Self-Censorship Appears to be an Effective Way of Reducing the Spread of Misinformation on Social Media

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Howe, Piers
Perfors, Andrew
Ransom, Keith James
et al.

Publication Date
2023

Peer reviewed

Self-Censorship Appears to be an Effective Way of Reducing the Spread of Misinformation on Social Media

Piers D. L. Howe[1,a], Andrew Perfors[1], Keith J. Ransom[1], Bradley Walker[2,3], Nicolas Fay[1], Yoshihisa Kashima[1], Morgan Saletta[4]

[1]School of Psychological Sciences, University of Melbourne; [a]pdhowe@unimelb.edu.au; [2]School of Psychological Science, University of Western Australia; [3]School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University; [4]Hunt Laboratory, University of Melbourne.

Abstract

There is increasing pressure on social media companies to reduce the spread of misinformation on their platforms. However, they would prefer not to be the arbiters of truth as the truth can be subjective or otherwise hard to determine. Instead, they would prefer that social media users themselves show better discernment when deciding which information to share. Here we show that allowing people to share only those social media posts that they have indicated are true significantly improves sharing discernment, as measured by the difference in the probability of sharing true information versus the probability of sharing false information. Because it doesn't require social media companies to be the arbiters of truth, this self-censorship intervention can be employed in situations where social media companies suspect that individuals are propagating misinformation but are not sufficiently confident in their suspicions to directly censor the individuals involved. As such, self-censorship can usefully supplement externally imposed (i.e. traditional) censorship in reducing the propagation of false information on social media platforms.

Keywords: censorship; fake news; accuracy prompt; truth; misinformation; echo chambers

Introduction

The propagation of false and misleading information on social media has been thrust into the public spotlight by a series of high-profile events, including the Capitol Hill Riots and the UK's Brexit referendum, where highly misleading and sometimes entirely fabricated news stories were widely circulated on social media (Pennycook et al., 2021; Pennycook & Rand, 2021). As a result of these and other occurrences, there are increasing calls for social media companies to do more to reduce the spread of misinformation on their platforms (Nix, 2022), and a majority of U.S. adults believe that tech companies should take such steps to reduce the spread of misinformation, even if this would result in losing some freedom to access and publish content (Mitchell & Walker, 2021). However, social media companies are reluctant to censor information on their platforms (Zakrzewski, Lima, & Harwell, 2023) and have long argued that they can't be the arbiters of truth (Borchers, 2018; Mosseri, 2016), in part because the truth might be subjective (Leetaru, 2019) and disputed (Tripodi, 2022). Instead, they argue that social media users themselves should discern truth from falsehood and voluntarily refrain from spreading false information online.

A number of approaches have been developed to encourage users to be more discerning when sharing information on social media (Kozyreva, Lewandowsky, & Hertwig, 2020; van der Linden et al., 2021). Of particular relevance to the present study is a novel approach that employs accuracy prompts to encourage users to consider whether the information is accurate before deciding whether to share it. This approach capitalises on the fact that a large majority of social media users report that it is important to share only accurate information (Pennycook et al., 2021). Pennycook *et al.* suggested that users share false information despite this belief because they do not consider accuracy when making the decision to share. This suggests that reminders to consider the accuracy of the information before deciding whether to share should increase sharing discernment.

Despite some impressive initial findings (Pennycook et al., 2021), the effectiveness of accuracy prompts has been disputed (Rathje, Roozenbeek, Traberg, Bavel, & Linden, 2022; Roozenbeek, Freeman, & Linden, 2022). In part to address these concerns, a meta-analysis was conducted (Pennycook & Rand, 2022). Defining sharing discernment as the difference in probability for sharing accurate and inaccurate information, Pennycook and Rand found that accuracy prompts reliably improved sharing discernment, but the effectiveness of this intervention was relatively small and varied across studies.

One reason why accuracy prompts may not be particularly effective is that some users appear to share news posts even when they don't believe they are accurate (Study 6 of Pennycook & Rand, 2021). This suggests that accuracy is not the only factor to influence sharing decisions. This makes some sense: people might also share in order to signal their identity, to mock or laugh at something, or because it would be interesting if it were true. Nevertheless, the act of sharing might still signal to the receiver that the shared item is true, so still spread misinformation.

