DE FACTO
Observatory
of Information

SHORT-TERM POLICIES TO FIGHT DISINFORMATION

Authors : Sergei Guriev, Emeric Henry, Théo Marquis, Ekaterina Zhuravskaya

Sciences Po - December 29, 2023 – Final Version

SciencesPo AFP CLEMI Le centre pour l'éducation aux médias et à l'information WIKI

Co-funded by the European Commission under the Connecting Europe Facility (CEF) - Telecommunications Sector (Agreement No. INEA / CEF / ICT / A2020 / 2394372)

Short-term Policies to Fight Disinformation

Sergei Guriev, Emeric Henry, Théo Marquis, Ekaterina Zhuravskaya

With growing evidence of the impacts of fake news on health outcomes, political processes, or hate crimes, fighting disinformation has become a key concern for policy makers. The public debate explores different policy options, including regulations of social media platforms, e.g. the European Union's Digital Services Act (EU, 2023). However, regulatory measures must navigate the delicate balance between curbing misinformation and upholding free speech. In the United States, constitutional limitations impede content moderation regulation, while the European Union's focus is on illegal content, which excludes a significant portion of political misinformation. Another major policy approach involves comprehensive digital literacy programs to empower individuals to discern accurate information from false news (Guess et al., 2020).

In this policy brief, we explore short-term policies that social scientists have proposed to slow down the circulation of misinformation. They are not intended to be substitutes for the aforementioned approaches but should instead be viewed as complements. We describe and discuss the effectiveness of these different policies in Section 1. We then compare them in Section 2 based on a recent paper (Guriev et al., 2023). We then conclude and provide policy recommendations in the last section.

1 Short-term policies

1.a Nudges

A widely advocated approach is to adopt policies intended to shift users' attention towards accuracy. A variety of methods have been proposed, as categorized by Pennycook and Rand (2022). The first method involves asking users to rate the accuracy of different content, with the potential for users to receive feedback on their accuracy ratings. The second method simply warns users that false content circulates on social media. The third method appeals to norms: for instance, participants are told that most other survey respondents think it is very important to only share accurate content. Finally, the fourth method provides tips to detect fake content (akin to digital literacy training).

While these different policies might differ in how easily they can be implemented, they

are generally cheap to apply at scale. Pennycook and Rand (2022) conducted a meta-analysis of over 20 studies measuring the impact of such policies exploiting nudges. They find that such procedures are efficient at reducing the sharing of false news. The average magnitude is a decrease of 10% relative to the control group in these studies. Furthermore, a majority of studies included in this meta-analysis have a significant impact on the sharing of these false headlines. The authors also show that these nudges tend to have a slightly positive effect on the sharing of true news. Arechar et al. (2023) show that these accuracy prompts seem to work regardless of the cultural context; they show the robustness of the results across 16 different countries, even though they highlight some differences.

1.b Fact-Checking

Many social media platforms have started implementing fact-checking of their content. To avoid being accused of censorship, they typically resort to outsourcing the evaluation of veracity, either to outside selected partners or to users themselves. For instance, Facebook (now Meta) set up in 2016 the Third Party Fact-Checking program. In a large panel of countries, they selected a number of partners, fact-checking organisations either part of larger media (such as AFP) or independent NGOs. These partners have freedom in selecting the contents they want to evaluate or the methods they employ. They can, for instance, rely on the algorithm developed by Facebook to detect false information. Once a fact check is produced, the partners have direct access to the platform to flag posts circulating the content. The impact of this program on the circulation of the targeted posts is examined in Cagé et al. (2024).

What is the evidence on the impact of such policies? While there is consistent evidence that fact checking cannot correct the shift in beliefs caused by the initial exposure to fake news (Barrera et al., 2020; Berinsky, 2017), extensive evidence shows that fact-checking reduces the sharing of the rated news. In the French context, Henry et al. (2022) show that exposure to fact checking reduces the sharing of false content by 45%. Moreover, they show that this effect is of a similar magnitude whether users are forced to read the fact-check or can voluntarily access the content. Merely being aware of the existence of a fact-check, without knowledge of its content, is sufficient to decrease the sharing of disinformation. In Section 2, we explain the mechanisms driving this effect.

