Harvard Kennedy School Misinformation Review[1]
February 2024, Volume 5, Issue 1
Creative Commons Attribution 4.0 International (CC BY 4.0)
Reprints and permissions: misinforeview@hks.harvard.edu
DOI: https://doi.org/10.37016/mr-2020-134
Website: https://misinforeview.hks.harvard.edu

Research Article

Debunking and exposing misinformation among fringe communities: Testing source exposure and debunking anti-Ukrainian misinformation among German fringe communities

Through an online field experiment, we test traditional and novel counter-misinformation strategies among fringe communities. Though generally effective, traditional strategies have not been tested in fringe communities, and do not address the online infrastructure of misinformation sources supporting such consumption. Instead, we propose to activate source criticism by exposing sources' unreliability. Based on a snowball sampling of German fringe communities on Facebook, we test if debunking and source exposure reduce groups' consumption levels of two popular misinformation sources. Results support a proactively engaging counter-misinformation approach to reduce consumption of misinformation sources.

Authors: Johannes Christiern Santos Okholm (1), Amir Ebrahimi Fard (2), Marijn ten Thij (2)
Affiliations: (1) Department of Political and Social Sciences, European University Institute, Italy, (2) Department of Advanced Computing Sciences, Maastricht University, Netherlands

How to cite: Santos Okholm, J. C., Ebrahimi Fard, A., & ten Thij, M. C. (2024). Debunking and exposing misinformation among fringe communities: Testing source exposure and debunking anti-Ukrainian misinformation among German fringe communities. *Harvard Kennedy School (HKS) Misinformation Review*, *5*(1).
Received: June 19th, 2023. Accepted: January 11th, 2024. Published: February 19th, 2024.

Research questions

- RQ1: Do counter-misinformation strategies reduce fringe communities' consumption of misinformation sources?
- RQ2: Is source exposure better at reducing fringe communities' consumption of misinformation sources than debunking?
- RQ3: Can exposing fringe communities' gatekeepers to counter-misinformation strategies reduce the communities' consumption of misinformation sources?

Essay summary

- In collaboration with the fact-checking organization VoxCheck, member of the International Fact Checking Community, we conducted an online field experiment testing whether counter-misinformation strategies can reduce fringe communities' consumption of misinformation media on Facebook. Using snowball sampling, we identified public Facebook groups that regularly consume German misinformation sources and posted either debunks of the anti-Ukrainian misinformation claims or exposed the two most prominent misinformation sources' bad track record for spreading misinformation.

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

- We found that debunking anti-Ukrainian misinformation claims does not reduce fringe communities' consumption of misinformation sources.
- We found that exposing sources' bad track record of spreading misinformation and biased reporting reduces fringe communities' consumption of misinformation sources.
- We found that exposing gatekeepers among fringe communities lowers their acceptance of content from targeted misinformation sources.
- Our findings support a more proactively engaging approach to counter misinformation of reaching out to fringe communities.
- Our findings indicate that source-focused counter-misinformation strategies are effective in addressing the growing network of misinformation sources.
- By showing the feasibility of independent online field experiments, our study opens up a field dominated by clinical experiments to more realistic testing of counter-misinformation strategies.

Implications

Existing counter-misinformation strategies have yet to explicitly address the relationship between fringe communities and misinformation sources. Focusing instead on verifying single claims for a broad audience (Parks & Toth, 2006; Pennycook et al., 2020; Roozenbeek et al., 2020), they do not address the infrastructure of misinformation sources that continuously provide misinformation to fringe audiences, who consume misinformation on a regular basis. As these communities are especially susceptible to misinformation (Lewandowsky et al., 2020; Nyhan & Reifler, 2010), a strategy to reduce their consumption of misinformation is needed.

We, therefore, present *source exposure* as a potential strategy, focusing on limiting consumption rather than limiting audience acceptance of misinformation. Source exposure focuses on the sources that contain misinformation. We argue that by exposing and highlighting misinformation sources' unreliability and debunking their track record to fringe communities, these communities will update their perspective of exposed sources and reduce their consumption to avoid being misled. As online communities face a complex stream of information, highlighting a source's unreliability can provide a point of reference similar to debunking (Lewandowsky et al., 2020), lower the costs of conducting source criticism (Ahlstrom-Vij, 2016), and activate members' media literacy (Steensen, 2019) and analytical thinking (McKernan et al., 2023). We tested this through an online field experiment to reduce fringe communities' consumption of misinformation sources using both debunking and a novel counter-misinformation strategy, source exposure, which focuses on the misinformation infrastructure. Through this, we hope to provide a more proactively engaging alternative to more reactive and passive debunking by engaging with vulnerable audiences and being focused on the credibility of sources of misinformation.

Though numerous studies show that highlighting the logical and factual fallacies of misinformation effectively reduces its influence on peoples' beliefs and activates their analytical thinking (Bode & Vraga, 2018; Bode et al., 2020; Chan et al., 2017; Lewandowsky et al., 2020; Martel et al., 2021; Parks & Toth, 2006; Roozenbeek et al., 2020; Skurnik et al., 2005; Skurnik et al., 2000; van der Linden et al., 2017), the literature and practice of debunking have three shortcomings when countering misinformation.

