DIGITAL PEACEBUILDING

2025 Issue 2

Countering Digital Misinformation with Automatic Fact-Checkers

Sophia Campbell

Executive Summary

Digital misinformation increases political violence by inspiring retribution for fabricated wrongdoings and intensifying societal ideological polarization. Automatic fact-checkers (AFCs), computer programs with the capacity to detect false claims, can mitigate the spread of misinformation and therefore reduce the correlated acts of violence. AFCs are cost-effective and generally correct in their judgements. However, they can be manipulated by cyber attacks and may contain inherent algorithmic biases. Developers and regulatory powers must take steps to ensure accurate, impartial, and effective judgements of veracity.

Introduction

Digital misinformation can inspire acts of political violence by increasing ideological or political polarization within a society and propogating targeted mistruths to provoke retribution. Automatic Fact Checkers (AFCs) have the capacity to verify pieces of digital media in near-live time, potentially reducing users' belief in false claims and limiting their spread altogether. While this technology could decrease instances of political violence fueled or instigated by misinformation, it also raises new concerns of inaccuracy and bias in claim verification. In this paper, we explore the capacities, benefits, and risks of AFCs as applied to digital political misinformation.

Digital Misinformation

Most generally, digital misinformation is information on the internet that is not correct.[1,2] There are a few key stipulations. First, "misinformation" refers to the incorrect piece of information itself, not a user's misunderstanding of correct information. Second, it need only be incorrect based on the best available contemporary evidence.[3] If there is general consensus among subject experts that a claim is true, but a lack of absolute evidence supporting it, that claim is not considered misinformation. Third, misinformation is not necessarily created or spread intentionally.[4] Users may share false claims in good faith. (This differentiates it from "disinformation," which is spread knowingly.)

It should be noted that while scholarly works may differ in their interpretation of misinformation, this paper will use the definition described above. Also– while there are plenty of genres of misinformation on

1. M. Rulis, "The Influences of Misinformation on Incidences of Politically Motivated Violence in Europe," The International Journal of Press/Politics 0 (2024), https://doi.org/10.1177/19401612241257873.

2. C. H. Au, K. K. W. Ho, and D. K. Chiu, "The Role of Online Misinformation and Fake News in Ideological Polarization: Barriers, Catalysts, and Implications," Information Systems Frontiers 24 (2022): 1332, https://link.springer.com/article/10.1007/s10796-021-10133-9/.

3. Emily K. Vraga and Leticia Bode, "Defining Misinformation and Understanding Its Bounded Nature: Using Expertise and Evidence for Describing Misinformation," Political Communication 37, no. 1 (2020): 139, https://doi.org/10.1080/10584609.2020.1716494.

4. Pramukh Nanjundaswamy Vasist et al., "The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative," Information Systems Frontiers 26 (2024): 664, https://doi.org/10.1007/s10796-023-10390-w.

Sophia Campbell is an MPA candidate at Brown University's Watson Institute and holds a BA in economics from Barnard College. She can be reached at sophia.ellen.campbell@gmail.com.

Editor: Dawn Brancati

the internet–we will limit our discussion to false political, ideological, or large-scale cultural statements. Finally, we will use the term "user" to refer to anyone consuming digital information, whether they believe it or not, and "claim" to refer to unverified pieces of digital media, including images and videos.

Misinformation as a Cause of Violence

Misinformation fuels political violence through direct inspiration and ideological polarization. Most overtly, misinformation may include malicious claims about a specific person or persons. Users who believe the false claims might take violent actions with a sense of vigilante justice, believing that they are acting in retaliation for or prevention of a perceived wrongdoing.[5] For instance, the two-day anti-Muslim riots in Myanmar in August of 2018 were instigated by an unsubstantiated rumor, spread on Facebook, that a Muslim man had sexually assaulted a Buddhist woman.[6]

Misinformation can also inspire acts of violence indirectly by exacerbating ideological polarization. Misinformed claims that are partisan in nature tend to contain hyperbolic, provocative, or malicious statements which, when believed, can push users to more extreme versions of their existing partisan beliefs. The resulting polarization broadens ideological gaps between civilians and can contribute to increasing social frictions, which can result in acts of violence including political instability, protests, and even domestic terrorism.[7]