Here we investigate a new possible intervention, which we call "self-censorship", in which users are allowed to share only those posts that they have indicated are true. This is *self*-censorship rather than externally imposed censorship because it is the user themself that decides which posts they are allowed to share. To be clear, there is nothing preventing the user from lying and labeling a post they believe is false as true so that they are allowed to share it. Nevertheless, we posit that most users will not lie, so this intervention will be more effective than accuracy prompts alone.

Study 1: Evaluating the Relative Effectiveness of Accuracy Prompts and Self-Censorship

The purpose of our first study was to determine how effective self-censorship is, relative to accuracy prompts, for in-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

creasing sharing discernment. There were three conditions: a Baseline condition, an Accuracy Prompt condition and a Self-Censorship condition. In the Baseline condition people decided whether to share a series of social media posts, one at a time, and there were no accuracy prompts nor any self-censorship. For the Accuracy Prompt condition, participants assessed the accuracy of each news post immediately before deciding whether to share it. Pennycook and Rand (2021) showed that this type of accuracy prompt is effective.

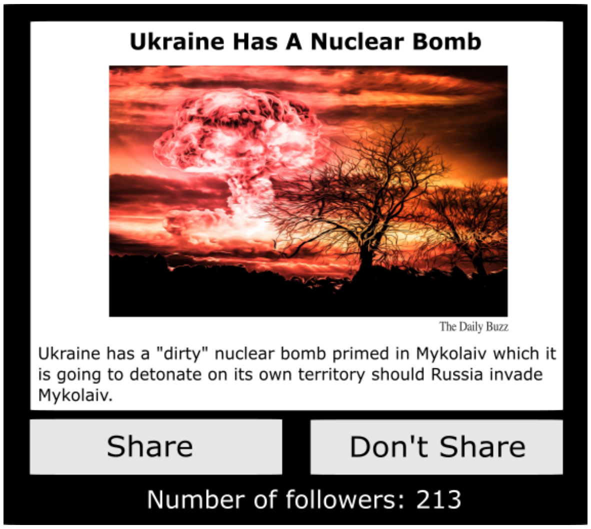
In the Self-Censorship condition, participants were informed in advance that they would not be able to share any news posts that they indicated were false. They then rated the accuracy of each news post. Only if they indicated that the post was true were they then asked whether they wished to share it. To be clear, there was nothing to prevent participants from lying and indicating that a news post was true simply because they wished to share it but we predicted that most participants would not do this because, as discussed above, we had no reason to suspect that our participants would be dishonest. As such, we expected sharing discernment to be greater in the Self-Censorship condition than in the Accuracy Prompt condition. We also predicted that sharing discernment would be greater in both these conditions than in the Baseline condition.

Study 1 had three aims: to determine whether sharing discernment was greater in the Accuracy Prompt condition than in the Baseline condition, to determine whether it was greater in the Self-Censorship condition than in the Baseline condition, and to determine whether it was greater in the Self-Censorship condition than in the Accuracy Prompt condition.

Method

Participants Participants were recruited from the USA via Mechanical Turk. The study was only open to the subset of MTurkers who had previously passed a test for English proficiency that also screened out bots. We therefore excluded only those participants who didn't finish the experiment. The study took approximately 5 minutes to complete, and participants were compensated US$1 for their time. 149 participants completed the experiment. Each was randomly assigned to one of the three conditions (Baseline: 46 participants, Accuracy Prompt: 53 participants, Self-Censorship: 50 participants). 93 participants self-identified as male, 53 as female, two as non-binary and one self-identified as "other". The mean age was 39.8 years old (sd = 11.5 years). 81 participants self-identified as Democrats, 28 as Republicans, 36 as independents and four as "other". All participants gave informed consent, and the study was approved by the University of Melbourne Human Ethics Advisory Group (ID: 23317).

Materials and Procedure Both this and the subsequent study were run using Psytoolkit (Stoet, 2010, 2017). Participants first completed a brief survey asking their age, gender and the political party with which they most strongly identified. They were then randomly assigned to one of the three conditions. In the Baseline condition, they were informed that they would be participating in a simulated social media experiment and that their task was to gain as many followers as possible. They were instructed that they could gain new followers by sharing social media news posts that people wanted to see. They were instructed that each news post would contain a photograph and that the source of each news post would be indicated directly under the photograph. Finally, they were informed that *Now8News*, *The Daily Buzz*, and the *World News Daily Report* often publish news that is not true whereas *The New York Times*, the *BBC*, and *The Wall Street Journal* almost never publish news that is not true. The first three (i.e. the unreliable) news sources were chosen as they often publish false news and have no consistent political bias. The latter three (i.e. the reliable) news sources were chosen as they are reputable, have neutral news coverage and represent a range of political opinions, with left-leaning, central and right-leaning editorial coverage respectively.



