Reducing Falsehoods at the Source: An experiment incentivizing Brazilian political elites to avoid online misinformation

Paul Atwell, Fernando B. Mello, and Simon Chauchard[a]

[a]UC3M - Departamento de CCSS; Instituto Juan March

ARTICLE HISTORY
Compiled November 12, 2024

ABSTRACT
Strategies to counter misinformation have, so far, mostly focused on demand-side approaches, encouraging and equipping media users to avoid it, detect it, fact-check it or report it (Ecker et al. 2022, Jerit and Zhao 2020). Yet, since elites play a central role in the dissemination of misinformation (Mosleh and Rand 2022), many governments have begun to enact supply-side approaches assessing legal penalties to individuals that share misinformation. Brazil's new regulations, introduced after the storming of the presidential palace in 2023, illustrate this trend. Since 2024, a key government body systematically monitors the online behavior of political candidates, and energetically prosecutes offenders. In this survey-experiment, we explore whether informing candidates for office of these penalties (n=873) is effective in mitigating the flow of misinformation they endorse. We go beyond this and additionally test whether other messaging approaches can have the same effects. We show that informing candidates of penalties causes significant improvement in their accuracy discernment: they are able to identify and avoid misinformation at greater rates when they are informed of potential costs. However, we also find that exposure to these countermeasures reduces the overall amounts of content posted. Finally, we show gentler nudges may be equally as effective as raising penalties. These findings have broad implications for how governments and online platforms seek to durably address misinformation: strong supply-side approaches may reduce it, but do so at the cost of a general chilling effect on social media engagement.

KEYWORDS
Misinformation, Compliance, Elections, Campaign strategy, Brazil

1. Introduction

Following a wave of misinformation around the 2022 federal election in Brazil, thousands of protesters attempted to overthrow the elected government of one of the world's largest democracies, ransacking both the supreme court and presidential palace (Frenkel 2023). Since then, Brazil and several other countries have broadened the legal

Corresponding author: Paul Atwell (p.atwell@uc3m.es). The authors thank Gilberto Kassab and Alexandre Gajardoni, from PSD, and all respondents for their collaboration. This project was supported by the European Research Council. A pre-analysis plan was pre-registered at OSF. The project was approved by the Committee on Ethics in Research of Universidad Carlos III de Madrid. Names are listed in alphabetical order.

definition of what constitutes misinformation - raising penalties, increasing prosecutions and pressuring platforms do reduce its impact on politics (Bateman and Jackson 2024). This study focuses on the behavior of key producers and propagators of political (mis)information in the context of heightened social costs and legal penalties. Can increasing the costs of misinformation reduce political elites' role in the broader propagation of falsehoods and help safe-guard democracies?

Though avenues to make voters resistant to misinformation have been thoroughly explored (Jerit and Zhao 2020), we move "upstream" from information consumers to candidates for public office. Despite their importance in online ecosystems, existing research on how to discourage elites from engaging in information is plainly limited (Blair et al. 2024). In one example, Nyhan and Reifler (2015) advise sitting U.S. Representatives that their posting will be fact-checked by independent websites and find a reduction in the amount of falsehoods shared. Yet, Ma et al. (2023) provide a recent replication of the former with state legislators and find no such effect. Thus evidence is not only inconclusive, but is focused on the American context, using a narrow (reputational) intervention to motivate behavioral change.

We argue that much like regular consumers of misinformation, at least one class of misinformation propagators, political elites, can be incentivized to avoid false stories. Re-incentivizing elites requires offsetting the combination of (1) perceived benefits of posting misinformation, and (2) their directional biases toward posting partisan congenial content. Many policy responses are beginning to focus on overwhelming the former with practical and predictable costs, which should have the effect of activating effortful thinking, sidelining the effect of biases in the latter. Informing elites of these costs, or alternatively making them salient again, should favor cost-benefit considerations of their online behavior. Under normal circumstances, this would result in the production of less misinformation, greater efforts to discern false from true, and active stances against misinformation.

To evaluate this, we conduct a survey experiment in Brazil in the context of the 2024 municipal elections. Collaborating with the Partido Social Democrático (PSD), we circulated a survey experiment to active candidates for municipal council in several states and received 873 viable responses. As a party, PSD ranges from ideologically diverse to ideologically void (Zucco and Power 2024), and has a strong foothold in municipalities large and small, setting wide external validity. As a country, Brazil is both an easy and hard test case. The violence of the 2022 election has been widely attributed to high levels of misinformation, meaning there is an ample supply to be reduced. At the same time, since 2022, prosecution rates have climbed sharply and been matched with much steeper penalties. This has made the topic highly salient, reaching its peak with the banning of Twitter from Brazil in August 2024.

Our design randomly assigns respondents a control condition, or one of four treatment conditions that highlight the risks of posting misinformation as a candidate. Two focus on actual *penalties*, while two offer informational or social *context* around misinformation. We then observe how respondents perform on a discernment task, reporting how likely they would be to share each headline from a pool of true and false stories that had recently circulated in Brazil. We also invite candidates to sign a real-world open letter hosted by an anti-misinformation NGO. Finally, we introduce richer outcomes including the amount of time respondents spend evaluating each headline, their knowledge of misinformation rules, and issue fatigue around the topic.

We find that treatment was effective in reducing reported likelihood of sharing misinformation by over 11%. Although it also appears to have had a slight chilling effect on posting over all– relative to the control group, treated respondents showed less interest in true headlines as well. This additional accuracy does not seem to have required greater effort, we find no evidence treated respondents spent more time deliberating. Observationally, we note knowledge levels are low across treatment and control, indicating that nuanced understanding of penalties may not necessary to achieving normatively better behavior. This dovetails with our finding that the context treatments were just as effective as those that focus on penalties in reducing interest in false headlines. Finally, we do no find that treated respondents were more likely to take a stance against misinformation, being no more likely to sign the open letter.

We make several contributions. Firstly, we show that upstream or "supply-side" approaches to endemic misinformation appear to have traction in our sample of local candidates for public office. Evidence on combating misinformation is heavily tilted toward the Global North and towards information consumers. We not only focus on the Global South, but also on political elites. This is just the third example of an experimental intervention targeting politicians to our knowledge. Our study stands in contrast to the null results of one (Ma et al. 2023), and successfully replicates Nyhan and Reifler (2015), showing notable decreases in interest in misinformation. This may be the result of the type of intervention. Whereas Ma et al. (2023) note their media-based fact checking of state-level legislators may have lacked teeth, the costs our respondents confront are assessed not by the voting public, but an active electoral court.

Additionally, we introduce an important corollary; at equilibrium, our results suggest that pricing-in the costs of misinformation would make candidates more cautious about posting any content, constraining discourse. While on balance this may still be a net positive, this depends on contextual or unknown counterfactual costs of misinformation on a social level. Therein, it raises the thorny question of how much accurate (and potentially banal) *information* can be sacrificed in the pursuit of reducing *misinformation.* It additionally provides strong evidence that platforms' engagement-based business models are at odds with their obligations to punish offenders (Cusumano, Gawer and Yoffie 2021). Yet, considering the very low revealed knowledge levels around misinformation rules *despite* the exceptional salience of the aggressive changes by the electoral court on the issue, it is possible that this chilling would not occur if rules are made clear. Further research is needed to understand if properly educating elites can empower them to make the most of online platforms.