To date, the academic literature overly relies on clinical experiments that are performed in simulated conditions which are removed from the reality of digital misinformation. Moreover, these experiments make use of unrepresentative participants. These facts undermine the generalizability of the findings of such experiments. For example, Roozenbeek et al. (2020) and Pennycook et al. (2020) make use of educational simulations and subsequent in-simulation evaluations which monopolizes participants' attention and does not account for the complexities of social dynamics and attention scarcity on social media platforms (Hendricks & Hansen, 2014). Moreover, Munger et al.'s (2021) finding of online recruitment services' issues with generalizability—due to heavy recruitment bias towards younger and more digitally literate participants—poses further questions for prevailing methods (e.g., Martel et al., 2021; Pennycook & Rand, 2019). While these studies provide valuable insights on the use of counter-misinformation strategies, they do not necessarily translate to high-risk audiences. Our study addresses this by conducting an online field experiment and showing the feasibility of testing counter-misinformation strategies in the online world.

Second, with a focus on general audiences, the debunking literature has not tested interventions among fringe communities, who are more susceptible to believe and consume misinformation. While this may largely be due to the difficulty of accessing fringe communities, who shun an academic elite, these understudied groups are nevertheless particularly relevant for the study of misinformation. Driven by distrust of authorities and belief in conspiracy theories (Bruder et al., 2013; Imhoff & Bruder, 2014), experiences of status loss (Bor & Petersen, 2022; Petersen et al., 2023), or feelings of marginalization (Freelon et al., 2020), such communities are more likely to fall for misinformation. As members strongly identify with their political beliefs, they are prone to rejecting debunking efforts that contradict their beliefs (Michael & Breaux, 2021; Nyhan & Reifler, 2010; Osmundsen et al., 2021; Shin & Thorson, 2017; Zollo, 2019). Strong ideologies and conspiracy theories provide elaborate worldviews that are difficult to debunk as they are based on values and beliefs. Based on this, fringe communities can be understood as communities that feel disenfranchised, distrust authorities, and hold strong ideological beliefs and world views that are considered extreme or fringe by the rest of society. Due to these strong biases, adherents are likely to accept false information that is aligned with their personal beliefs. This has structural consequences for their exposure to information, as online information flows remain within ideological homogenous networks (Marchal, 2021; Pogorelskiy & Shum, 2019; Rathje et al., 2021; Zollo, 2019) and are worsened by fringe communities migrating to alternative social platforms to avoid fact-checkers and digital censorship (Guhl et al., 2020; Nouri et al., 2021; Trujillo et al., 2020). Moreover, a recent study by Aslett et al. (2023) found that such confirmation bias remains persistent when online users try to use search engines as a proxy for debunking, as online searches also give users access to multiple low-quality sources of information.

While this positions fringe communities outside the reach of fact checkers, our study indicates the effectiveness of a more proactively engaging approach by reaching out to fringe communities to reduce their misinformation consumption. Despite the abovementioned literature on biases among fringe communities and the little effect that corrections have on beliefs, we find that counter-misinformation strategies can be effective in changing consumption patterns. While beliefs may be hard to change, our results indicate that new misperceptions can be stopped from reaching these audiences. A common fear among fact checkers is that such initiatives would be stopped by the communities' gatekeepers (e.g., by deleting posts). However, we found that group administrators who rejected the posting of our treatments (i.e., source exposure and debunking posts), were subsequently found to allow less content from exposed misinformation sources in their Facebook groups. Though gatekeepers may halt outreach, their position as curators of online discussion can be co-opted, as they themselves are also susceptible to corrective interventions. What motivates this change in curation practices is unclear. The change may either be driven by altruistic motivations of informing one's community or be due to fears of censorship, triggered when interventions undermine the notion of these Facebook groups being safe spaces for alternative discourse by highlighting their public nature.

Thirdly, the focus on single claim validity does not address the large infrastructure of misinformation sources (e.g., media, blogs, influencers, and other communication channels) that sustain fringe communities' consumption of misinformation (DiResta & Grossman, 2019; Starbird, 2017). Recent studies by Donovan et al. (2022) and Rothschild (2021) describe how this infrastructure absorbs new trending misinformation claims and pushes them on a global scale, for example, anti-vax movements during the COVID pandemic (Burki, 2020; Cinelli et al., 2020; Johnson et al., 2020). While this production of misinformation on an industrial scale outpaces fact-checking of claims, our study shows that consumption can be reduced by source exposure, but not by debunking. It shows that even fringe audiences are susceptible to corrective interventions and will update their consumption of misinformation sources if they are notified of the sources' bad track records. Though social media platforms already use similar methods of labelling content "state-controlled media," independent studies of such methods repeat similar mistakes as above of clinical conditions (Nassetta & Gross, 2020; Pennycook et al., 2020).