Ukraine Has A Nuclear Bomb

The Daily Buzz

Ukraine has a "dirty" nuclear bomb primed in Mykolaiv which it is going to detonate on its own territory should Russia invade Mykolaiv.

Figure 1: An example of a (false) social media news post shown to participants in the Baseline condition. For copyright reasons, the image used in the post has been substituted by a copyright-free image by George Hodan (https://www.publicdomainpictures.net/en/view-image.php?image=198316&picture=nuclear-bomb-explosion).

The instructions for the Accuracy Prompt and Self-Censorship conditions were identical to those for the Baseline condition except that participants were also told that before deciding whether they would share a news post they would first need to indicate whether it was true. The instructions for the Self-Censorship condition further specified that participants would only be able to share those news posts that they had indicated were true.

After reading these instructions, participants were then quizzed on their understanding of what they had read, and any misunderstandings were corrected. They were then shown a series of 20 simulated social media news posts in a random order. Ten of the news posts made entirely false claims while the remainder made entirely true claims. These news posts were inspired by similar new posts found on reputable news sites (*The New York Times*, the *BBC*, and *The Wall Street Journal*), sites known to propagate fake news (*Now8News*, *The Daily Buzz*, and the *World News Daily Report*), and a fact-checking website (*Snopes.com*). All the news posts were modified to be in the same format which comprised a heading, an image, a lede sentence and a source attribution, which appeared directly below the image. An example of a false news post is shown in Figure 1.

In the BASELINE condition, participants needed to decide whether to share each news post. Every post they shared, regardless of whether it was true or false, gained them a random number of new followers, where this number was drawn from a normal distribution with a mean of 100 and a standard deviation of 20. The ACCURACY PROMPT condition was identical to the BASELINE condition except that each post initially had two buttons beneath it ("True" and "False"). Once the participant had indicated whether the post was true or false, the buttons were replaced with "Share" and "Don't Share", and the participant then needed to indicate whether they wished to share the post. The SELF-CENSORSHIP condition was identical to the ACCURACY PROMPT condition except that the "Share" and "Don't Share" buttons appeared only if the participant indicated that the post was true. After the participant had viewed the 20 news posts, the experiment finished, they were debriefed and invited to leave any comments they had. The stimuli, raw data and analysis code are available at OSF: https://osf.io/zm4rb/

Results The data in both experiments in this paper were analysed in R (R Core Team, 2022) using the lme4 software package (Bates, Mächler, Bolker, & Walker, 2015). A mixed effects model was used to predict each participant's response to each news post. The dependent variable was either whether the participant shared the news post or whether they thought it was true. Although both dependent variables were binary, we used a linear model because our goal was to obtain unbiased estimates of the causal effects of our predictor variables, rather than maximising the predictive power of our model (Gomila, 2021). The predictors were `veracity`, `condition` and an interaction between the two, with random intercepts for `participant` and `news post`. The variable `veracity` denotes whether the news post was actually true or not regardless of how it was perceived. Adopting the terminology of Pennycook and Rand (2021), we defined sharing discernment as the difference in probability for sharing true versus false news posts. Thus, if the interaction between `veracity` and `condition` was significant, this would show that sharing discernment varied across conditions.

The data is summarised in Figure 2. Subplot A shows the mean `share rate`, defined as the proportion of participants who chose to share each news post, averaged over the actually false (light blue) and actually true (gold) news posts separately. The `share rate` for false news posts was significantly less than the `share rate` for true news posts ($\chi^2(1,149) = 30.6$, $p < .001$) and varied as a function of `condition` ($\chi^2(2,149) = 16.3$, $p < .001$). Sharing discernment varied across the conditions ($\chi^2(2,149) = 21.2$, $p < .001$). Sharing discernment was not significantly different between the BASELINE and ACCURACY PROMPT conditions ($\chi^2(1,99) = 1.88$, $p = .17$) but was significantly different between the ACCURACY PROMPT and SELF-CENSORSHIP conditions ($\chi^2(1,103) = 10.6$, $p = .001$) and the BASELINE and SELF-CENSORSHIP conditions ($\chi^2(1,96) = 19.8$, $p < .001$). Considering only participants who self-identified as either Democrats or Republicans, the later result was not significantly affected by party affiliation ($\chi^2(1,73) = 0.018$, $p = .89$), with sharing discernment being significantly greater in the SELF-CENSORSHIP condition than in the BASELINE condition when Republicans ($\chi^2(1,18) = 4.4$, $p = .04$) and Democrats ($\chi^2(1,55) = 16.0$, $p < .001$) were considered separately.