Secondly, our comparison of context-focused interventions (including reminding respondents there are new rules, and social norms stand against dishonesty) with penalty-focused approaches (noting exceptional prosecution rates and the ease with which access to online accounts can be revoked) shows that in a context with firm misinformation penalties already in existence, reliance on similarly firm messaging may not be necessary to draw elites away from engaging in misinformation. This instead leads us to consider that the main effect of all interventions was to raise the salience of the broader discourse around misinformation in Brazilian society and how to manage it.

Thirdly, we show that political elites can arrive at higher levels of accuracy (versus engagement with misinformation) without additional deliberation. Whereas the literature on accuracy nudges has either not considered effort or implied that accuracy requires effort and cognitive resources, we find that better performance was achieved with less effort (time). We argue this indicates that effortful deliberation may be already occurring in the control group, as respondents continue to consider posting potential misinformation. In contrast, the intervention seems to have made the choice obvious– do not share false stories.

Finally, results do not appear to be conditional on the kind of intervention message, the type of headline content, education levels, or gender of the respondent. The main results hold in each sub-sample and disaggregation of treatment and outcomes. Nevertheless, we observe some difference in the magnitude of effects that point to potential avenues for future research. Namely, the potential role of education with ultimately receiving information about policy changes and how to move accuracy on non-political misinformation. Taken together with our sample belonging to a large and diverse centrist party, this study suggests regulating misinformation at the source (political elites) is widely effective and likely a valuable tool in shoring up democracy in the 21st Century.

2. Countering Misinformation: the Potential Role of Supply-side Measures

Misinformation in online spaces is widely recognized as a critical challenge for modern democracies (Hochschild and Einstein 2015, Persily, Tucker and Tucker 2020), driving not just political polarization, but also playing a role in fostering intergroup violence and undermining faith in fundamental institutions (Blair et al. 2024). We take misinformation to be purportedly factual claims that are either demonstrably false, widely contradicted by existing evidence and expert opinion, or false and as of yet unverifiable (Vraga and Bode 2020). While many who post and consume rumors and misinformation are genuinely unable to discern its misleading or false nature, many others are able but fail to act according to their best appraisal of facts (Bullock et al. 2015). Thus we consider individuals who, given the correct incentives, would accurately identify and avoid misinformation to also be engaging in it.

Most research on this phenomenon has focused squarely on "demand side" factors as a means of reducing its influence. Studies have experimentally evaluated attempts to reduce information consumers' propensity to believe or share misinformation, through informational, educational, or sociopsychological interventions (Blair et al. 2024, Jerit and Zhao 2020). This includes successful nudges and education efforts to build media literacy and resilience to attempts to mis- or disinform (e.g. Lewandowsky and Van Der Linden (2021), Guess et al. (2020), and Pereira et al. (2023)). Others have succeeded in warning consumers that information may be false or misleading (Prike, Butler and Ecker 2024), which stands in contrast to evidence highlighting the exceptional resistance of consumers to fact-checking (Batista Pereira et al. 2022).[1]

At the same time, the creation, propagation, and global prevalence of misinformation has been empirically linked to social and political elites. On the level of social norms and roles, voters are shaped by what they observe from elites (Lenz 2012, Pérez 2015, Zaller 1992), as well as look to them to model positions and attitudes (Druckman 2001).

[1] We acknowledge that the boundary between misinformation consumers and producers not easily defined and that the category of consumers will often include political elites. Nevertheless, it is an empirical reality that content creation (including misinformation) is not evenly distributed across users (Mosleh and Rand 2022).

More specific to the issue of misinformation, the online behavior, e.g. sharing, reaction, posting etc., of political and social elites directly shapes the content their followers see (Lasser et al. 2022, Mosleh and Rand 2022). In empirical observations of real online networks, the propagation of misinformation has been shown to emanate from a stable core of central spreaders (Shao et al. 2018, Zhang, Chen and Lukito 2023). More cynically, this population is often the original source of rumors and misinformation (Hameleers 2020), including intentional campaigns to misinform (Mosleh and Rand 2022).

It follows that as policymakers and others seek to find scale-able solutions to misinformation, moving upstream from misinformation consumers to address primary and secondary sharers of information should be effective in reducing the global supply of misinformation. Building on Berinsky's (2017) finding that partisan source effects can help undo misinformation, turning these actors into advocates may have further positive effects. Moreover, as Blair et al. (2024) notes, these interventions "...which arguably hold the greatest promise because they aim to reduce the supply of misinformation, have been studied the least, with no studies on politician messaging or journalist training in Global South countries to date" (p. 2). Targeting political elites may be a more direct, if not necessary, approach to getting ahead of misinformation.

We describe efforts to change elites' relationship to misinformation as "upstream" interventions. While some studies observationally document elites' relationship to misinformation (Mosleh and Rand 2022), to our knowledge only two exist that have experimentally tested an approach to reducing the supply of information among these actors. Both Nyhan and Reifler (2015) and Ma et al. (2023) study the impact of warning active U.S. legislators that their posting will be fact checked, working with samples of national and state representatives, respectively. While the earlier Nyhan et al.'s study finds a reduction in fact-checked claims, the latter finds no such effects. Thus, despite providing some key initial indications showing that addressing the supply of misinformation may be a viable solution, the base of evidence is both limited in scope and in the type of intervention, as well as in their geographical bounds (the United States). In this paper, we expand on these by introducing a new type of intervention in a younger democracy among local political elites.

What prevents political elites from using their influence to mitigate the prevalence of misinformation?[2] Firstly, like any other individual on social media, they may lack the factual background to identify misinformation (Pasek, Sood and Krosnick 2015).[3] In all other conditions, we assume that with effortful reflection, accurate discernment is possible (Pennycook and Rand 2019, 2022), but the individual either lacks incentive and/or attention to do so, or is more incentivized to favor information congruent to their worldview (Flynn, Nyhan and Reifler 2017, Jerit and Zhao 2020).

This imbalance of incentives and disincentives occurs for many reasons. Misinformation can garner votes and politicians may consciously choose to engage in it (Bräuninger and Marinov 2022, Farhall et al. 2019, Grossman and Helpman 2023). Others will face an informational deficit around the costs of posting misinformation. For example, reminding individuals that sharing misinformation is a deviation from social norms and carries social costs, can increase rates of reporting it to platforms (Gimpel et al.

[2] While a relative outlier, in the case of Brazil, the steep penalties, ease of reporting someone, and prosecution rates assume that in no circumstances is it rational to engage in misinformation.

[3] Despite typically counting on higher levels of formal education, political elites are likely to lack the attention and factual priors to effectively identify misinformation in many cases.

