Studies in Philosophy Politics and Economics

Josef Huber (2020) *Fighting Fake News Successfully: Why Rini's Display of Reputation Scores Fails and How to Remedy Her Account*, Studies in Philosophy, Politics and Economics, 2(1), 39-44

Fighting Fake News Successfully: Why Rini's Display of Reputation Scores Fails and How to Remedy Her Account

Josef Huber*

* Philosophy, Logic and Scientific Method, London School of Economics, Houghton Street, London, WC2A 2AE, UK. Email: josefhuber4@gmx.at. URL: https://www.linkedin.com/in/josef-huber-813507124/

Abstract

How can we prevent sharing of Fake News on social media? This paper critiques Regina Rini's proposal (2017) to introduce reputation scores displaying individuals' propensity to share Fake News. Based on empirical data, I argue that Rini's approach is too narrow in its interpretation of users' motivations and may ultimately backfire. Given additional motivations to spread Fake News highlighted by Bucher (2012) and Chaput (2010), I suggest using reputation scores to diversify feeds without displaying the scores to users. I conclude that more empirical work needs to be done to understand why people consume and share Fake News.

Keywords: Fake News; Social Media; Misinformation; Democracy; Rini

In this essay, I analyse and critique Regina Rini's (2017) proposal to curb the spread of Fake News (henceforth FN) on social media by displaying reputational scores which depend on users' propensity to share FN articles. I argue that this approach could backfire and ultimately exacerbate the spread of FN because it relies on a narrow definition of motivations to share FN. My discussion begins by defining FN and assessing its role in undermining democracy. Next, I introduce Rini's notion of partisan epistemology as a mechanism leading people to share FN online. Following my critique of Rini's solution and an evaluation of additional reasons to consume and share FN, I proceed to present an alternative proposal, that keeps Rini's institutional angle. Feeding reputational scores to social media algorithms without displaying them to users would curb the spread of FN through exposure to a diverse pool of sources and opinions while, at the same time, preventing backfire effects. I conclude that a successful strategy again FN requires more empirical work to better understand why people really consume and share FN.

1. What Is Fake News, and Why Should We Care?

LSE Houghton St Press

Rini (2017) defines FN as a piece of news designed to be shared widely and to deceive by mirroring mainstream reportage, while known, by its authors, to be false. For the purpose of this paper, I assume this definition sufficiently captures FN as a political phenomenon (see Edson et al. 2018 for a typology of definitions).

FN poses a salient threat to core principles of democracy. Democratic participation requires that voters can access *actionable, intelligible* and *accurate* information (Fung 2013). For it is only when citizens are well-informed that they can make political decisions based on their interests. Given voters' time and resource constraints, they thus depend on intermediaries, such as the media, for reliable information. FN undermines this process because it misinforms by inducing readers to confidently believe in falsehoods (Kuklinski et al. 2000). This may lead voters to make decisions counter to what they would have decided had they been perfectly informed. It is often assumed that only *naive* consumers of FN are at risk of misinformation, because they fail to identify FN when confronted with it (Levy 2017). However, even sophisticated readership may see their attitudes vis-à-vis mainstream sources altered after reading FN (Levy 2017: 21). It follows that FN, by distorting citizens' opinions and decision-making, poses a threat to democratic processes and should be fought. A first step would be to combat FN on social media, which has become perhaps its most important channel (Hunt and Gentzkow 2017).

2. Why Is Fake News Shared?

Per Rini (2017), users share FN because they believe news forwarded by their politically affiliated peers. The lack of behavioural norms on social media render the sharing of FN a *bent* form of testimony (Rini 2017): In analogue life, we usually accept propositions because they are endorsed by another person. However, on social media it is not fully clear if a user's sharing of FN or other content automatically implies an endorsement. Despite this ambiguity, Rini argues that sharing FN can be reasonable given *partisan epistemology*: Because it is reasonable for individuals to trust their partisans more than members of rival political groups, and in light of information and time constraints, it may be reasonable for individual social media users to forward FN shared by partisans. Note, at this point, the similarity to the aforementioned account of *naive* FN consumption.