The findings in Henry et al. (2022) echo the results in other studies. Yaqub et al. (2020) show that putting labels below a news headline—indicating that the news has been fact-checked and shown to be false (even though the actual fact-check is not shown)—decreases the self-reported intention to share (see also Kreps and Kriner (2022)). Pennycook et al. (2020a) carried out an online experiment where the participants were shown true and false statements. They find that adding the "false" label to a statement significantly reduces participants' self-reported intention to share the statement on social media.

Some concerns have been raised that fact-checking and the labeling of news that it implies may have an "implied truth effect," whereby adding tags on specific pieces of content may increase the confidence of users in those that are not flagged. Evidence appears mixed, Pennycook et al. (2020b) show evidence consistent with this effect, while it does not seem to be present in Guriev et al. (2023).

1.c Frictions

Finally, a last type of policy that has been proposed is to increase the cost of sharing content for users by adding an intermediary step before sharing, either by requiring a confirmation click or by requiring a pause before the content is shared. For instance, before the 2020 US election, Twitter modified the default sharing option, prompting users to add a comment to the content they wanted to share. According to Ershov and Morales (2024), this was intended to encourage more thoughtful consideration before sharing.

Henry et al. (2022) show descriptively that a policy requiring an extra click decreases the sharing of false news. Guriev et al. (2023) show that the extra click policy decreases the sharing of false news while leaving the sharing of true news unaffected. Ershov and Morales (2024) show that the Twitter policy mentioned above reduced the overall sharing of news, with left-wing outlets being particularly affected. However, they cannot conclude on the relative effect on sharing true and false news. The evidence on this category of policies thus appears mixed.

2 Comparing the different policies

The literature has evaluated separately these different types of policies. The different studies mentioned use a wide range of methods, making comparisons difficult. For instance, some studies use randomized survey experiments where the outcome of interest is self-reported sharing intention, some use actual sharing behavior, while others use observational data from social media platforms. We conclude this overview by describing a recent paper (Guriev et al., 2023) that uses a unified framework to assess the impact of all the policies considered above on the circulation of both accurate and false news. The study also sheds light on the mechanisms through which these different policies operate.

2.a Methodology

During the 2022 U.S. mid-term election campaign, the authors conducted a randomized controlled experiment involving 3,501 American Twitter users to closely emulate a real sharing experience on the platform. Within a survey environment, participants were presented with

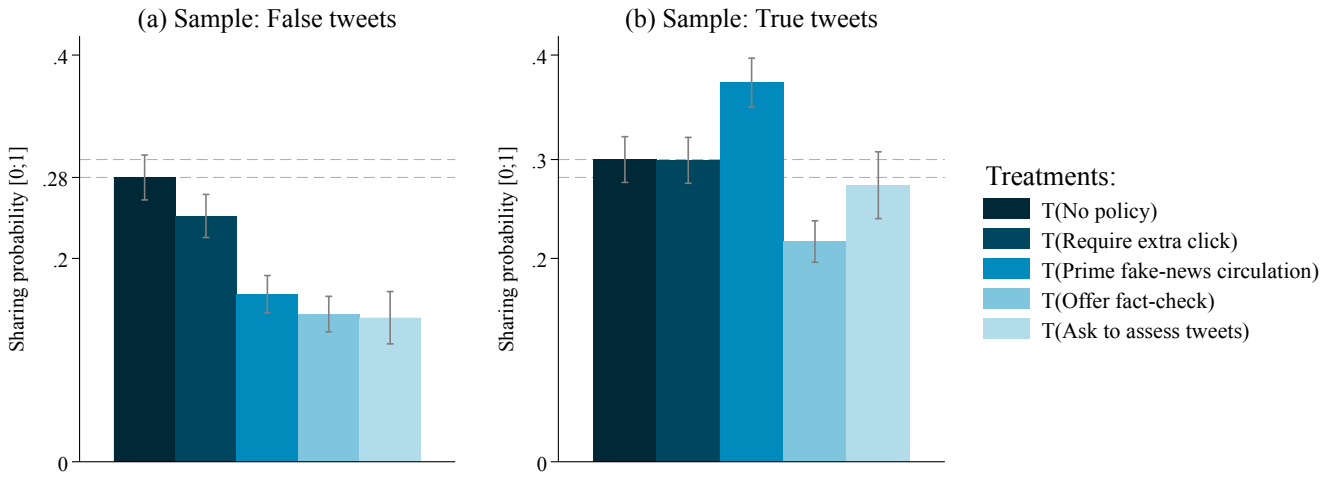
four political information tweets—two containing misinformation and two with accurate facts. The tweets were accompanied by screenshots, and participants were provided with direct links to access the tweets on Twitter. After seeing these tweets, random subgroups of participants underwent treatments simulating policy measures discussed in Section 1. Subsequently, participants were asked if they would share one of these tweets on their Twitter accounts. Those who agreed were directed to Twitter, where they could confirm the retweet of their chosen tweet. The authors also collected information on participants' perceptions of tweet characteristics: they were asked to evaluate each tweet's accuracy and partisan leaning.