2021, Prike, Butler and Ecker 2024). Likewise, ignorance of the legal and regulatory environment may lead to sub-optimal behavior, and communicating such penalties can lead to less engagement with misinformation (Yildirim et al. 2023).

Many will be aware of the theoretical costs of misinformation, but this information will either lack salience or credibility as they consciously or unconsciously consider the decision to engage in misinformation. Online environments are rife with visual and auditory stimuli (Bayer, Trieu and Ellison 2020), and the associated cognitive load may suppress the salience of information that would otherwise facilitate more optimal decision making (Bordalo, Gennaioli and Shleifer 2022). Finally, salient priors about rules and penalties may be accurate, but be perceived to be low-probability events, leading again to lower compliance (Castro and Scartascini 2015, Slemrod, Blumenthal and Christian 2001).

In this study, we consider four approaches to raising perceived costs, credibility of those perceived costs, and the salience thereof. Specifically, we alternately highlight the existence of changes in rules around misinformation, point to social norms against dishonesty, note that an account suspension from Meta may cripple a campaign, and detail the frequency with which posters are reported and prosecuted. In practice, we either pool all treatment groups, or bundle them as cues on the *context* of posting misinformation versus *penalties* that may result from posting misinformation. We they should collectively improve behavior through (1) causing a rational aversion to false or risky posting, (2) increasing effort to discern headline accuracy before sharing, and (3) increased salience and recall of existing knowledge around misinformation and its regulation (and therefore the expected costs).

We also note misinformation interventions often have perverse effects (see Nyhan (2021) or Jerit and Zhao (2020)). On the one hand, if beliefs about the costs of misinformation rise too high, this could have a *chilling effect* on all types of posting–truncating online posting across the board. On the other, highlighting the regulation (and costs) of risky content may overwrite or weaken accurate priors or intuitions about rules, potentially causing more violations. Conditional on prior exposure to anti-misinformation campaigns, additional attempts to shift beliefs may also further issue-fatigue, reducing receptivity to such messages and willingness to comply with rules (Gurr and Metag 2021, Morrison, Parton and Hine 2018). Finally, we consider that likelihood of responding to any misinformation intervention is strongly conditioned on expected returns to posting or not posting in a later period. If an candidate is cynical about their electoral chances, this should shift the balance of incentives toward apathy around following regulations and moderate treatment effects.

Brazil and Misinformation

Many democracies are grappling with misinformation, but Brazil is notable due to the scale of the problem and the scale of its approaches to combating it (Margolis and Muggah 2023). The exceptional rates of social media and messenger app use in Brazil have provided ample infrastructure for a misinformation crisis to emerge.[4] Among candidates for municipal office, official data show nearly all have multiple social media

[4] 66% of its population of 212 million is on social media and over 96% uses WhatsApp. Source: https://datareportal.com/reports/digital-2024-brazil and https://www.statista.com/topics/7731/whatsapp-in-brazil/

channels.[5]

Following the contentious 2022 elections, which marked a peak in political polarization (Samuels and Belarmino 2024), misinformation reached critical levels, culminating in protesters attempting to initiate a coup, ransacking the Presidential Palace, the Supreme Court, and other government buildings in the process. This violent episode set the stage for immediate efforts by authorities to durably curb the spread of false information, particularly in the realm of electoral politics.

A major prong of the response was led by the Electoral Supreme Court (hereafter TSE), with Justice Alexandre de Moraes at the forefront of the campaign against misinformation. Moraes set the stakes, calling misinformation the "plague of the 21st century," and arguing social media usage and exposure to false information are an existential threat to democracies.[6] In March 2024, the TSE introduced a raft of regulations to minimize misinformation in the upcoming municipal elections (Resolution No. 23,732/2024).[7] This included stricter penalties for spreading misinformation, as well as clear obligations for platforms to address and report on the issue if they are to do business in Brazil.[8]

The robustness of the response, the characterization of Moraes as a dictator, and generally high levels of political polarization made the issue a cultural flash point and one of the most salient political themes of 2024 (perhaps culminating in the banning of Twitter due to non-compliance with the new rules). Meta, the parent company of Facebook, Whatsapp, and Instagram, has independently launched PSA campaigns against misinformation, both on their networks and in public spaces. A number of NGOs have been formed around the issue, and many serve as rapid fact-checking services. Leaders like `Aos Fatos` and `Lupa` are known by most Brazilians online. Despite this wide salience of the issue and attempts by government bodies to communicate new rules, as well as channels to report violations, knowledge of the particulars of the regulations remained limited, even among candidates for office. In a separate survey of over 500 incumbent mayors in May 2024, we found almost 60% failed to explicitly identify a basic criterion for when sharing AI content is legal.[9]

Thus, while the need to fight misinformation is clear, it has remained unclear whether the legal tools available will have an effect and in what way that may be. It is also unobvious whether the simple knowledge that these tools exist, or their grounding in social norms against dishonesty and misinformation, might be enough on their own to shift elites' behavior.

[5] The TSE now requires candidates to submit the URLs of all their social media accounts when registering their campaign, and we found nearly every mayor complied at least partially. In a prior 2024 survey of incumbent mayors from across Brazil (N = 504), we found most of those accounts are on Instagram followed next by Facebook.

[6] https://www.tse.jus.br/comunicacao/noticias/2023/Setembro/em-evento-no-stf-moraes-afirma-que-noticias-falsas-201csao-a-praga-do-seculo-21201d

[7] Source: https://idrc.sun.ac.za/judicial-activism-or-democratic-safeguard-the-new-limits-of-digital-electoral-propaganda-in-brazil/

[8] Other institutions joined the fight against misinformation. One of President Luiz Inácio (Lula) da Silva's first actions after returning to office in 2023 was to establish the National Prosecutor's Office for the Defense of Democracy. By assigning the Attorney General's Office (Advocacia Geral da União - AGU) to oversee this new institution, the government signaled its strong resolve to tackle fake news. Source: https://agenciabrasil.ebc.com.br/en/politica/noticia/2023-03/attorney-general-vows-redouble-efforts-against-fake-news-brazil

[9] Specifically, they were asked to evaluate the accuracy of "In this years election, it is legal to post AI content if you label it as such."

Brazil is a good case to test the effects of misinformation campaigns using political parties. We study members of the Social Democratic Party (Partido Social Democrático, hereafter PSD), a significant player in Brazil's electoral politics. Positioned as a centrist party, the PSD has demonstrated flexibility by aligning with both left-wing administrations, such as the Workers' Party (PT) governments, and right-wing administrations, like that of Jair Bolsonaro. Although the party matured in just São Paulo state, it achieved a notable milestone in 2024 by electing the highest number of mayors across the entire country, making it the biggest winner in the municipal elections. Unlike more conservative parties, the PSD has not been the origin of scandals involving misinformation and has remained largely outside the debate surrounding misinformation versus freedom of expression, setting it apart from more right-wing counterparts.[10]

All in all, the 2024 election is in many ways a vital test of a supply-side approach to solving misinformation, for two reasons. First, misinformation is a pervasive social issue deeply intertwined with politics, influencing public opnion and political behavior. Second, the issue is a high priority for the country's political elite and is actively addressed in policy discussions. Despite these efforts, no systematic evaluations have been conducted to assess the impact of anti-misinformation measures, leaving a gap in understanding the effectiveness of these strategies.