3. Rini's Strategy Against Fake News

Given the reasonableness of sharing by individuals due to partisan epistemology, Rini (2017) locates the responsibility to combat FN in institutions. Direct censorship of FN seems problematic in light of technical difficulties and the dangers of interference with press freedom. Instead, Rini suggests introducing nudges which help users overcome information constraints and better distinguish untrustworthy virtual testimony received from ideologically proximate friends on social media. Her solution is based on an existing initiative by Facebook but differs in an essential aspect. In 2016, the platform introduced notifications warning users before sharing links deemed to contain FN by certified third-party fact-checking agencies (Facebook, accessed on 17 April 2020). Instead of flagging FN articles, Rini suggests monitoring and displaying the testimonial reputation of individual users. If a user regularly shares disputed stories, Facebook should lower their reputation score. Rini argues that her proposed solution would help users think on their feet and refrain from sharing articles posted by users with low reputation scores - an assumption I critique in the next part of this paper.

4. Rini's Solution Is Not Successful Enough

Rini's proposal relies on a narrow interpretation of users' motivation for sharing FN: Partisan epistemology induces belief in the story itself and leads social media users to forward FN. When nudged by reputation scores users will change their behaviour. I argue that this approach focuses too much on the *naive* consumer of FN. There may be more to the story. While it is impossible to precisely identify, *ex ante*, users' motivations for spreading FN, Bucher (2012) and Chaput (2010) highlight additional considerations, which challenge Rini's account. Sharing FN on social media may enable users to *use their voices* to contribute to developing political narratives and *be seen.* Those consuming and sharing FN may do so because they profoundly distrust mainstream institutions and want to support political narratives which are more reflective of their own, contrasting perceptions of the world. In the United States, for instance, only 41% of respondents in a recent study claimed to trust mainstream media sources (Brenan 2019). Individuals may share FN not only because they trust its testifier and fail to recognise it as FN, but also because they believe partisans' testimony *qua its contrast to mainstream reportage*. If that is indeed the case, then Rini's solution could backfire.

Let us consider why this is the case. Nyhan and Reifler (2010) find that a so-called *backfire effect* causes corrective alerts on articles promoting misleading claims to reinforce misinformation among ideological subgroups. This may be because readers' forceful efforts to argue against corrective information lead them to reinforce their original position (Lodge and Taber 2000). Similarly, rather than update misperceptions, Rini's reputational scores on Facebook may strengthen belief in FN. It seems unlikely that users, who reasonably trust their partisans and distrust the mainstream, should throw partisan epistemology overboard in favour of Facebook's reputation scores. After all, Facebook itself is widely seen as a powerful operator and part of the mainstream (Bell 2019). A recent survey on public attitudes towards technology companies concludes that 72% of US Americans think Facebook has too much power, and that most people trust Facebook less than other tech-giants (Newton et al. 2017). Instead of highlighting untrustworthiness, a bad testimonial reputation score may thus exacerbate the spread of FN by signalling to some partisan groups that a low-ranked user regularly shares articles they should trust *because* they diverge from those propagated by the mainstream.

One could argue that Rini's proposal may still be tenable in light of this criticism because the above backfire effects concern a specific type of user only. However, even those who generally do not mistrust the mainstream may not be helped by the display of reputation scores. Pennycook et al. (2020) find that the display of nudges in the form of content warnings renders individuals reliant on aides to identify FN and increases perceived accuracy on unflagged FN articles (they label this an *implied truth effect*). Analogously, in the case of reputation scores, users may begin to rely on Facebook's scores too much rather than think on their feet, as Rini intended.

So far, I have shown that Rini's proposal faces several issues. Its focus on the partisan epistemology of FN sharing ignores other potential motivations for sharing FN online. I now propose an alternative solution.