Our findings, therefore, give clear indications of effective counter-misinformation strategies to stakeholders. The effectiveness of source exposure among fringe communities and gatekeepers teaches us that there is added value for practitioners and organizations to 1) adopt a proactively engaging approach towards regular consumers of misinformation and 2) go beyond claim-based countermeasures. Moreover, the fact that consumption is reduced also highlights there is benefit of allowing fact checkers to independently access and engage with regular consumers of misinformation on social platforms. We do stress that this should complement content moderation and existing fact-checking initiatives.

Limitations

While the design of our study is adapted to the real online environment of fringe communities, this also meant adapting to shifting conditions. First, as the European Union banned Russian misinformation sources in March 2022, Facebook purged several fringe groups and misinformation sources, decimating our original pool of public Facebook groups. This forced us to balance between enough groups with regular consumption of the same misinformation sources and focusing on a limited number of sources (n = 35) in accordance with VoxCheck resources, at the expense of our generalizability. This small sample size also underscores the practical challenges of studying fringe communities who usually shun academic investigations as part of the oppressive establishment. Second, as German authorities arrested members of the far-right Reichsbürger Movement on December 7th of 2022, the week after VoxCheck's postings, we saw a surge in general posting activity among all surveilled experiment groups. This surge increases the risk of canceling out the effect of certain treatments (e.g., debunking), which is only accentuated by our small sample. Hence, our findings should be taken with some reservations.

In addition, the observational nature of the field experiment holds a limitation on mechanism, as well as the methodological set-up of studying group reactions. Though our study shows a strong correlation between source exposure, targeting gatekeepers with corrective interventions, and lower consumption, we cannot fully explain the individual reasoning behind this change in behavior. As gatekeepers' reasons for behavioral change cannot be explored without conducting in-depth interviews, this lies outside our current scope but will be a natural next step in studying counter-misinformation strategies among fringe communities.

Findings

Finding 1: Counter-misinformation strategies reduce fringe communities' consumption of misinformation sources.

Our analysis indicates that interventions on misinformation can be effective in lowering consumption among fringe groups (see Methods section "Selecting Facebook groups"). Our results in Figure 1 indicate that groups who received treatments (i.e., source exposure and debunking posts) did indeed have a lower consumption than the control group 2 and 2–4 weeks after the treatment. However, looking at Table 1, we see that the effect is only statistically significant in week 2.

The treatment coefficient of -0.630 $(SE = 0.301, p < 0.05)$ indicates that groups who were given either debunking or source exposure posts had a significant effect of reducing consumption of the two misinformation sources by 63%. Though this is significant, we caution against overstating the model's coefficient due to our small n. In our analysis, we controlled for group sizes ($\log$(Group Size)) and pre-treatment consumption frequencies (2 weeks pre-treatment). As larger groups are likely to be more active and posting frequency may crowd out our treatment, we see a return of a positive correlation 4 weeks after the treatment, coinciding with the effect of counter-strategies disappearing. We see that a 1% increase of group size leads to a 38% increase in consumption of misinformation media. We also see that prior high consumption of misinformation is significantly correlated with higher consumption levels both 2 and 4 weeks after the treatment is given. Though being low, the coefficients for "2 weeks pre-treatment" of $0.131\ (SE = 0.022, p < 0.01)$ and $0.159\ (SE = 0.030, p < 0.01)$ do show that high-consumption of misinformation remains correlated with continued consumption despite counter-misinformation strategies. Hence, a core of dedicated consumers persists despite interventions.

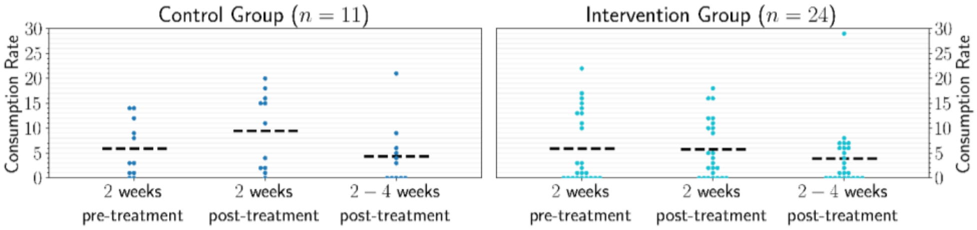


Figure 1. Observed consumption rates of misinformation media for all Facebook groups. *The left panel depicts groups that were part of the control group, and the right panel shows all groups in which we placed an intervention. Black dashed lines indicate the mean of the observed values.*

Finding 2: Source exposure reduces fringe communities' consumption of misinformation sources.

On a closer look at the effect of our treatments, we find that debunking statements (displayed in red in Figure 2) do not change consumption levels of the misinformation source. Meanwhile, we see that exposing sources' bad track record (displayed in green in Figure 2) to fringe groups indeed lowers their consumption. This shows the utility of source exposure to address the continued consumption of misinformation among high-consuming audiences.