Hypotheses

We focus this study on testing several key relationships described in Table 1. The first class of outcomes pertains to respondents' ability and willingness to identify misinformation from a set of true and false headlines (hypotheses 1a, 1b, and 1c). On the one hand, we expect that raising the salience and expected magnitude of the costs of misinformation will induce respondents to eschew false stories at greater rates (1a and 1b). On the other hand, we expect as a perverse side-effect a chilling in overall posting, made most evident in a reduction in openness to true headlines (1c). [11]

Next, we consider specifically the effort that respondents exert in examining news (1d and 1e). As indexed by the *average time* spent evaluating news (where more time is associated with greater effort), we predict treated respondents will be more wary of engaging with misinformation and exert more effort identifying and avoiding misinformation.[12] However, we also believe that a small component of this deliberation will correspond to the decision of whether to share misinformation, once identified. Hence we predict treated respondents will spend more time with false headlines relative to true headlines as they weigh the now more salient costs of doing so (1e).

Turning to other domains, we predict that treated respondents will be more likely to send costly signals of a stance against misinformation (2a) in the form of signing an open letter against misinformation. We also expect them to demonstrate higher levels of knowledge around misinformation rules (2b), not because of information contained in a treatment, but through the increased salience of penalties generating more effort to avoid penalties. With an eye toward how repeated messaging on a discursive theme

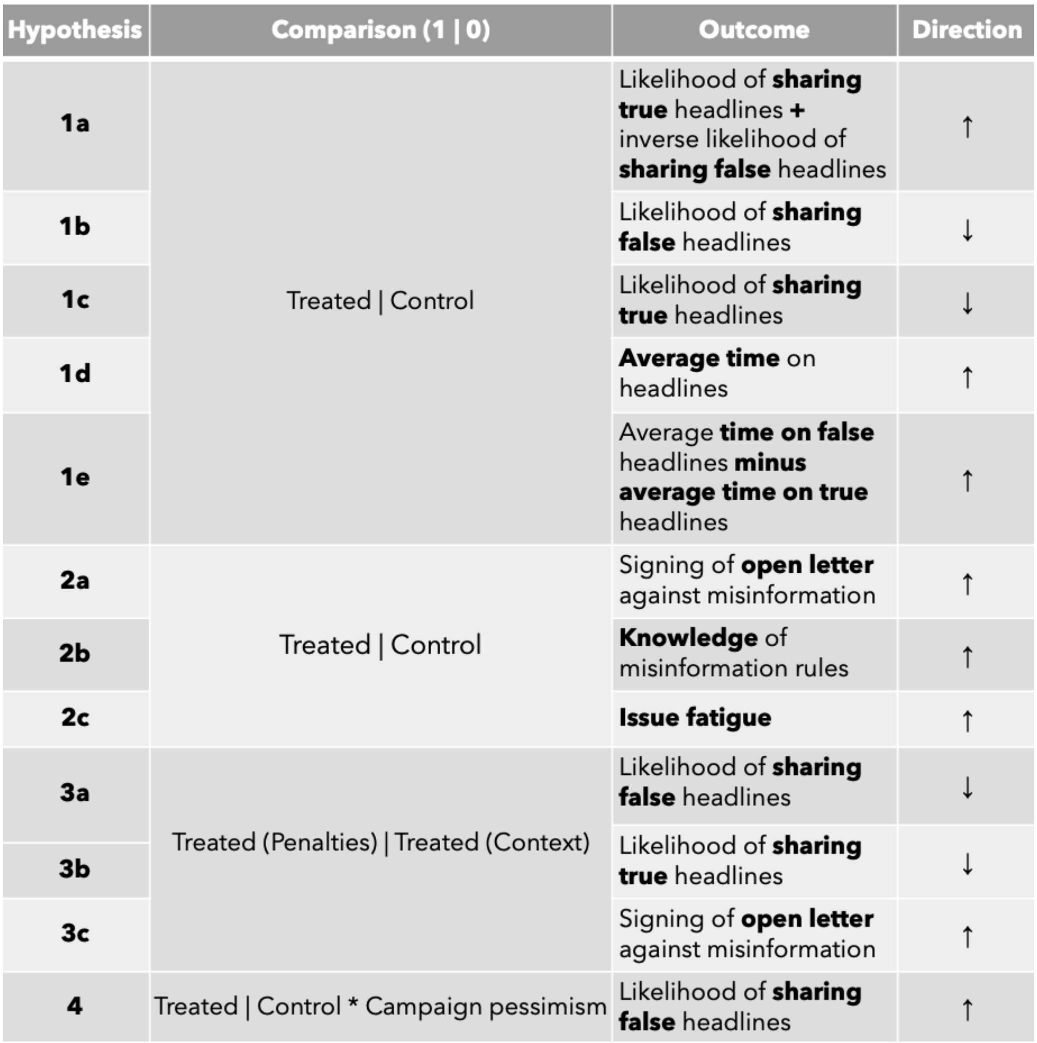
[10] It was, however, recently the subject of one that saw misinformation attacks originating from Bolsonaristas against its leader, Gilberto Kassab. `Source.`

[11] Notably, as the intervention will not carry any information that would help assess headline veracity, distinguishing effects on **1b** from **1c** would suggest that information on the costs of misinformation sharing may be as important in mitigating it as factual priors about a given headline.

[12] Time is an imperfect proxy for deliberation and effort. However, Alós-Ferrer and Buckenmaier (2021) provide experimental evidence that "deliberation times can be fruitfully used" as an index of cognitive effort (pp. 559)

can not just erode the impact of an information campaign, but actually cause a potential backlash, we predict treatment will cause greater rates of issue fatigue around misinformation (2c).

To explore the behavioral value of penalizing misinformation and put a finer edge on hypotheses 1b, 1c, and 2a, we predict that respondents that have been discouraged from posting misinformation in terms that highlight concrete costs (i.e. legal cases and loss of campaign media) will have a greater impact than other approaches when compared to a control group.



Hypothesis	Comparison (1 | 0)	Outcome	Direction
1a	Treated | Control	Likelihood of **sharing true** headlines **+** inverse likelihood of **sharing false** headlines	↑
1b		Likelihood of **sharing false** headlines	↓
1c		Likelihood of **sharing true** headlines	↓
1d		**Average time** on headlines	↑
1e		Average **time on false** headlines **minus average time on true** headlines	↑
2a	Treated | Control	Signing of **open letter** against misinformation	↑
2b		**Knowledge** of misinformation rules	↑
2c		**Issue fatigue**	↑
3a	Treated (Penalties) | Treated (Context)	Likelihood of **sharing false** headlines	↓
3b		Likelihood of **sharing true** headlines	↓
3c		Signing of **open letter** against misinformation	↑
4	Treated | Control * Campaign pessimism	Likelihood of **sharing false** headlines	↑

Figure 1. Pre-registered Hypotheses. We also pre-registered four additional hypotheses, which we present and analyze in Appendices D and I

Finally, we explore two additional relationships. In addition, we consider the impact of cynicism as a moderator in elites' relationship to misinformation. In particular, we predict that respondents who are more optimistic about their electoral chances will be more responsive to treatment.