There were five treatment groups. Participants in the control group labeled "No policy," proceeded directly to the sharing decision without receiving any treatment. The second group underwent the "Extra click" treatment, introducing an additional confirmation click, as discussed in Section 1.c. The third group was subjected to the "Prime fake news circulation" treatment, where, before sharing, participants received a warning message inspired by nudges discussed in Pennycook and Rand (2022) and analyzed in Section 1.a: "Please think carefully before you retweet. Remember that there is a significant amount of false news circulating on social media." The fourth treatment, "Offer fact check," informed participants that the two false tweets had been fact-checked by PolitiFact.com, a reputable fact-checking NGO. They were provided with a link to access the fact-checking of these tweets, as discussed in Section 1.b. It resembles the voluntary fact check implemented in Henry et al. (2022). In the last treatment, "Ask to assess tweets," participants were prompted to evaluate the accuracy and partisan leaning of the four tweets before sharing, introducing considerations about accuracy and potential partisan biases associated with the content. Everybody else did these evaluations after sharing.

2.b Results

Figure 1 illustrates the reduced-form effects of the treatments on sharing true and false news. In the group with no policy intervention, 28% of participants shared one of the false tweets, while 30% shared one of the true tweets. Consistent with findings in prior research, all treatments led to a reduction in the sharing of false tweets. Specifically, the sharing rates of false news in the (i) extra click treatment, (ii) priming fake news circulation treatment, (iii) offering fact-check treatment, and (iv) the treatment that prompts participants to assess content before sharing were 3.6, 11.5, 13.6, and 14.1 percentage points lower than in the no policy group, respectively.

Figure 1: Average Treatment Effects on Sharing for False and True Tweets



Source: Guriev et al., 2023.

However, the treatments have very different impacts on the sharing of true tweets. Requiring an extra click and prompting participants to assess tweets before sharing show no discernible effect. Offering a fact-check decreases the sharing of true tweets by 7.8 percentage points compared to the no policy group. In contrast, the priming treatment boosts the average sharing rate of true tweets by 8.1 percentage points. These findings establish a clear hierarchy of policy effectiveness in enhancing the accuracy of shared political content. The priming fake news circulation treatment emerges as the most potent strategy.

2.c Mechanisms

To understand the underlying mechanisms behind the differential effects of treatments on the sharing of true and false news, the authors construct and structurally estimate a model of sharing political information on social media. In this model, the sender evaluates the costs and benefits of sharing. The costs encompass the effort involved in sharing, such as the number of clicks required or the mental energy spent processing fact-checking information. The benefits of sharing are driven by three distinct motives: political persuasion of content's audience, signaling own partisan affiliation, and the benefits of maintaining a reputation as a credible and trustworthy source.