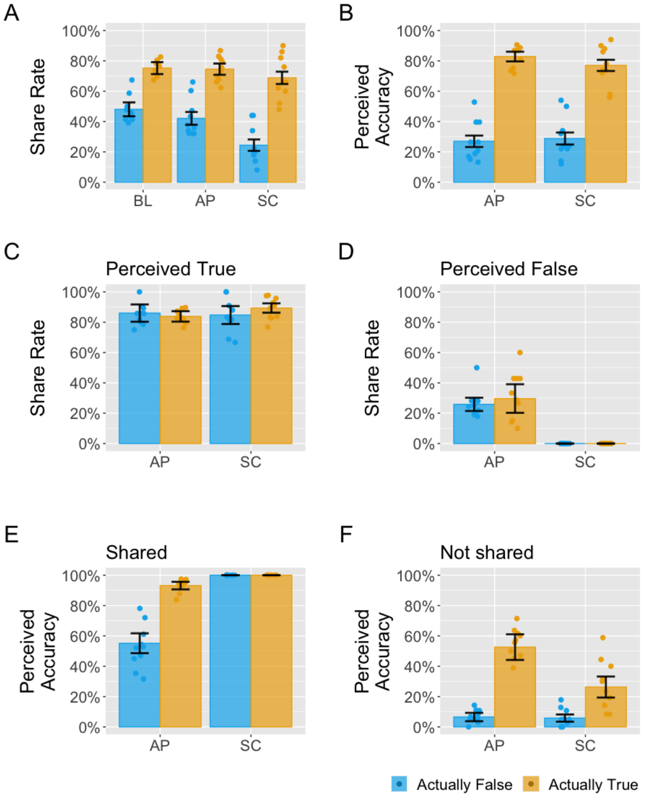


Figure 2: The data from the first study. Subplot A shows the `share rate`, defined as the proportion of participants who chose to share each news post, for the BASELINE (BL), ACCURACY PROMPT (AP), and SELF-CENSORSHIP (SC) conditions for news posts that are actually false (light blue) or actually true (gold). Subplot B shows the `perceived accuracy`, defined as the proportion of participants who indicated that each news post was true, for the AP and SC conditions. Subplots C and D show the `share rate` for the news posts that were perceived to be true and false respectively. Subplots E and F show the `perceived accuracy` for the news posts that were shared and not shared respectively. In all subplots, error bars represent 95% CI and dots represent ratings for individual news posts.

Subplot B shows the `perceived accuracy`, defined as the proportion of participants who indicated that each news post was true averaged over the actually false (light blue)

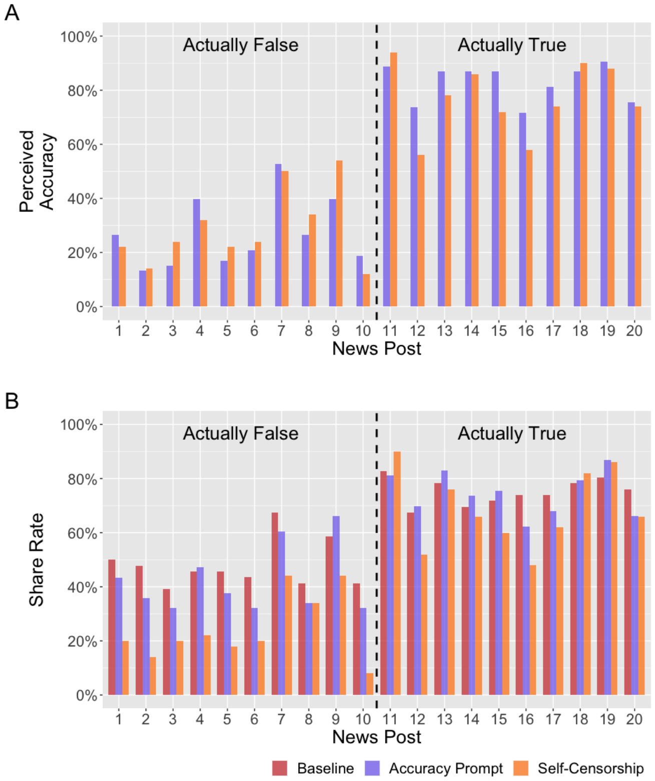


Figure 3: The data from the first study for the individual news posts for the Baseline (BL), Accuracy Prompt (AP), and Self-Censorship (SC) conditions. News posts 1-10 are actually false whereas news posts 11-20 are actually true. Subplot A shows the `perceived accuracy` and Subplot B shows the `share rate`.