This intuitive relationship is supported by statistical evidence. Vasist, Chatterjee, and Krishnan used an expansive, cross-country database of manually collected data to examine correlations between instances of hate speech, misinformation, censorship, and indicators of societal polarization; they determined that online falsehoods carry a particular "central role" in polarizing societies.[8]

Through these mechanisms, misinformation increases political violence in a significant, observable pattern. By conducting statistical analysis on a fused dataset of confirmed pieces of misinformation and instances of political conflicts across Europe, Rulis found that the presence of translational digital misinformation on social media had a significant positive correlation with occurrences of (1) verbal and material conflict between citizens and government entities, and (2) material–but not necessarily verbal–conflict between civilians.[9]

TapTheForwardAssist, 2021.DC Capitol Storming. https://commons.wikimedia.org/wiki/File:DC_Capitol_Storming_IMG_7965.jpg via Wikimedia Commons.

Anecdotal evidence indicates that this relationship is not limited to Europe. For instance, the January 6th insurrection attempt in the United States, in which perpetrators enacted material violence on government property and expressed a desire to assault government personnel as a result of misinformation spread through social media, is a clear example of material and verbal violence from civilians to the government. Civilian-civilian material violence is evident in cases such as the mob lynchings of suspected child abductors–accused by unsubstantiated rumor on Whatsapp–in India in 2018.[10]

5. Sumitra Badrinathan, Simon Chauchard, and Niloufer Siddiqui, "Misinformation and Support for Vigilantism: An Experiment in India and Pakistan," American Political Science Review 118, no. 1 (2024): 1, https://doi.org/10.1017/S0003055424000790.

6. Claire Wardle and Hossein Derakhshan, Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making, Council of Europe Report, 2017, 41, https://firstdraftnews.org/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-de%CC%81sinformation-1.pdf.

7. Pramukh Nanjundaswamy Vasist et al., "The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative," Information Systems Frontiers 26 (2024): 666, https://doi.org/10.1007/s10796-023-10390-w.

8 Ibid. 677.

9. Rulis, "The Influences of Misinformation...."

10. Rohit Chopra, "Misinformation and Violence," Santa Clara University Markkula Center for Applied Ethics, November18, 2021, https://

Misinformed and Patriotic: Public Support for the War on Terror

When the United States invaded Iraq, about 80 percent of the Americans who supported the invasion ranked "Iraq's connection with groups like Al-Qaeda" as the main reason for their approval. Only a few months prior, more than half the country was reported to believe that Iraq was colluding with terrorist organizations, or had weapons of mass destruction, or that the rest of the world would support an American invasion. This was not true.[I]

A research team at the University of Maryland's Program on International Policy Attitudes ran a multivariate analysis to determine the correlation between belief in the three misperceptions above and likelihood of supporting the invasion. The results: Those who believed at least othese myths were 4.3 times more likely to be supportive than those who had no misperceptions.[II] A study on the same subject by the American Psychological Association came to an equally blunt conclusion: "Dissemination and control of information are indispensable ingredients of violent conflict." [III]

This is interdisciplinary agreement that public support for the war was a function of widespread misperceptions. In this case, as in the case of conspiracy theories, belief in false claims fuels support for violence even among civilians.

I. Steven Kull, Clay Ramsay, and Evan Lewis, "Misperceptions, the Media, and the Iraq War," Political Science Quarterly 118, no. 4 (Winter 2003–2004): 569–598, https://doi.org/10.1002/j.1538-165X.2003.tb00406.
II. Ibid.
III. Lewandowsky et al., "Misinformation, Disinformation, and Violent Conflict..."

And, the agitating effect of misinformation is not limited to the violent actors alone: A study of conspiracy theories run by the Harvard Kennedy School found significant statistical evidence that, even if believers are not moved to acts of violence themselves, they are more they are more likely to support violence enacted by others.[11]

This relationship, too, may not be limited to within the test case. Lewandowsky et. al's study of public support for the American invasion of Iraq indicates that the correlation between misinformation support for violence extends to widespread beliefs as well as conspiracies.[12] (Misinformation does not necessarily have to be of civilian creation.)

"Dissemination and control of information are indespensible ingredients of violent conflict."

While cases of violent actors inspired by misinformation can be documented and evaluated, the effect of passive actors in quiet support of the violence cannot be quantified. The statistical and anecdotal evidence discussed here may actually underrepresent the actual relationship between misinformation and violence.