Looking at Table 2 gives us a more nuanced understanding of these findings. With a statistically significant coefficient of $-1.176\ (SE = 0.592, p < 0.05)$ 2–4 weeks after treatment, indicating a 117.6% decrease of consumption of misinformation, source exposure had a long-term reduction of groups' consumption of misinformation sources. The absence of short-term effects may, however, be due to the increase in consumption of the two targeted misinformation sources, following the foiled coup d'état of the Reichsbürger Movement.

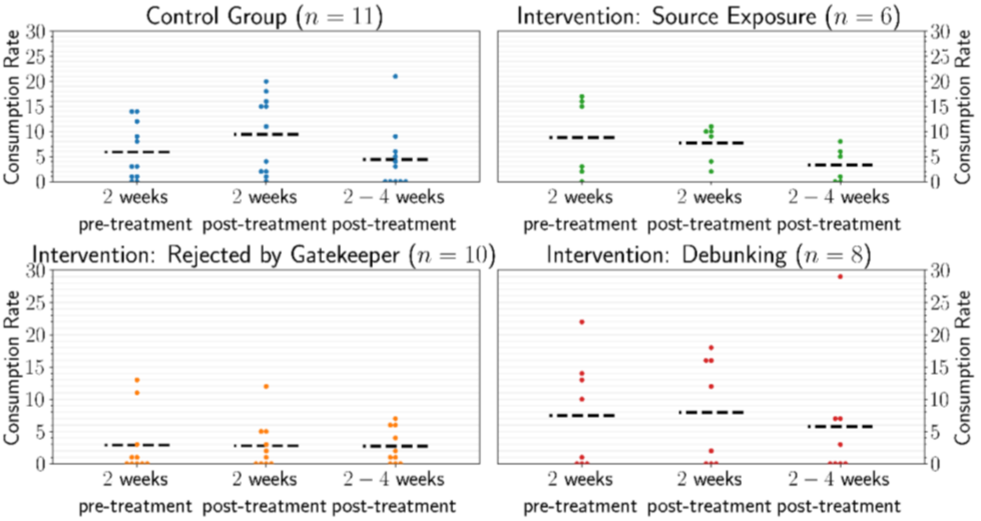


Figure 2. Observed consumption rates of misinformation media for all Facebook groups. *The top left panel depicts groups that were part of the control group and the groups that received the three different interventions, i.e., rejected by gatekeeper, source exposure, and debunking. These are displayed in the bottom left, top right, and bottom right panels, respectively. Black dashed lines indicate the mean of the observed values.*

Finding 3: Notifying group administrators lowers the consumption of misinformation media.

More surprisingly, however, we see that exposing group administrators to counter-misinformation strategies was associated with the most significant reduction of consumption. This indicates that despite denying our treatment to be posted in the groups, these gatekeepers would followingly modify which sources they allowed to be posted in their forums in the future.

Exposure to group administrators had a statistically significant coefficient of -0.836 $(SE = 0.390, p < 0.05)$ the first two weeks after the treatment, indicating an 83% short-term reduction of consumption. This indicates that despite an increase in demand among the fringe community due to the foiled coup, gatekeepers were less likely to allow posts with the two targeted misinformation sources. Meanwhile, coefficients for group size and previous consumption did not change in statistical significance and only had marginal changes in coefficients, similar to Table 1.

Methods

As our research questions focus on the effects of interventions on misinformation, specifically among fringe communities, we begin by identifying the online presence of fringe communities on Facebook by measuring the frequency at which these groups share links to specific misinformation sources. We test whether our interventions are effective using a Negative Binomial Regression Model (see Hilbe, 2011), as it is less vulnerable to extreme outliers. The data for this analysis was collected with CrowdTangle and contains all posts on the Facebook pages of the selected groups that were posted in December 2022 and earlier. For our analysis, we selected all German fringe communities and investigated their consumption of two sources of misinformation (i.e., *Deutschland Kurier* and *Reitschuster*) both before and after our intervention in November 2022. This section describes our method in more detail.

Data

Selecting Facebook groups. To build our collection of Facebook groups used by fringe communities, we use a snowball sampling starting from a list of multiple prominent misinformation sources. By identifying one group through its relation to others, this sampling method is particularly effective in identifying and mapping hidden and stigmatized communities, such as fringe communities of misinformation (Browne, 2005; Petersen & Valdez, 2005), and has previously been used to study the online spread of misinformation among communities (Badaway et al., 2019; Hindman & Barash, 2018).

This list, containing 17 outlets, was constructed through a literature review of academic articles, government reports, investigative journalism, and fact checkers in the fall of 2021. All selected groups or blogs were subsequently identified as far-right/left, conspiracy, anti-vax, and anti-establishment, and as regular spreaders of misinformation, based on their online content, website content, and the labels used by reviewed literature. Due to an emphasis on far-right communities in the reviewed literature, the sample has a right-wing bias, with groups often being anti-immigrant, Islamophobic, and anti-left wing. Finally, the list and labels used were verified by the German fact checker Correctiv, a member of Poynter's International Fact-Checking Network, as the main spreaders of online misinformation.