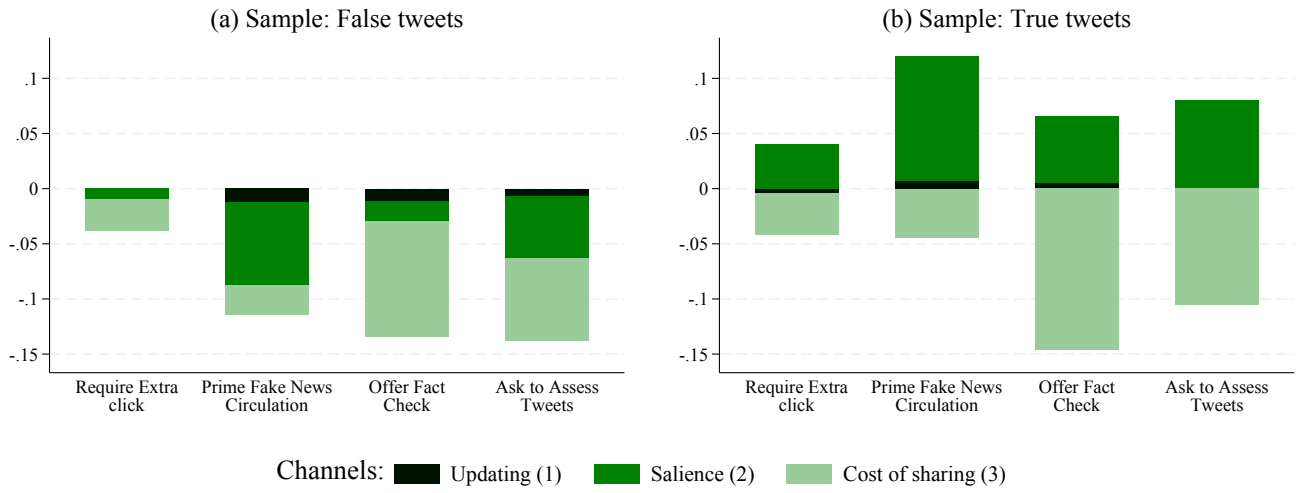
The paper shows that the decision to share a specific piece of news increases with the perceived veracity and the perceived partisan alignment of the news. Veracity positively influences the utility of sharing because of the reputation motive—well-informed receivers are more likely to view the sender as credible. Both partisan motives—persuasion and signaling—indicate that the sender gains a higher payoff when sharing news aligned with their views. Sharing decision also depends on the interaction between veracity and alignment. The direction of this effect depends on which partisan motive predominates. Persuasion is stronger when the news is more

likely to be true, as receivers are inclined to adjust their actions accordingly. Conversely, signaling partisanship is stronger when the news is likely to be false, as a likely-false partisan message conveys a more credible signal of partisanship than a true partisan message. The results of the structural estimation of the model using experimental data show that the reputation motive is a crucial driver of sharing, while partisan motives also play a role. Moreover, partisan persuasion dominates signaling.

This analysis explains the mechanisms through which the different policy interventions considered in Section 1 affect sharing. Overall, anti-misinformation policies can influence sharing through three channels: (1) Updating, (2) Salience, and (3) Cost of sharing.The Updating channel operates through treatments that potentially prompt the sender to revise their beliefs about the content's veracity and partisan alignment. For example, fact-checking aims to alter users' perceptions of content accuracy. The Salience channel works through treatments altering the relative salience of reputation concerns compared to the partisan motives – so that the participants put a larger weight on the veracity of the news when deciding whether to share. The nudges, for instance, are designed to affect salience. Finally, each treatment affects the cost of sharing. For instance, adding an extra confirmation click increases the sharing cost in terms of the number of clicks for all types of content, whether true or false.

Figure 2 presents the decomposition of the effects of various treatments into these three channels. Surprisingly, the Updating channel contributes minimally to the impact of treatments, despite some treatments affecting the sender's estimates of news veracity and partisan alignment. Instead, the overall effect of each treatment comes from the combination of how they influence the salience of reputation and the cost of sharing. Specifically, the salience channel drives the difference in treatment effects on sharing false and true news. Raising the salience of reputation positively impacts the sharing of true news and adversely affects the sharing of false news. To varying degrees, all treatments increase the salience of reputation, with priming fake news circulation having the most substantial effect. Simultaneously, the costs of friction associated with different treatments reduce the sharing of both true and false news. Notably, the additional costs in the priming treatment are considerably lower than in the fact-checking treatment, making priming more effective in increasing the sharing of true news. This analysis explains why the priming treatment emerges as the most cost-effective policy.