and actually true (gold) news posts separately. `Perceived accuracy` was significantly higher for true news posts than for false news posts ($\chi^2(1,103) = 33.7$, $p < .001$), but did not vary by condition ($\chi^2(2,103) = 3.82$, $p = .051$). Additionally, there was no evidence that participants were more likely to indicate that the false news posts were true in the Self-Censorship condition than in the Accuracy Prompt condition ($\chi^2(1,103) = 0.14$, $p = .71$). Subplot C shows the `share rate` for the news posts that were perceived to be true. It shows that if a news post was perceived to be true, it was highly likely to be shared. Conversely, subplot D shows that if a news post was perceived to be false it was less likely to be shared in the Accuracy Prompt condition and, by design, could not be shared in the Self-Censorship condition. For completeness, subplots E and F show the `perceived accuracy` for the news posts that were shared and were not shared, respectively. Subplot E confirms that in the Accuracy Prompt condition participants shared news posts that they believed were false. By design, this could not occur in the Self-Censorship condition.

Figure 3 shows the `perceived accuracy` (Subplot A) and `share rate` (Subplot B) for the individual news posts. Subplot A shows that, for each news post, the `perceived accuracy` was similar for the Accuracy Prompt and Self-Censorship conditions. Comparing Subplot A to Subplot B we see that for news posts that were actually false, for the Accuracy Prompt condition, the `share rate` was always higher than the `perceived accuracy`, which demonstrates that people share news posts that they perceive to be false. Conversely, for the Self-Censorship condition, the `share rate` was always less than the `perceived accuracy` because people were prevented from sharing news posts that they believed were false.

In contrast to previous findings (Pennycook & Rand, 2022), our data shows that sharing discernment was not significantly greater in the Accuracy Prompt condition than in the Baseline condition. However, sharing discernment was significantly greater in the Self-Censorship condition than in either the Baseline condition or the Accuracy Prompt condition. The reason for this is that in the Accuracy Prompt condition participants shared news posts that they had identified as false, whereas they were prevented from doing this in the Self-Censorship condition. These results demonstrate that self-censorship increases sharing discernment more than accuracy prompts.

Study 2: Relative Preference of Accuracy Prompts Versus Self-Censorship

Study 1 shows that self-censorship may be an effective way of increasing sharing discernment beyond what can be achieved by accuracy prompts. But would people accept self-censorship in lieu of accuracy prompts? Study 2 addressed this question.

Method

Participants As before, the study was only open to the subset of MTurkers who had previously passed a test for English proficiency that also screened out bots. MTurkers who had participated in the previous experiment were not allowed to participate in this experiment, and we excluded only those participants who didn't finish the experiment. The study took approximately 5 minutes to complete, and participants were compensated US$1 for their time. Of 100 participants, 56 self-identified as male and the remainder as female. The mean age was 41.5 years old (sd = 11.1 years). 49 participants self-identified as Democrats, 17 as Republicans, 31 as independents and three as "other". All participants gave informed consent, and the study was approved by the University of Melbourne Ethics Advisory Group (ID: 23317).

Materials and Procedure The study started in the same manner as the previous one with participants being asked to state their age, gender, and political affiliation. As before, it was then explained that they would participate in a simulated social media study where their task was to gain as many followers as possible by sharing news posts that people wished to see. It was explained which news sources are unreliable and which ones are generally considered to be reliable. Participants were told that they would need to indicate whether each news post was true or not before deciding whether they wished to share it. They were told that there were two conditions and that they could choose which condi-