5. Fighting Fake News Successfully

A successful strategy against FN must first attempt to better understand the motivations behind users' sharing behaviour. This requires more empirical research. Nevertheless, a first step can be made in the right direction by adjusting Rini's proposal to avoid its greatest pitfalls. Reputation scores could be fed to Facebook's algorithm to diversify news walls without showing the scores to users. This would prevent backfire and implied truth effects caused by the open display of scores. Instead of reinforcing notorious echo-chambers which strengthen users' ideological preconceptions (Passe et al. 2018), score-fed algorithms would ensure that newsfeeds are sufficiently diverse in terms of content by users with low and high propensity to share FN articles. Exposing individuals to a heterogeneous set of sources and views, enhances their openness to new ideas and fosters depolarisation, as a study by Beam et al. (2018) suggests. Reputation-score-fed algorithms would thus help regain the trust of those who have come to deeply mistrust mainstream news outlets.

My proposal extends the process of partisan epistemology to social peers, assuming that users may reasonably trust and share testimony not only from their political partisans but also from their virtual social peers, due to the role of social endorsement in reducing partisan selective exposure on social media (Messing and Westwood 2014). Indeed, Facebook users' virtual peers tend to be at least somewhat ideologically diverse. According to Bakshy et al. (2015), on average more than a fifth of every individual Facebook users' virtual friends identify with opposing political views. Given an average of 338 Facebook friends per user (and many more for active users; Smith 2019) this creates ample opportunity for virtual exposure of individuals not only to social peers with a diverse range of views but also numerous different sources in their newsfeeds. My proposal thus retains an institutional angle and may be able to tackle a wider range of users' motivations to share FN while avoiding backfire or implied truth effects. Instead of displaying reputation scores users may misinterpret or rely on, score-fed algorithms would make certain individuals are confronted with posts shared by peers with diverse reputation scores. This, in turn, could diversify the set of news users share away from FN.

Reputational scores will have to be reliable, unbiased, and include as much objective information as possible. While a lot remains to be said about this, Facebook could start by being fully transparent about the fact-checking agencies it hires, as well as the data it uses and how it uses it to compute scores. To maximise transparency and trust, Facebook could also cooperate with academic institutions to ensure objective rigour of this process, or even reach out to its users, whose trust it needs most urgently. Finally, in order to help foster behavioural norms and counteract bent testimony-giving, users could be required to comment on the news they share, or, at least, express (dis-)approval, for instance through the use of the appropriate emojis.

Critics may argue that my approach is paternalistic and could undermine trust in Facebook further by making users feel censored. However, my proposal explicitly does *not* seek to censor users' contributions or shares. Rather, it attempts to ensure that users are exposed to a maximum of diverse sources and opinions online. Furthermore, these concerns could be quelled by introducing reputation-score-fed algorithms as a default setting users can opt out of. While, of course, some may make immediate use of this function, thus undermining the positive effect of reputation-score-fed algorithms, it seems reasonable to assume that at least some users would not opt out. Ultimately, however, fighting FN successfully may require a plethora of different policies and adjustments to tackle a wide breadth of possible motivations to share FN. For instance, to fight misinformation around the coronavirus pandemic, Facebook relies on fact-checkers, nudges, alerts, and algorithmic brakes on its spread (Wong 2020).

6. Conclusion

In this essay, I argued that Rini's proposal to curb FN on social media is based on a narrow interpretation of users' motivations to share FN. The display of users' scores may worsen the issue through unintended side-effects. Instead, I suggested feeding reputation scores to Facebook's algorithms without showing them to users. Finally, I assert that all strategies against FN will involve trade-offs. To minimise these as much as possible, future scholarship should conduct further empirical research into why, exactly, FN is shared.

References

Bakshy, E., Messing, S. and Adamic, L. A. 2015. "Exposure to ideologically diverse news and opinion on Facebook". *Science* 348(6239): 1130-1132.

Beam, M. A., Hutchens, M. J. and Hmielowski, J. D. 2018. "Facebook news and (de)polarization: reinforcing spirals in the 2016 US election". *Information, Communication & Society* (21)7: 940-958.