Automatic Fact-Checkers as a Countermeasure

Given the many negative aspects of political violence, it's decidedly important to mitigate the spread of misinformation. Automatic fact-checkers (AFCs) are an emergent technology with the capacity to detect digital misinformation. AFCs are both relatively new and usually proprietary, so their precise individual mechanics are continually evolving and typically unseen by the public.

Most generally, however, AFCs use two main approaches to detecting misinformation. First, they interrogate the claim itself. An AFC might find a logical fallacy within the claim, or even digital evidence

www.scu.edu/ethics/internet-ethics-blog/misinformation-and-violence/.

11. Adam M. Enders et al., "The Relationship between Conspiracy Theories and Political Violence," Misinformation Review, 2022, http://misinforeview.hks.

12. Stephan Lewandowsky et al., "Misinformation, Disinformation, and Violent Conflict: From Iraq and the 'War on Terror' to Future Threats to Peace," American Psychologist 68, no. 7 (2013): 487–501, https://doi.org/10.1037/a0034515.

of tampering.[13]AFCs can also be coded to find more subtle flags for misinformation. Their programming can include artificial intelligence structures, Natural Language Processing frameworks, or adjacent designs that permit them to "learn" complex patterns in media.[14] In development, the nascent AFCs are fed "training data," large datasets that include both true and false claims, from which they "learn" characteristics or digital footprints that correlate with misinformation.[15] For instance, an AFC might learn from its training data that false claims are more likely to include hyperbolic statements (eg, "greatest," "bestest," "of all time") than true claims. That AFC will be more likely to flag statements containing hyperbolics as misinformation.

Second, AFCs can utilize web-scraping techniques to compare claims to other data on the internet. When provided a textual claim–or image, or video–the AFC can search for supportive or contradictory evidence on other websites.[16]

AFCs using such language processing models are generally accurate in determining the truth or falsehood of claim when equipped to verify the data they handle.[17] A major benefit of these programs is their capacity to verify claims faster and cheaper than human fact checkers.[18, 19] This gives them market appeal and may make them more likely to be implemented among less affluent (or less truth-focused) companies. AFCs can effectively verify images, videos, and audial media–including AI-generated or "deepfake" material–as well as text.[20] And, critically, users tend to believe them: AFCs reduce a user's likelihood of believing false information in both the short-and long-term, and may even make them less likely to believe misinformation in future exposures.[21]

For each one of these capacities comes a limitation. For instance: If a claim is neither strictly true nor entirely false, an AFC geared toward a binary verdict cannot make an accurate judgement.[22, 23] Or, since AFCs are privately developed, different programs will operate with different metrics, be trained on different data, and potentially return different judgements of veracity on the same claim.[24]

Perhaps the most pervasive issue is intrinsic bias. Proponents of AFCs may claim that these programs, by virtue of being programs, eliminate the inherent judgement bias of human fact checkers. But, training datasets themselves can introduce an element of bias. If a disproportionate amount of the dataset's false claims contain the word "security," the AFC might be more inclined to label any claims containing that word as false. Or, in a more malicious scenario, an AFC developer could easily manipulate an AFC's go-to comparison sources to influence verification with a partisan or ideological slant.[25]

13. Mubashara Akhtar et al., "Multimodal Automated Fact-Checking: A Survey," in Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023, 5433, https://aclanthology.org/2023.findings-emnlp.361/.

14. H. Akin Unver, "Emerging Technologies and Automated Fact-Checking," SSRN Electronic Journal, 2023, 2, https://edam.org.tr/Uploads/Yukleme_Resim/pdf- 28-08-2023-23-40-14.pdf.

15. Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu, "Misinformation in Social Media: Definition, Manipulation, and Detection," ACM SIGKDD Explorations Newsletter 21, no. 2 (2019): 87, https://doi.org/10.1145/3373464.3373475.

16. Akhtar et al., "Multimodal Automated Fact-Checking," 5434.

17. Dorian Quelle and Alexandre Bovet, "The Perils and Promises of Fact-Checking with Large Language Models," Frontiers in Artificial Intelligence 7 (2024): 3, https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1341697/full.

18. Ibid., 2.

19. Gionnieve Lim and Simon T. Perrault, "XAI in Automated Fact-Checking? The Benefits Are Modest and There's No One-Explanation-Fits-All," arXiv preprint, 2023, 1, https://arxiv.org/pdf/2308.03372.