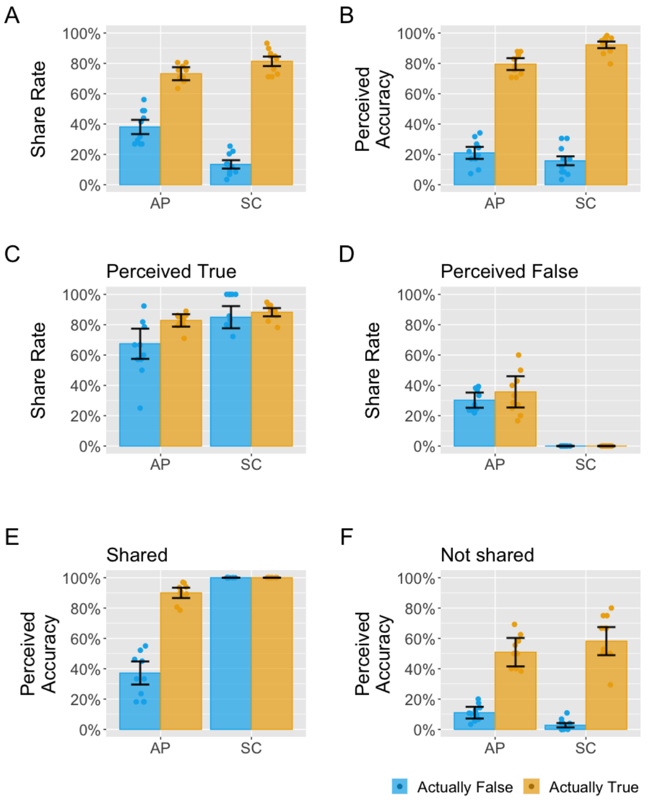


Figure 4: The data from the second study. Subplot A shows the `share rate` for the Accuracy Prompt (AP) and Self-Censorship (SC) conditions for news posts that are actually false (light blue) or actually true (gold). Subplot B shows the `perceived accuracy` of the news posts. Subplots C and D show the `share rate` for the news posts that were perceived to be true and false respectively. Subplots E and F show the `perceived accuracy` for the news posts that were shared and not shared respectively. In all subplots, error bars represent 95% CI and dots represent ratings for individual news posts.

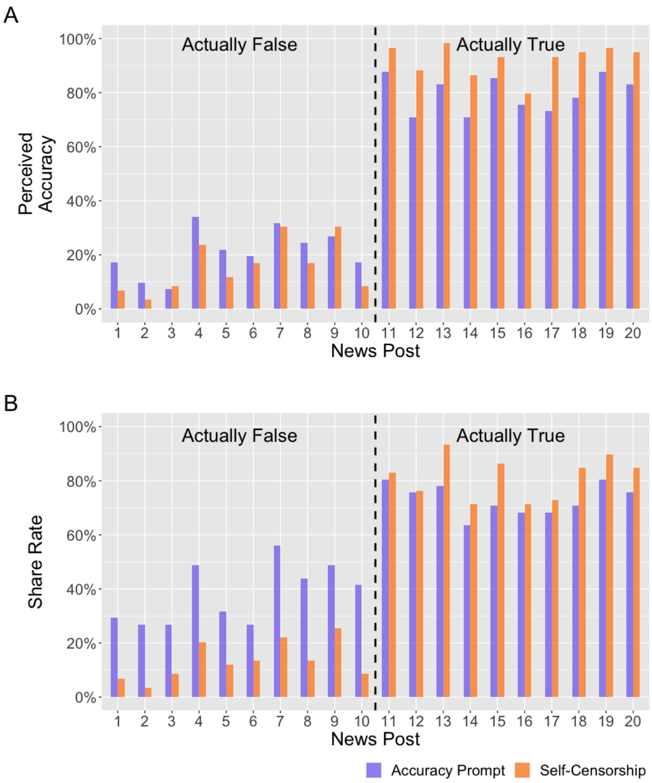


Figure 5: The data from the second study for the individual news posts for the Accuracy Prompt (AP) and Self-Censorship (SC) conditions. As before, news posts 1-10 are actually false whereas news posts 11-20 are actually true. Subplot A shows the `perceived accuracy` and Subplot B shows the `share rate`.

tion they wished to participate in. They were told that in one condition they would be able to share any news post regardless of whether they had indicated whether it was true (i.e. the Accuracy Prompt condition). Conversely, they were told that in the other condition they would only be able to share those news posts that they had indicated were true (i.e. the Self-Censorship condition). To avoid biasing them, participants were not informed of the names of the two conditions. Participants were informed that they were less likely to receive untrue news posts in the second condition than in the first, to reflect the fact that if self-censorship was introduced on a social media platform then participants would be less likely to receive misinformation, as implied by the results of Experiment 1. Participants then completed an understanding check, and any misunderstandings were corrected. They then chose their preferred condition. Despite what participants were told, both conditions used the same news posts as in the previous experiment.