Bell, E. 2019. "Facebook and Twitter are growing into the mainstream". *The Guardian*, June 2 2019. URL: https://www.theguardian.com/media/commentisfree/2019/jun/02/social-platforms-facebook-debate-regulation

Brenan, M. 2019. "Americans' Trust in Mass Media Edges Down to 41%". *Gallup*. September 26 2019. URL: https://news.gallup.com/poll/267047/americans-trust-mass-media-edges-down.aspx

Bucher, T. 2012. "Want to be on top? Algorithmic power and the threat of invisibility on Facebook". *New Media & Society* 14(7): 1164-1180.

Chaput, C. 2010. "Rhetorical circulation in late capitalism: Neoliberalism and the overdetermination of affective energy". *Philosophy & Rhetoric* 43(1): 1-25.

Edson C. T. Jr., Lim, Z. W. and Ling, R. 2018. "Defining "Fake News"", *Digital Journalism* 6(2): 137-153. DOI: 10.1080/21670811.2017.1360143

Fung, A. 2013. "Infotopia: Unleashing the Democratic Power of Transparency". *Politics & Society* 41(2): 183-212.

Facebook. n.d. "How is Facebook addressing false news through third-party fact-checkers?" Accessed April 17, 2020. URL: https://www.facebook.com/help/1952307158131536

Hunt, A., and Gentzkow, M. 2017. "Social Media and Fake News in the 2016 Election". *Journal of Economic Perspectives* 31(2): 211-236.

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D. and Rich, R.F. 2000. "Misinformation and the Currency of Democratic Citizenship". *The Journal of Politics* 63(3): 790-816.

Levy, N. 2017. "The Bad News About Fake News". *Social Epistemology Review and Reply Collective* 6(8): 20-36.

Lodge, M., and Taber, C. S. 2000. "Three steps toward a theory of motivated political reasoning". In *Elements of reason: Understanding and expanding the limits of political rationality,* edited by Lupia, A., McCubbins, M. D ad Popkin, S. L. Taken from: Nyhan, B. and

Reifler, J. 2010. "When Corrections Fail: The Persistence of Political Misperceptions". *Political Behaviour* 32 (2): 303-330.

Mena, P. 2019. "Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook". *Policy and Internet* 12: 165-183. DOI:10.1002/poi3.214

Messing, S., and Westwood, S. J. 2014. "Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online". *Communication Research* 41(8): 1042-1063.

Newton, C., Statt, N. and Zelenko, M. 2017. "The Verge Tech Survey. How Americans really feel about Facebook, Apple, and more". *The Verge*. October 27 2017.
URL: https://www.theverge.com/2017/10/27/16550640/verge-tech-survey-amazon-facebook-google-twitter-popularity

Nyhan, B. and Reifler, J. 2010. "When Corrections Fail: The Persistence of Political Misperceptions". *Political Behaviour* 32(2): 303-330.

Passe, J., Drake, C. and Mayger, L. 2018. "Homophily, Echo Chambers, & Selective Exposure in Social Networks: What Should Civic Educators Do?". *The Journal of Social Studies Research* 42(3): 261-271.

Pennycook, G., Bear, A., Collins, E. T. and Rand, D. G. 2020. "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings". *Management Science Articles in Advance*: 1-14.

Rini, R. 2017. "Fake News and Partisan Epistemology". *Kennedy Institute of Ethics Journal* 27(2): E-43.

Smith, K. 2019. "53 Incredible Facebook Statistics and Facts". *Brandwatch*. June 1 2019.
URL: https://www.brandwatch.com/blog/facebook-statistics/

Wong, J.C. 2020. "Coronavirus: Facebook will start warning users who engaged with 'harmful' misinformation". *The Guardian.* April 16 2020. URL:
https://www.theguardian.com/technology/2020/apr/16/coronavirus-facebook-misinformation-warning#maincontent