20. Akhtar et. al, "Multimodal Automated Fact-Checking," 5438.

21. Jennifer Allen et al., "Evaluating the Fake News Problem at the Scale of the Information Ecosystem," Proceedings of the National Academy of Sciences 118, no. 15 (2021): e2104235118, https://pubmed.ncbi.nlm.nih.gov/32284969/.

22. Lasha Kavtaradze, "Challenges of Automating Fact-Checking: A Technographic Case Study," Emerging Media 2 (2024): 1365–1389, https:/www.researchgate.net/publication/

23. Quelle and Bovet, "The Perils and Promises..."

24. Sian Lee, Aiping Xiong, Haeseung Seo, and Dongwon Lee, "Fact-Checking Fact Checkers: A Data-Driven Approach," Harvard Kennedy School Misinformation Review 4, no. 5 (2023): 1–20, https://misinforeview.hks.harvard.edu/wp-content/uploads/2023/10/lee_fact-checking_fact_checkers_20231026.pdf.

25. Unver, "Emerging Technologies...," 3.

Snopes vs. PolitiFact: A Digital Disagreement[IV]
Snopes and PolitiFact are popular fact checkers broadly considered to be relatively reliable. In the case of AFCs, "reliable" might mean a consistent judgement of truth.

A team at the Harvard Kennedy School tested this theory by comparing the respective judgements that the two AFCs had made about differently-worded claims communicating the same concept. (Eg, Snopes might have verified that "Donald J. Trump is the current president," while Politifact may have judged, "The current US president is Donald Trump.") After examining the AFCs' judgements on 749 parallel claims, researchers found that Snopes and Politifact attributed only 69.6% of the pairs of parallel claims the same rating. Over 30% of the claims were judged to be differently truthful between the respective programs.

These different judgments were ultimately attributed to slight differences in the AFCs' rating systems and subtle semantic differences between the claims. This demonstrates a potential issue with the use of language processing frameworks: Judgements of veracity may be influenced by the way the claim is worded, not only its content.

IV. Lee et al., "Fact-Checking Fact Checkers...."

And, unlike human fact checkers, AFCs can be influenced by cyber attacks. "Planting," an ostensibly popular method, occurs when an attacker tampers with a source that AFCs use to verify claims. If an AFC was programmed to compare claims to a specified news source, for example, a hacker could alter the statements on that page to align with the misinformation they wish to promote. Planting can be effective when as low as one sentence is inserted.[26]

Alternatively, hackers could flood a social media site with AI-generated statements in support of the false claim, potentially causing the AFC to find that the claim is consistent with general consensus.

While this may seem like a lot of work to disrupt one claim verification, planting could theoretically change widespread public perception if the attacked AFC is being used on a large scale. In this way, AFCs can actually backfire by verifying misinformation that has been intentionally created by a malicious party.

Recommendations
AFCs can–and should–be used effectively to prevent the spread of digital misinformation. However, their vulnerabilities necessitate precautionary measures. Mitigating these risks calls for the participation of not only the developers, but any applicable regulatory power and the users themselves.

1. **Developers** should code AFCs with the capacity to disclose the external sources used to verify a claim to the user. Sites using AFCs often state when claims have been checked; those statements should include an option to view a list of the pages it was checked against. This is not a mechanically complex fix. However, it would assuage a few key issues: Any selection bias in sources used to verify would be immediately evident to the user; it would be evident if few sources were in agreement; and, in the case of cyber attack, it would be immediately clear which sources had been tampered with.

2. A **regulatory power** should commit to certifying the impartiality of privately-developed AFCs that will be used by the public. Before becoming available on the market, a qualified entity–be it governmental or otherwise–should review the AFC's hardcoding (relationships learned from training data, etc) and ensure no ingrained bias. This is a significantly more difficult recommendation to adopt. However, a company's own report of an unbiased product may not be reliable.

3. **Users** must understand that AFCs are not invulnerable. While their judgments are generally correct, statements supported by AFCs alone cannot be blindly accepted as fact.

26. Sahar Abdelnabi and Mario Fritz, "Fact-Saboteurs: A Taxonomy of Evidence Manipulation Attacks Against Fact-Verification Systems," in Proceedings of the 32nd USENIX Conference on Security Symposium (SEC '23), 2023, 6719–6736, https://www.usenix.org/conference/sec23/presentation/abdelnabi.