Results Out of the 100 participants, 59% choose the Self-Censorship condition. A binomial test found that this result was not significantly different from 50% ($p = .09$), suggesting no strong preference for either condition. Figure 4 shows the data collapsed across news posts. As shown by Subplot A, sharing discernment was again higher in the Self-Censorship condition than in the Accuracy Prompt condition ($\chi^2(1,100) = 21.2$, $p < .001$). As shown by Subplot B, participants were less likely to indicate that the false news posts were true in the Self-Censorship condition than in the Accuracy Prompt condition ($\chi^2(1,100) = 17.0$, $p < .001$). As shown by Subplots D and E, this was because participants in the Accuracy Prompt condition would share news posts that they had labelled as false whereas this could not occur in the Self-Censorship condition. Figure 5 shows the data for the individual news posts. For every false news post, the `share rate` was greater than the `perceived accuracy` in the Accuracy Prompt condition but not in the Self-Censorship condition.

General Discussion

This work suggests that self-censorship is an effective way of increasing sharing discernment. We also found people are not against using it. Indeed, there was no evidence that people preferred accuracy prompts to self-censorship, implying that they perceived that the disadvantage of self-censorship (i.e. being restricted in which posts they could share) was compensated for by its advantage (i.e. receiving fewer news posts that were false). That said, participants indicated their preference *before* experiencing the condition. Future work will need to demonstrate whether these results can be replicated in a more realistic setting and whether people would be willing to continue to accept self-censorship when they have experienced it. As it stands, self-censorship represents a promising but unproven approach for reducing the sharing of misinformation on social media.

Our results are consistent with other findings indicating that accuracy prompts may not be an effective way of increasing sharing discernment, such as the recent meta-analysis finding that the size of the effect of accuracy prompts on sharing discernment was relatively small and varied significantly across experiments (Pennycook & Rand, 2022). This meta-analysis reported that the average effect size, as measured by the coefficient on the interaction between `condition` and `veracity`, was 0.04. Considering only the BASELINE and the ACCURACY PROMPT conditions from Experiment 1, we found an interaction coefficient of 0.05. For comparison, we found an interaction coefficient of 0.17 when considering only the BASELINE and the SELF-CENSORSHIP conditions. This suggests that accuracy prompts have a reliable but smaller effect on sharing discernment than self-censorship.

We had expected that at least some participants in the SELF-CENSORSHIP condition would lie and indicate that some of the news posts that they thought were false were true so that they would then be allowed to share them. Surprisingly, there was no evidence that this occurred: the proportion of false news posts rated as true was not higher in the SELF-CENSORSHIP condition than in the ACCURACY PROMPT condition in either Experiment 1 or Experiment 2. Despite being entirely honest, participants did not always prioritise accuracy when deciding whether to share a post: in Experiments 1 and 2, 21% and 18% respectively of the posts that were shared in the ACCURACY PROMPT condition had been labelled as 'false' by the participant. These results show that while the participants were not willing to lie and indicate that a post they think is false is true just so that they could share it, they were willing to share posts they had labelled as 'false'. This may explain why sharing discernment was greater in the SELF-CENSORSHIP condition than in the ACCURACY PROMPT condition: participants were prevented from sharing posts they had labelled as 'false' in the former condition but could do so in the later condition.

To be clear, we don't believe that social media users should *always* be required to indicate that they believe a social media post is true before being allowed to share it. Unlike in this study, in real life, not all social media posts can be categorized as either 'true' or 'false'. For instance, some don't make any factual statements. Additionally, most posts will focus on non-contentious topics and some may even be satirical. It would not be sensible to insist that social media users categorize such posts as 'true' before being allowed to share them. Rather, we are advocating that social media posts are analysed by a natural language model (NLM) and that self-censorship is employed only when the NLM believes that a factual statement has been made, that the claim is not satirical and that it is about a sufficiently important topic (e.g. an upcoming election) that self-censorship is warranted. Furthermore, if the social media user does not wish to share the post, there is no need for them to indicate whether it is true or not, thereby further reducing the load on the user.