3. A causal test of when elites comply with misinformation rules

Using a survey experiment we attempt to causally identify the impact of a short message on the costs of misinformation on political elites' behavior and attitudes around misinformation. We do so with a sample of active candidates for câmara municipal (municipal council) in several Brazilian states. Our main outcomes include on a 10-trial headline discernment task (Pennycook and Rand 2022) and an opportunity to sign an open letter against misinformation.

Sample

We collaborate with a major political party (PSD) to provide specific information on the risks of misinformation to its candidates for municipal office during the 2024 national election. The PSD is a centrist and relatively young party, but has a large electoral footprint, drawing support from both the Left and the Right. Notably, despite having national presence, this party has its main roots in São Paulo state and this is where it runs the most candidates and has most of its success.

We study this informational effort among current candidates for municipal council (vereador), running under the banner of the PSD. Because the strength of the intervention is arguably weak, we targeted specifically candidates for municipal council as a large pool of respondents from which we could study the impacts of single interventions. To give a sense of the scale, PSD ran 411 candidates for mayor in São Paulo state, while the same number for vereadores exceeded 6,498.

Under a collaboration with central party leadership, respondents were recruited via internal party communication networks. This was typically initiated from the central communications office, which sent an invitation to participate in the survey to state-level directors, who were instructed to pass the same message on to local (usually municipal) coordinators to circulate. This message explained the research collaboration, encouraged participation, and provided a link to the survey platform. Importantly, the message did not bear any mention of misinformation, but instead focused on building evidence to grow the party in the future. Local coordinators were encouraged to remind candidates several times after the initial sending. No respondent was provided with any in-kind or financial compensation.

Our sample spans several of Brazil's most populous states, including São Paulo, Paraná, and Minas Gerais. However, the core of our sample is squarely located in São Paulo, home to 645 municipalities. The sample contains at least one respondent from 35% of these municipalities. The number of vereadores in each municipality varies according to its population. Councilors are up for (re)election every four years (staggered two years from the general election) and do not face term limits. Voters select either a single candidate or a party, but seats are eventually awarded proportionately to the votes each party won.

Vereadores are important figures in local politics, serving as the legislative body of the municipal government, providing collaboration or opposition to the mayor. They are distinguished from national elites by their often more humble socioeconomic backgrounds, including lower levels of formal education. Table 1 presents some individual level traits, as well as balance by treatment. For example, only half of our sample (51%) completed secondary school. Women are comparatively well represented among

Table 1. Pre-treatment covariates

Trait	**Mean**	**S.D.**	% of T_1	% of T_0	**Range**
Gender (woman)	0.41	0.49	0.38	0.44	0,1
Completed Secondary	0.51	0.5	0.47	0.53	0,1
Elections competed in	2.04	1.45	2.04	2.04	1- 10
São Paolo state	0.63	0.48	0.63	0.62	0,1
N			464	409	

Note: Data only includes individuals present in experimental analyses, i.e. current candidates for vereador.

vereadores– 41% of our sample were women. Most were quite "green" in the world of politics, with a mean number of elections contested of 2 (including the current one).

Among those who responded, only current candidates for municipal council were given a experimental assignment. Though mayoral candidates are not outside our frame of theoretical interest, because we understood they may serve as local coordinators and thus potentially spoil the experiment, they also were not exposed to any treatment condition. In both of these cases, those respondents are excluded from all experimental analyses. Our final sample contains 873 individuals.

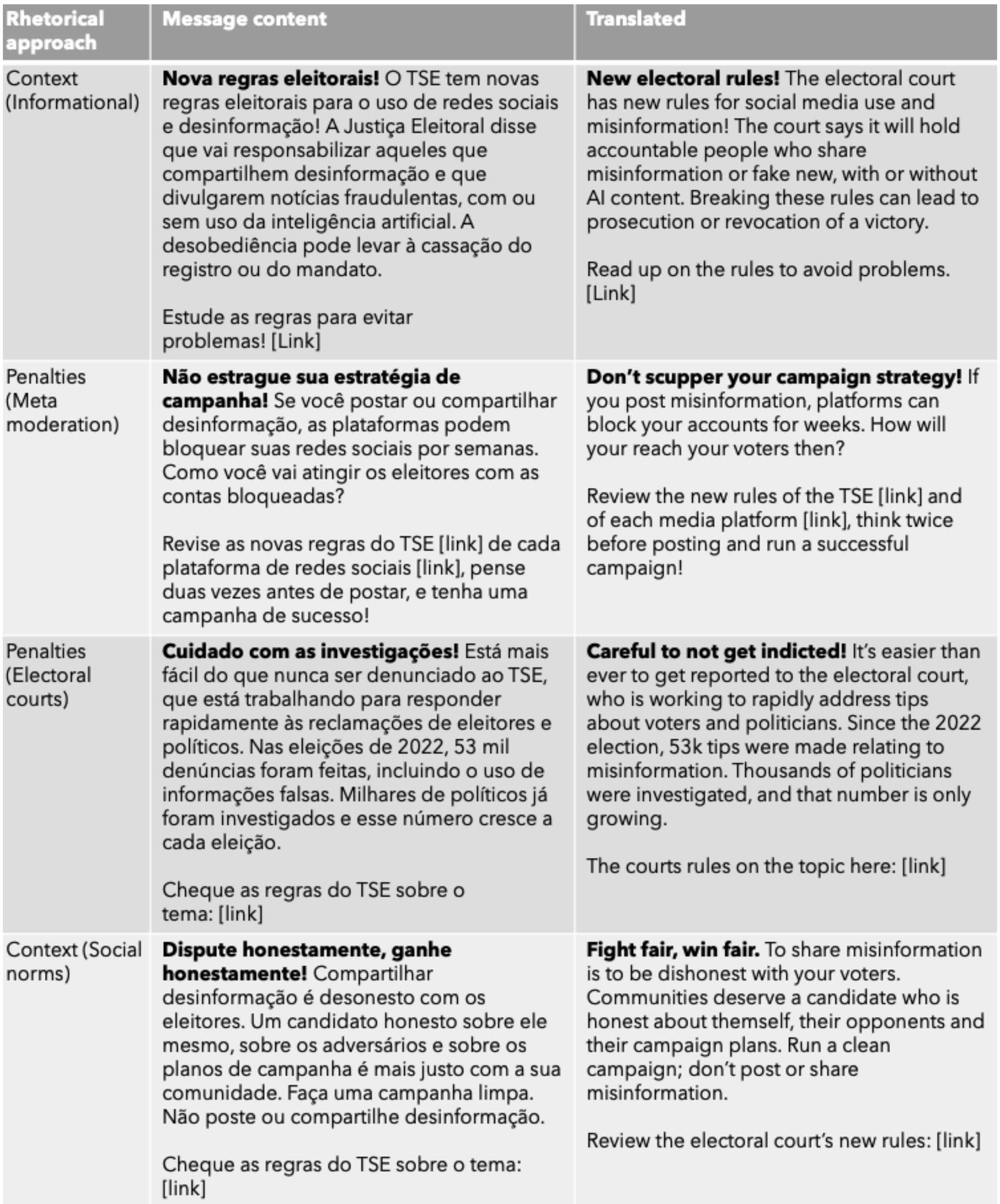
Empirical design

We explore our hypotheses through an online survey experiment. Respondents who consented to participate were assigned to one of 6 possible experimental groups (four treatment, placebo, and pure control) that determined whether they viewed a message attempting to raise the perceived costs of posting/reposting misinformation or if they are in a control condition. Prior to observing the main treatment message, respondents were presented with a small note on party letterhead from the party's president, noting that misinformation has been a big issue and that he wants to share one example of a message he found helpful, followed by his actual signature.