Figure 2: Decomposition of the treatment effects into the three channels



Source: Guriev et al., 2023.

3 Conclusion

In this policy brief, we have discussed the effectiveness of a number of short-term policies intended to slow down the circulation of disinformation. It is important to evaluate their impact on both the circulation of true and false news. Policies based on nudges, intended to shift the users' attention towards accuracy, appear to be the most effective: they decrease the sharing of false news while increasing the sharing of true content. Furthermore, the other policies considered, such as fact-checking, also turn out to operate mainly through a salience effect: presenting fact-checks prompts individuals to think about accuracy, regardless of whether they see the content of the fact-check.

The following policy requirements emerge:

1. Encourage platforms to implement nudges prompting users to consider accuracy. The exact format of these nudges would have to be determined and could be changed over time to avoid habituation.

2. Encourage methods that facilitate speedy fact-checking, such as algorithmic methods, even at the cost of increased error rates since the main effect of fact-checking is a salience effect.

We further note that short-term policies such as accuracy prompts and fact-checking should be seen as complements rather than substitutes for other more long-term methods such as digital literacy training.

References

Arechar, Antonio A., Jennifer Allen, Adam J. Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G. Lu, Robert M. Ross, Michael N. Stagnaro, Yunhao Zhang, Gordon Pennycook, and David G Rand, "Understanding and combatting misinformation across 16 countries on six continents.," *Nature Human Behaviour*, Jun 2023.

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya, "Facts, alternative facts, and fact checking in times of post-truth politics," *Journal of Public Economics*, 2020, *182*, 104–123.

Berinsky, Adam J, "Rumors and health care reform: Experiments in political misinformation," *British journal of political science*, 2017, *47* (2), 241–262.

Cagé, Julia, Moritz Hengel, Émeric Henry, and Nathan Gallo, "Fact-Checking and Misinformation. Evidence from the Market Leaderh," *Working Paper*, 2024.

Ershov, Daniel and Juan S Morales, "Sharing News Left and Right: Frictions and Misinformation on Twittera," Technical Report 2024.

EU, "Digital Services Act: Application of the Risk Management Framework to Russian disinformation campaigns," Technical Report 2023. Permanent link: https://op.europa.eu/en/publication-detail/-/publication/c1d645d0-42f5-11ee-a8b8-01aa75ed71a1.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar, "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India.," *Proceedings of the National Academy of Sciences of the United States of America*, Jun 2020.

Guriev, Sergei, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya, "Curtailing False News, Amplifying Truth," *CEPR Discussion Paper 18650. CEPR Press*, 2023.

Henry, Emeric, Ekaterina Zhuravskaya, and Sergei Guriev, "Checking and Sharing Alt-Facts," *American Economic Journal: Economic Policy*, August 2022, *14* (3), 55–86.

Kreps, Sarah E and Douglas L Kriner, "The COVID-19 infodemic and the efficacy of interventions intended to reduce misinformation," *Public Opinion Quarterly*, 2022, *86* (1), 162–175.

Pennycook, Gordon, Adam Bear, Evan Collins, and David Rand, "The Implied Truth Effect: Attaching Warnings to a Subset of FakeNews Headlines Increases Perceived Accuracy of Headlines Without Warnings," *Management Science*, 2020, *66*, 4921–5484.

_ **and David Rand**, "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation," *Nature Communications*, 2022, *13* (2333).

_ **, Jonathon Mcphetres, Yunhao Zhang, and David Rand**, "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention.," *Psychological Science*, 2020, *31*, 770–780.

Yaqub, Waheeb, Otari Kakhidze, Morgan Brockman, Nasir Memon, and Sameer Patil, "Effects of Credibility Indicators on Social Media News Sharing Intent," *CHI '20:*

Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020.