To identify fringe groups, we use CrowdTangle to select public Facebook groups that post or share links with URLs on our list on a weekly basis throughout 2021. This led to a list of 148 public Facebook fringe groups, whose historic data of posts we collected in May 2022.

Measuring consumption. To compute the consumption rate of a top-level domain (e.g., https://reitschuster.de) in a Facebook group, we count the times this top-level domain appears in the posts during the predefined period of t and then divide it by the length of this period. Here, we set this inspection period to two weeks, i.e., identifying consumption on a bi-weekly basis, as previous studies (McCabe & Smith, 2002; Osmundsen et al., 2021; Zerback et al., 2021) find intervention effects to wane after this time period.

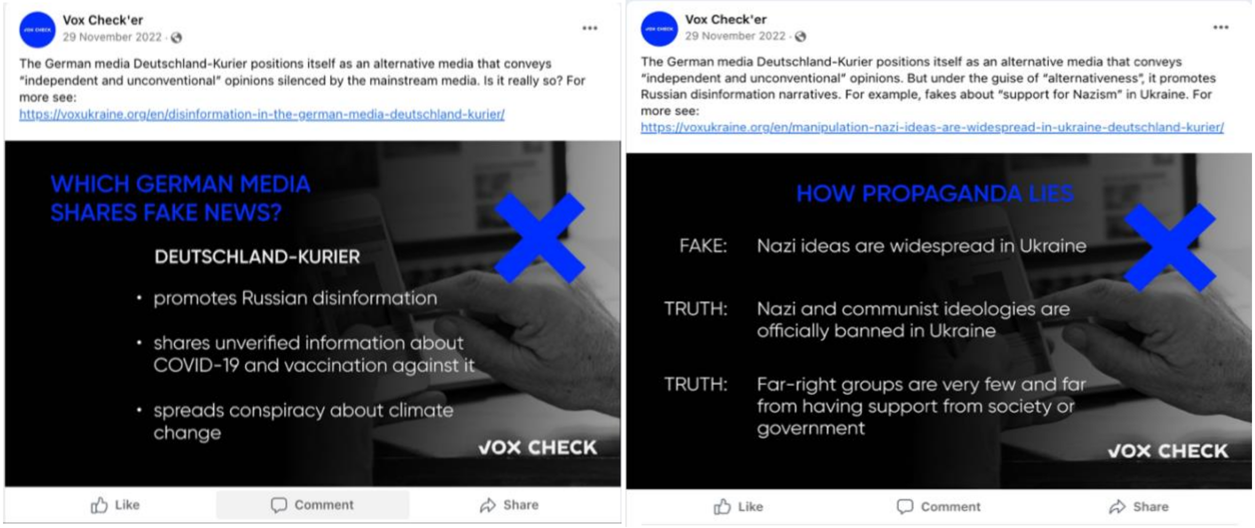
Obtaining consumption levels. After updating our historical data for all 148 groups with CrowdTangle in November 2022, we found that two misinformation sources (*Deutschland Kurier* and *Reitschuster)* were most frequently consumed in all groups. More specifically, 35 groups consumed these sources regularly (i.e., at least once per two weeks) between September and October 2022. We selected these groups as our final sample. Finally, we repeated the data-gathering process again in December 2022 to obtain the consumption levels of the groups after the intervention.

Experimental design

Intervention groups. We divide our final sample of 35 Facebook groups into three groups: a control set consisting of 11 Facebook groups, a source exposure treatment set containing 12 Facebook groups, and a debunking treatment set containing 12 Facebook groups.

Treatment. Treatments consisted of two posts made by VoxCheck through a Facebook profile (as shown in Figure 3), clearly stating relation to the fact-checking organization in its bio and using the VoxCheck logo. To make posts comparable, they focused on the same source and followed the same design of 4-5 lines of English text accompanied by a link to an explanatory article by VoxCheck. The length was chosen to increase accessibility to online community members, who might be deterred by lengthy text and follow previous findings of shorter interventions being effective (Martel et al., 2021). Adding a link that allows for more contextual debunking was also found to be more effective by Chan et al. (2017) and Parks and Toth (2006). Further, the link is accompanied by an eye-catching visualization, that reinforces audiences' memory and stands out in the users' online feed (Lewandowsky et al., 2020).

The source exposure statement highlighted a source's lack of credibility and reliability by proving their bad track record of spreading misinformation, biased reporting, or lack of impartiality, whereas the debunking post disproved misinformation claims about Ukrainian refugees from the same outlet. Anti-Ukrainian refuge misinformation was chosen as this topic had garnered much attention in the wake of Russia's invasion of Ukraine in 2022 and both outlets had posted such content within the fringe community. To remain visible within the flow of each group, we planned to post the same statement twice within the span of one week.



Vox Check'er
29 November 2022

The German media Deutschland-Kurier positions itself as an alternative media that conveys "independent and unconventional" opinions silenced by the mainstream media. Is it really so? For more see:
https://voxukraine.org/en/disinformation-in-the-german-media-deutschland-kurier/

WHICH GERMAN MEDIA SHARES FAKE NEWS?