We acknowledge that our proposed intervention requires a NLM to identify to which posts self-censorship should be applied. However, we don't believe that this requirement is unduly onerous for three reasons. First, the NLM does not need to be 100% accurate. It wouldn't matter if it occasionally overestimated the importance of a topic and applied self-censorship to a topic that wasn't sufficiently important to warrant self-censorship as the cost to the human user of applying self-censorship is minimal. Second, in situations where the NLM was unsure as to whether self-censorship should be applied it could err on the side of caution and not impose self-censorship. Applying self-censorship for only some of the posts where it is warranted would still be a substantial improvement over the status quo. Finally, NLM's have advanced tremendously in recent months (e.g. ChatGPT). Even if NLM's are not yet capable of doing what we require, it is likely that they may soon be able to.

In conclusion, subject to further research, we believe that self-censorship is a potentially viable alternative to externally imposed censorship. While externally imposed censorship may still be needed, we believe that self-censorship can be employed in situations where social media companies are not sufficiently confident in their suspicions to censor the individuals involved. Specifically, self-censorship can be applied whenever the social media company suspects a post might be making an untrue factual statement about a contentious and important topic. In this way, we believe that self-censorship has the potential to supplement externally imposed censorship and thereby play a significant role in reducing the spread of misinformation in online settings. Crucially, it can do this without requiring social media companies to be the arbiters of truth and in a user-centered manner that is unlikely to induce psychological reactance (Brehm, 1966), which might otherwise encourage people to adopt the behaviour opposite to what the social media company desires (i.e. the social media user may purposely share information that they believe is false because they believe that the social media company is attempting to induce them not to). Given the increasing pressure on social media companies to restrict the flow of misinformation on their platforms as embodied by recent initiatives such at the European Commission 2022 Code of Practice on Disinformation (European Commission, 2022), we expect that social media companies will find self-censorship a useful tool to help them meet their obligations, at least in some circumstances.

Acknowledgments

The research was supported by an Office of National Intelligence (ONI) and Australian Research Council (ARC) grant (NI210100224), and the Western Australian Government (Defence Science Centre).

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Borchers, C. (2018, February 8). Twitter executive on fake news: 'we are not the arbiters of truth'. *The Washington Post*.

Brehm, J. W. (1966). *A theory of psychological reactance.* Academic Press.

European Commission. (2022). *The 2022 code of practice on disinformation.* https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation (Jan 24, 2023).

Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *JEP: General*, *150*(4), 700–709.

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: confronting digital challenges with cognitive tools. *Psychol. Sci. Public Interest*, *21*, 103–156.

Leetaru, K. (2019, June 25). Is there such a thing as objective truth in data or is it all in the eye of the beholder? *Forbes*.

Mitchell, A., & Walker, M. (2021). *More americans now say government should take steps to restrict false information online than in 2018.* Retrieved from https://www.pewresearch.org/fact-tank/2021/08/18/more-americans-now-say-government-should-take-steps-to-restrict-false-information-online-than-in-2018/ (Jan 23, 2023).

Mosseri, A. (2016). *Addressing hoaxes and fake news.* Retrieved from https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/ (Jan 23, 2023).

Nix, N. (2022, October 27). Big tech is failing to fight election lies, civil rights groups charge. *The Washington Post*.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590–595.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends In Cog. Sci*, *25*(5), 388–402.

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Comm.*, *13*(2333), 1–12.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rathje, S., Roozenbeek, J., Traberg, C. S., Bavel, J. J. V., & Linden, S. (2022). *Letter to the editors of Psychological Science: meta-analysis reveal that accuracy nudges have little to no effect for U.S. conservatives: regarding Pennycook et al. (2020).* Retrieved from https://journals.sagepub.com/page/pss/letters-to-the-eds (Jan 23, 2023).

Roozenbeek, J., Freeman, A. F., & Linden, S. (2022). How accurate are accuracy nudge interventions? A pre-registered direct replication of pennycook et al. (2020). *Psychol. Sci.*, *32*(7), 1169–1178.

Stoet, G. (2010). Psytoolkit - a software package for programming psychological experiments using linux. *Behav. Res. Methods*, *42*(4), 1096–1104.

Stoet, G. (2017). Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psych.*, *44*(1), 24–31.

Tripodi, F. B. (2022). *The propagandists' playbooook: How conservative elites manipulate search and threaten democracy*. New Haven, CT: Yale University Press.

van der Linden, S., Roozenbeek, J., Maertens, R., Basol, M., Kácha, O., Rathje, S., & Traberg, C. S. (2021). How can psychological science help counter the spread of fake news? *Span. J. Psychol.*, *24*, 1–9.

Zakrzewski, C., Lima, C., & Harwell, D. (2023, January 17). What the Jan. 6 probe found out about social media, but didn't report. *The Washington Post*.