Those assigned to view a message were shown one of four messages that were developed by the National Confederation of Municipalities (CNM) and sent out to incumbent mayors in a parallel field experiment. The messages were developed to reflect different rhetorical approaches to communicating the potential costs of misinformation, collectively touching on the need to be informed, to avoid having campaign accounts suspended, to avoid being indicted and prosecuted, and to not disappoint voters. The specific messages are presented in Table 2.

While each was intended to be persuasive, a key purpose of this experiment was to understand their clarity and efficacy. To this end, we measured the time that the respondents spent on the page that contained their assigned treatment. In Figure 3, we present the distribution of time spent on the respective messages (including the 5 seconds they were required to stay on the page).[13] We note that the social norms treatment was viewed quickly compared to all others, but later present strong evidence this does not seem to have impacted efficacy. The message on potential legal penalties

[13] The links in the messages were shortened, but unclickable. Doing so prevented response attrition, but was informed by a parallel experiment in which we observed exceedingly low click-throughs. We exclude from this figure times that exceeded 100 seconds, which would indicate either a severe distraction or an attempt to visit the unclickable links included in messages.



Rhetorical approach	Message content	Translated
Context (Informational)	**Nova regras eleitorais!** O TSE tem novas regras eleitorais para o uso de redes sociais e desinformação! A Justiça Eleitoral disse que vai responsabilizar aqueles que compartilhem desinformação e que divulgarem notícias fraudulentas, com ou sem uso da inteligência artificial. A desobediência pode levar à cassação do registro ou do mandato. Estude as regras para evitar problemas! [Link]	**New electoral rules!** The electoral court has new rules for social media use and misinformation! The court says it will hold accountable people who share misinformation or fake new, with or without AI content. Breaking these rules can lead to prosecution or revocation of a victory. Read up on the rules to avoid problems. [Link]
Penalties (Meta moderation)	**Não estrague sua estratégia de campanha!** Se você postar ou compartilhar desinformação, as plataformas podem bloquear suas redes sociais por semanas. Como você vai atingir os eleitores com as contas bloqueadas? Revise as novas regras do TSE [link] de cada plataforma de redes sociais [link], pense duas vezes antes de postar, e tenha uma campanha de sucesso!	**Don't scupper your campaign strategy!** If you post misinformation, platforms can block your accounts for weeks. How will your reach your voters then? Review the new rules of the TSE [link] and of each media platform [link], think twice before posting and run a successful campaign!
Penalties (Electoral courts)	**Cuidado com as investigações!** Está mais fácil do que nunca ser denunciado ao TSE, que está trabalhando para responder rapidamente às reclamações de eleitores e políticos. Nas eleições de 2022, 53 mil denúncias foram feitas, incluindo o uso de informações falsas. Milhares de políticos já foram investigados e esse número cresce a cada eleição. Cheque as regras do TSE sobre o tema: [link]	**Careful to not get indicted!** It's easier than ever to get reported to the electoral court, who is working to rapidly address tips about voters and politicians. Since the 2022 election, 53k tips were made relating to misinformation. Thousands of politicians were investigated, and that number is only growing. The courts rules on the topic here: [link]
Context (Social norms)	**Dispute honestamente, ganhe honestamente!** Compartilhar desinformação é desonesto com os eleitores. Um candidato honesto sobre ele mesmo, sobre os adversários e sobre os planos de campanha é mais justo com a sua comunidade. Faça uma campanha limpa. Não poste ou compartilhe desinformação. Cheque as regras do TSE sobre o tema: [link]	**Fight fair, win fair.** To share misinformation is to be dishonest with your voters. Communities deserve a candidate who is honest about themself, their opponents and their campaign plans. Run a clean campaign; don't post or share misinformation. Review the electoral court's new rules: [link]

Figure 2. Treatment messages

gave respondents the most to think about, having a relatively uniform distribution up to 25 seconds.

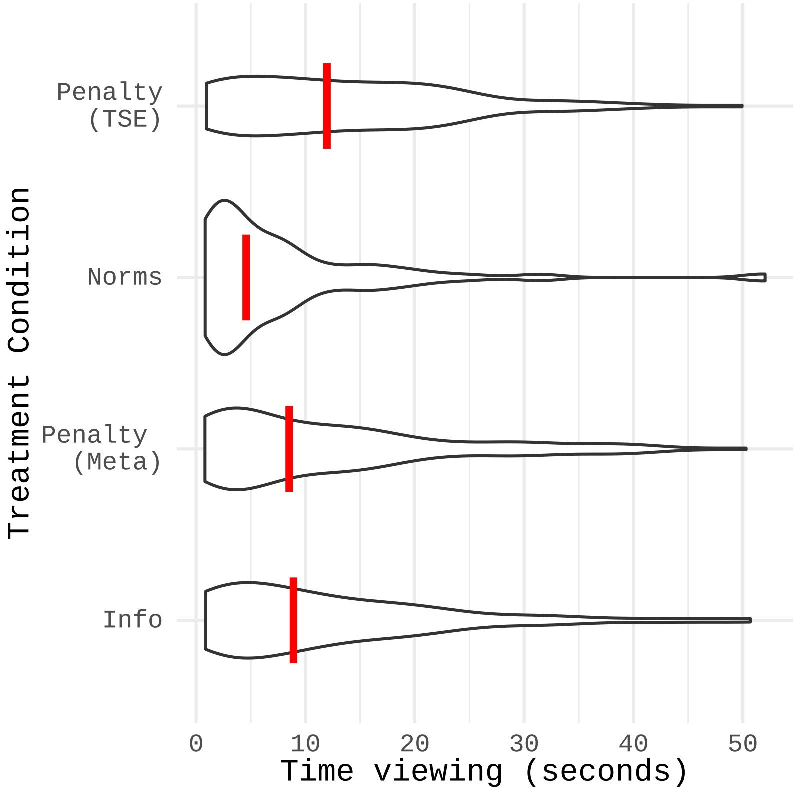


Figure 3. Time spent viewing each treatment. Violins are histograms of individual time spent on the treatment. Red lines represent the median of each distribution.

