professional Journalists. SPJ Code of Ethics. Available online: http://www.spj.org/ethicscode.asp (accessed on 23 October 2021).
13. Basol, M.; Roozenbeek, J.; van der Linden, S. Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. *J. Cogn.* **2020**, *3*, 2. [CrossRef] [PubMed]
14. Shapiro, A. Who Pays the Auditor Calls the Tune?: Auditing Regulation and Clients' Incentives. *Seton Hall Law Rev.* **2005**, *35*, 1029.