DEUTSCHLAND-KURIER

- promotes Russian disinformation
- shares unverified information about COVID-19 and vaccination against it
- spreads conspiracy about climate change

Vox Check'er
29 November 2022

The German media Deutschland-Kurier positions itself as an alternative media that conveys "independent and unconventional" opinions. But under the guise of "alternativeness", it promotes Russian disinformation narratives. For example, fakes about "support for Nazism" in Ukraine. For more see:
https://voxukraine.org/en/manipulation-nazi-ideas-are-widespread-in-ukraine-deutschland-kurier/

HOW PROPAGANDA LIES

FAKE: Nazi ideas are widespread in Ukraine

TRUTH: Nazi and communist ideologies are officially banned in Ukraine

TRUTH: Far-right groups are very few and far from having support from society or government

Figure 3. Examples of interventions posted by VoxCheck in the Facebook groups. *The left panel shows an example of a source exposure post, and the right panel shows an example of a debunking post.*

Gatekeeper rejection. Following VoxCheck's first post on November 28th of 2022, its profiles were reported and taken down by the administrators of the public Facebook groups. Serving as gatekeepers of posts in the groups, that is deciding which posts to be published or not, ten administrators blocked our statements (hence referred to as Gatekeeper Rejection). While this reality of natural experiments decreased the number of groups, we executed the debunking and source exposure interventions to eight and six groups, respectively. Alternatively, this allowed us to distinguish between exposing fringe communities and their gatekeepers for counter-misinformation strategies.

As this exposure echoes most experimental methods of exposing subjects to corrective treatments in the debunking literature, it can be expected that administrators will emulate similar changes in behavior and be motivated to keep their group members informed. This means that exposing administrators to treatments can be expected to reduce the number of published posts that include content from the targeted misinformation source.

Statistical analysis

As debunking's effect has been found (McCabe & Smith, 2002; Osmundsen et al., 2021; Zerback et al., 2021) to wane after two weeks, we measured consumption levels at three periods: 2 weeks pre-treatment, 2 weeks post-treatment, and 2–4 weeks post-treatment. As group administrators varied in time before accepting our posts, these three periods started from the day treatment posts appeared in every group. For the control groups, this was counted from the first post appearing in all groups.

The initial data analysis of our dependent variable (i.e., consumption) showed it followed an over-dispersed Poisson distribution, forcing us to apply a more conservative negative binomial regression model (see Hilbe, 2011), which is less vulnerable to extreme outliers and accepting deviations from the Poisson model's requirements of variance and mean being equal. We model the post-intervention consumption rate y (both 2 weeks post-treatment and 2–4 weeks post-treatment) as $y \sim \beta_0 + \beta_1 \cdot \mathbb{1}(Treatment) + \beta_2 \cdot \log(Group\ Size) + \beta_3 \cdot CR(2\ weeks\ prior)$ or $y \sim \beta_0 + \beta_1 \cdot \mathbb{1}(Debunking) + \beta_2 \cdot \mathbb{1}(Gatekeeper\ Reject) + \beta_3 \cdot \mathbb{1}(Source\ Exposure) + \beta_4 \cdot \log(Group\ Size) + \beta_5 \cdot CR(2\ weeks\ prior)$, controlling for both the pre-intervention consumption rate and Facebook group size. All outcomes are reported by their corresponding value and the corresponding standard error, denoted as $\beta_i = x(SE = y)$.

Bibliography

Ahlstrom-Vij, K. (2016). Is there a problem with cognitive outsourcing? *Philosophical Issues*, *26*(1), 7–24. https://doi.org/10.1111/phis.12072

Aslett, K., Sanderson, Z., Godel, W., Persily, N., Nagler, J., & Tucker, J. A. (2023). Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, *625*, 548–556.https://doi.org/10.1038/s41586-023-06883-y

Badawy, A., Lerman, K., & Ferrara, E. (2019). Who falls for online political manipulation? In *WWW'19: Companion proceedings of the 2019 world wide web conference* (pp. 162–168). Association for Computing Machinery. https://doi.org/10.1145/3308560.3316494

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication*, *33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

Bode, L., Vraga, E. K., & Tully, M. (2020). Do the right thing: Tone may not affect correction of misinformation on social media. *Harvard Kennedy School (HKS) Misinformation Review, 1*(4). https://doi.org/10.37016/mr-2020-026

Bor, A., & Petersen, M. B. (2021). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, *116*(1), 1–18. https://doi.org/10.1017/s0003055421000885

Browne, K. (2005). Snowball sampling: Using social networks to research non-heterosexual women. *International Journal of Social Research Methodology*, *8*(1), 47–60. https://doi.org/10.1080/1364557032000081663

Bruder, M., Haffke, P., Neave, N., Nouripanah, N., & Imhoff, R. (2013). Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy mentality questionnaire. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00225

Burki, T. (2020). The online anti-vaccine movement in the age of COVID-19. *Lancet Digital Health*, *2*(10), e504–e505. https://doi.org/10.1016/S2589-7500(20)30227-2

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracin, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-73510-5

DiResta, R., & Grossman, S. (2019). *Potemkin pages & personas: Assessing GRU online operations, 2014-2019* [white paper]. Freeman Spogli Institute for International Studies, Stanford University. https://fsi.stanford.edu/publication/potemkin-think-tanks

Donovan, J., Dreyfuss, E., & Friedberg, B. (2022). *Meme wars: The untold story of the online battles upending democracy in America*. Bloomsbury Publishing.

Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., & Adams, K. (2020). Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review*, *40*(3), 560–578. https://doi.org/10.1177/0894439320914853

Guhl, J., Ebner, J., & Rau, J. (2020). *The online ecosystem of the German far-right.* Institute for Strategic Dialogue. https://www.isdglobal.org/wp-content/uploads/2020/02/ISD-The-Online-Ecosystem-of-the-German-Far-Right-English-Draft-11.pdf

Hendricks, V. F., & Hansen, P. G. (2014). *Infostorm: How to take information punches and save democracy*. Springer.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hindman, M., & Barash, V. (2018). *Disinformation, 'fake news' and influence campaigns on Twitter*. Knight Foundation. https://knightfoundation.org/reports/disinformation-fake-news-and-influence-campaigns-on-twitter/

Imhoff, R., & Bruder, M. (2014). Speaking (un-)truth to power: Conspiracy mentality as a generalised political attitude. *European Journal of Personality*, *28*(1), 25–43. https://doi.org/10.1002/per.1930

Johnson, N. F., Velasquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., El Oud, S., Zheng, M., Manrique, P., Wuchty, S., & Lupu, Y. (2020). The online competition between pro- and anti-vaccination views. *Nature*, *582*(7811), 230–233. https://doi.org/10.1038/s41586-020-2281-1

Lewandowsky, S., Cook, J., Ecker, U., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E., Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., Vraga, E., ... Zaragoza, M. S. (2020). *Debunking handbook 2020*. Databrary. https://doi.org/10.17910/b7.1182

Marchal, N. (2021). "Be nice or leave me alone": An intergroup perspective on affective polarization in online political discussions. *Communication Research*, *49*(3), 376–398. https://doi.org/10.1177/00936502211042516

Martel, C., Mosleh, M., & Rand, D. G. (2021). You're definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online. *Media and Communication*, *9*(1), 120–133. https://doi.org/10.17645/mac.v9i1.3519

McCabe, D. P., & Smith, A. D. (2002). The effect of warnings on false memories in young and older adults. *Memory & Cognition*, *30*(7), 1065–1077. https://doi.org/10.3758/bf03194324

McKernan, B., Rossini, P., & Stromer-Galley, J. (2023). Echo chambers, cognitive thinking styles, and mistrust? Examining the roles information sources and information processing play in conspiracist ideation. *International Journal of Communication, 17,* 1102–1125. https://ijoc.org/index.php/ijoc/article/view/19244

Michael, R. B., & Breaux, B. O. (2021). The relationship between political affiliation and beliefs about sources of "fake news". *Cognitive Research: Principles & Implications*, *6*(1). https://doi.org/10.1186/s41235-021-00278-1

Munger, K., Gopal, I., Nagler, J., & Tucker, J. A. (2021). Accessibility and generalizability: Are social media effects moderated by age or digital literacy? *Research & Politics*, *8*(2). https://doi.org/10.1177/20531680211016968

Nassetta, J., & Gross, K. (2020). State media warning labels can counteract the effects of foreign disinformation. *Harvard Kennedy School (HKS) Misinformation Review, 1*(7). https://doi.org/10.37016/mr-2020-45

Nouri, L., Lorenzo-Dus, N., & Watkin, A.-L. (2021). Impacts of radical right groups' movements across social media platforms: A case study of changes to Britain first's visual strategy in its removal from Facebook to gab. *Studies in Conflict & Terrorism*. https://doi.org/10.1080/1057610x.2020.1866737

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization Is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, *115*(3), 999–1015. https://doi.org/10.1017/s0003055421000290

Parks, C. M., & Toth, J. P. (2006). Fluency, familiarity, aging, and the illusion of truth. *Aging, Neuropsychology, & Cogition*, *13*(2), 225–253. https://doi.org/10.1080/138255890968691

Pennycook, G., Bear, A., Collins, E. T., & 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*, *66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/https://doi.org/10.1016/j.cognition.2018.06.011

Petersen, M. B., Osmundsen, M., & Arceneaux, K. (2023). The "need for chaos" and motivations to share hostile political rumors. *American Political Science Review*, *117*(4), 1486–1505. https://doi.org/10.1017/s0003055422001447

Petersen, R. D., & Valdez, A. (2005). Using snowball-based methods in hidden populations to generate a randomized community sample of gang-affiliated adolescents. *Youth Violence and Juvenile Justice*, *3*(2), 151–167. https://doi.org/10.1177/1541204004273316

Pogorelskiy, K., & Shum, M. (2019). News we like to share: How news sharing on social networks influences voting outcomes. The Warwick Economics Research Paper Series (TWERPS) 1199, University of Warwick, Department of Economics. https://ideas.repec.org/p/wrk/warwec/1199.html

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Acadamy of Sciences*, *118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School (HKS) Misinformation Review, 1*(2). https://doi.org/10.37016//mr-2020-008

Rothschild, M. (2021). *The storm is upon us: How QAnon became a movement, cult, and conspiracy theory of everything*. Mellville House.

Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, *67*(2), 233–255. https://doi.org/10.1111/jcom.12284

Skurnik, I., Yoon, C., Park, Denise C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*(4), 713–724. https://doi.org/10.1086/426605

Skurnik, I. W., Park, D. C., & Schwarz, N. (2000). *Repeated warnings about false medical information can make it seem true: A paradoxical age difference.* Eighth Cognitive Aging Conference, Atlanta, GA.

Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 230–239. https://doi.org/10.1609/icwsm.v11i1.14878

Steensen, S. (2018). Journalism's epistemic crisis and its solution: Disinformation, datafication and source criticism. *Journalism*, *20*(1), 185–189. https://doi.org/10.1177/1464884918809271

Trujillo, M., Gruppi, M., Buntain, C., & Horne, B. D. (2020). What is BitChute? In *Proceedings of the 31st ACM Conference on Hypertext and Social Media* (pp. 139–140). Association for Computing Machinery. https://doi.org/10.1145/3372923.3404833

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008. https://doi.org/10.1002/gch2.201600008

Zerback, T., Töpfl, F., & Knöpfle, M. (2020). The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media & Society*, *23*(5), 1080–1098. https://doi.org/10.1177/1461444820908530

Zollo, F. (2019). Dealing with digital misinformation: A polarised context of narratives and tribes. *Special Issue: Proceedings of the Third EFSA Scientific Conference: Science, Food and Society, 17*(S1), e170720. https://doi.org/10.2903/j.efsa.2019.e170720

Acknowledgements
This study owes great thanks to the VoxCheck community, who remained committed to the project despite trials and tribulations of war and whose expertise in debunking misinformation was indispensable.

Funding
No funding has been received to conduct this research.

Competing interests
The authors have no competing interests.

Ethics
This study was approved by the Ethics Board of the European University Institute (20211112_Rasmussen).

Copyright
This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are properly credited.

Data availability
All materials needed to replicate this study are available via the Harvard Dataverse: https://doi.org/10.7910/DVN/ROXL6I

Appendix

Table 1. *Outcomes of the regression model which only considers treatment vs non-treatment. For each group, we look at whether or not this group received an intervention or not, which we denote with the indicator function ($\mathbb{1}$).*

Consumption rate	2 weeks post-treatment	2-4 weeks post-treatment
Constant (β_0)	$0.680\ (SE = 1.046)$	$-2.756^{*}\ (SE = 1.504)$
$\mathbb{1}$(Treatment) (β_1)	$-0.630^{**}\ (SE = 0.301)$	$-0.464\ (SE = 0.391)$
log(Group Size) (β_2)	$0.072\ (SE = 0.124)$	$0.380^{**}\ (SE = 0.173)$
2 weeks pre-treatment (β_3)	$0.131^{***}\ (SE = 0.022)$	$0.159^{***}\ (SE = 0.030)$
Observations	35	35
Log-likelihood	-94.001	-74.786
θ	$2.046^{**}\ (SE = 0.844)$	$1.369^{**}\ (SE = 0.637)$
Akaike Inf. Crit.	196.003	157.572

$^*p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$

Table 2. *Outcomes of the regression model which only considers treatment vs specific treatments. For each group, we look at which intervention this group received.*

Consumption rate	2 weeks post-treatment	2-4 weeks post-treatment
Constant (β_0)	$0.817\ (SE = 1.048)$	$-2.912^{**}\ (SE = 1.482)$
$\mathbb{1}$ (Debunking)(β_1)	$-0.633\ (SE = 0.388)$	$-0.850\ (SE = 0.541)$
$\mathbb{1}$ (Gatekeeper Rejection) (β_2)	$-0.836^{**}\ (SE = 0.390)$	$0.083\ (SE = 0.482)$
$\mathbb{1}$ (Source Exposure) (β_3)	$-0.395\ (SE = 0.412)$	$-1.176^{**}\ (SE = 0.592)$
log(Group Size) (β_4)	$0.060\ (SE = 0.124)$	$0.373^{**}\ (SE = 0.171)$
2 weeks pre-treatment (β_5)	$0.125^{***}\ (SE = 0.023)$	$0.193^{***}\ (SE = 0.033)$
Observations	35	35
Log-likelihood	-93.553	-72.881
θ	$2.117^{**}\ (SE = 0.872)$	$1.401^{**}\ (SE = 0.595)$
Akaike Inf. Crit.	199.106	157.761

$^*p < 0.1$, $^{**}p < 0.05$, and $^{***}p < 0.01$