We compare treatment groups against two types of control conditions. The first is a pure control, in which respondents are thanked for their time so far and skip to the outcome questions. The second is a placebo condition, in which respondents are exposed to a different letter from the party's president that thanks them for their time and wishes them a good campaign. The inclusion of a placebo allows us to study the potential impact of source effects versus the content of messages (Traberg et al. 2024), by introducing a condition in which a figure of institutional authority whose mention can activate a subject's interest in performing "well", potentially through a sense of duty or pride for the party *without* delivering any cues on misinformation. This also mimics most PSA-type campaigns, to which public figures often lend their credibility. Moreover, a pure control allows us to assess more directly the impact of the information alone. In most analyses, we pool the placebo and control for power, though we also compare them directly in pre-registered hypotheses in Appendix D.

We most often pool all treatment conditions, though we conduct pre-registered comparisons of the two "penalty" conditions versus the *informational* and *social norms*. In addition we present exploratory treatment-specific effects in Appendix K.

We estimate average treatment effects via an OLS regression in a pre-registered empirical model. This takes the following form where $Treated_i$ is a binary treatment indicator, $\mathbf{X}_i$ is a vector of individual controls (gender, secondary education, and number of elections contested), $postelection_i$ is an indicator that takes on the value 1 if the response was completed after the October 6th election, and δ_i is a state-level fixed effect:

$$y_i = \beta_0 + \beta_1 Treated_i + \beta_2 \mathbf{X}_i + \beta_3 postelection_i + \beta_4 \delta_i + \epsilon_i$$

Standard errors clustered at the state-level.

Outcomes

Our feature set of outcomes is derived from a discernment task. Typically, discernment tasks ask a respondent to evaluate the veracity of a given news story/post/headline, presenting items that are both true and false (e.g. Pennycook and Rand (2022)). We modify this slightly (and therein set the bar higher) by asking respondents how likely they would be to share each of 10 headlines (from 0-9 on a slider).[14] These are presented in a fully randomized order, and respondents had to spend a minimum of 3 seconds on each.

Headlines are presented as Google News headlines (see Figure 4 below), containing a pixelated (and fake) source, the headline, a photo, and an age of story. Because a story being many days old might convey that it has survived moderators and is less likely to be false, we randomly assigned the full set of ages between 1 and 5 days, and did so with balance between true and false headlines.

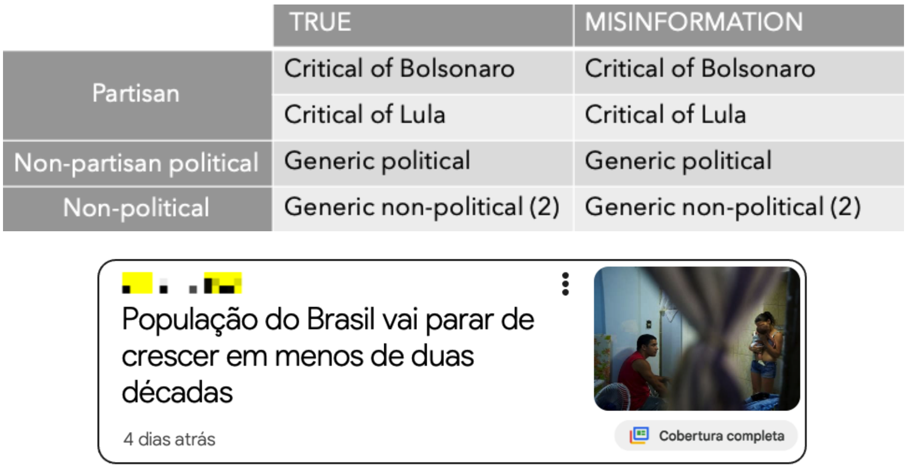
The essential goal of this typology was to present candidates with headlines that are consistent with their partisan ingroup bias, headlines that are non-partisan but political, as well as capture their openness to generic misinformation.[15] This not only extends the reach of potential findings, but also more accurately reflects content on a social media feed. We selected and pretested headlines from a number of online sources, keeping only ones that appeared to be of only moderate salience in the media, such that most respondents would not know their veracity with high certainty, but could still be expected to have some intuition as to their veracity. The headlines themselves complete the matrix of criteria in Figure 4.

From this exercise, we calculate three outcomes. First, and in following with previous studies, is a combined sharing (performance) score which consists of adding the likelihood of sharing true stories and the inverted likelihood of sharing false stories (i.e. interest in true stories and distinterest in false stories). We sum these and rescale this variable to fall between 0 and 1, and do the same with the next two measures. The second is the likelihood of sharing just false stories, which more precisely represents the effect of treatment on a respondent's discernment and likelihood of engagement with false headlines. The final is the likelihood of sharing true headlines, which allows us to assess whether there is a "chilling effect" on posting overall.[16]

[14] We set the scale from 0-9 so as to not allow a midpoint, but also avoid the colloquial assumption that 5 is the midpoint of any scale ending at 10.

[15] "Partisan" bias in the context of Brazil is conceptually distinct from the notion as developed in American politics (Green, Palmquist and Schickler 2004). Instead, partisan affect is mostly divided by whether they favor the Workers' Party (Lula) or the PL (Bolsonaro) (Samuels, Mello and Zucco 2024). We expect both in PSD's ranks, as it regularly forms municipal coalitions with both parties.

[16] Prior to rescaling, we exclude all respondents that either spent 40 seconds or more on three items, or more than 80 on one item. We flagged these as likely "googlers" or otherwise severely distracted respondents for



	TRUE	MISINFORMATION
Partisan	Critical of Bolsonaro	Critical of Bolsonaro
	Critical of Lula	Critical of Lula
Non-partisan political	Generic political	Generic political
Non-political	Generic non-political (2)	Generic non-political (2)

[redacted]

População do Brasil vai parar de crescer em menos de duas décadas

4 dias atrás

Figure 4. Matrix of headline types with an example of a true non-political headline

Though we faced a risk of selecting uninteresting headlines that no respondent would care to share, this was not the case. In Figure 5, we plot the headline-level distribution of self-reported sharing likelihoods. The median response always falls roughly between 5 and 6 (on a scale of 1-10). The distribution of the time spent evaluating each headline suggests the same (see Appendix A).

Next, we collaborate with a non-partisan NGO working on improving civility and decreasing misinformation online (Redes Cordiais) to give respondents the chance to sign a brief open letter that will be published before the election (but remain open to signing after). It covers basic themes on misinformation and includes a commitment of signers to avoid misinformation in the campaign. We conceive of signing the letter as a costly behavioral signal of their beliefs around misinformation (even if potentially performative).

Respondents were first presented with an explanation of the letter and its text before being asked if they would like to sign. On the next page, they encountered an open field in which they could sign in any chosen manner in response to the prompt, "Please enter your name as you wish it to appear on the letter." For ethical reasons, and as a point of theoretical interest, we allowed respondents to skip this question without answering, though it only appeared to those that first indicated they wished to sign. Approximately 67% agreed to sign. Only 3% of those who agreed made no attempt to provide a signature, while 88% provided a valid signature.

From this signing exercise, we create a four-point signing measure that we consider a quasi-continuous outcome where higher scores correspond to a stronger affirmation of the letter:

- **0:** Did not wish to sign
- **1:** Wished to sign, but did not provide a name that could be reasonably used to identify them
- **2:** Signed using at least a first and last name

whom responses had become a joint product of environmental influences as much as the survey prompts.

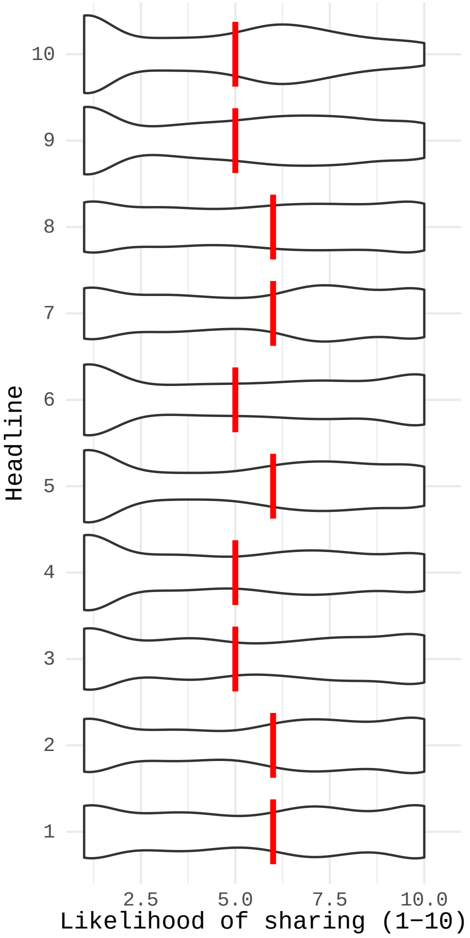


Figure 5. Distribution of reported likelihood of sharing by item.

- **3:** Signed using a first and last name and other identifying information (e.g. municipality, candidate number, or ballot name)

We also study the impact of treatment on effort in the discernment task. In the context of this study, this is captured by the amount of time spent on each item.[17] We calculate (1) the mean time spent on all statements, and (2) the mean on false statements minus the mean on true statements. These are each simple arithmetic means. Using a preregistered procedure we exclude individuals that spent more than 40 seconds on any three items, or 80 seconds on any one item.[18]

Our final two outcomes are a score on a misinformation rules knowledge index (-6,6) and an issue fatigue rating (-2,5). The former is described in detail in Appendix B and consists of three items which are scored -2 to 2, where two indicates high and accurate certainty. These items are then summed into a single score. The latter, described in the same appendix, sums scores on two items, one probing how often the respondent has

[17] While an imperfect proxy for effort, experimental evidence indicates cognitive effort is the primary source of variation in time measures of trials involving deliberation (Alós-Ferrer and Buckenmaier 2021).

[18] While this behavior may reflective of real world processes, it is highly likely to introduce confounding information between treatment and outcome that makes comparison difficult. However, in Appendix L we show our main results are robust to their inclusion.

received misinformation PSAs (0 to 3) and the other how much they think politicians would benefit from further messaging on misinformation (-2 to 2). These are again summed into a single score.

4. Main results

In broad terms, our results show that making clear and making salient the potential costs of misinformation can rapidly and widely shape behavior, both reducing interest in misinformation and introducing potential negative societal byproducts.

Our first results focus on the effect of treatment on behavior in the discernment exercise and are presented in Figure 6. We find that being assigned to treatment significantly improved performance on the combined performance measure. Treated individuals were more likely to demonstrate a combination of avoiding false stories and favoring true ones. The next (and our preferred) outcome focuses specifically the effect of treatment on willingness to share false stories. Here we observe that respondents that received an anti-misinformation message were, on average, 11% less likely to share a given false headline than untreated respondents. Together, these first findings demonstrate that without giving these respondents any information that would help them ascertain the veracity of headlines, they seem to have performed remarkably better in doing so.

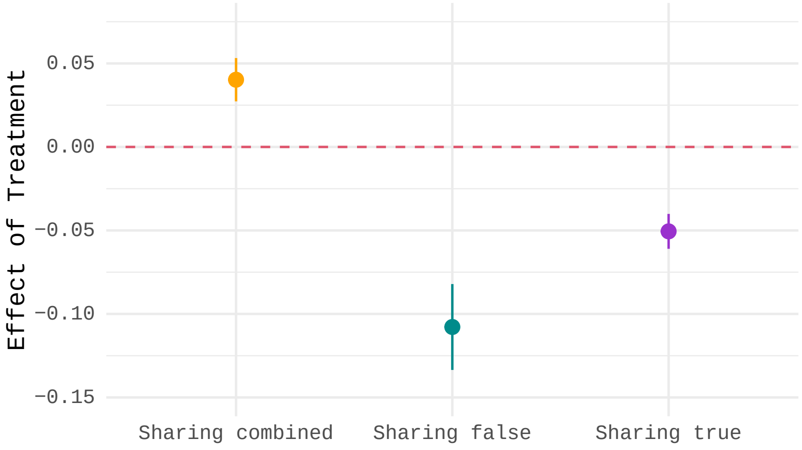
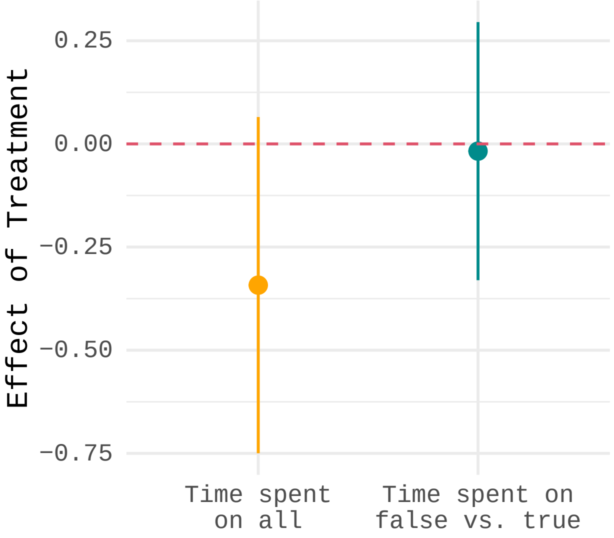


Figure 6. Main effect of treatment on sharing in discernment exercise. Coefficients map to H1a, H1b, and H1c, moving left to right. All models include demographic covariates and state fixed-effects. Standard errors are clustered at the state-level.

Yet, those effects would be interpreted differently if there were a symmetrical effect on their interest in true stories. In other words, the combined measure could mask a chilling effect, in which sharing is reduced across the board. The third coefficient in Figure 6 represents effect on sharing of true headlines, which, absent a chilling effect, should not be significant. There is, however, a moderate chilling effect. As consolation, we note that this effect is almost precisely half the magnitude of the effect on sharing false stories. This firstly adds to evidence that anti-misinformation interventions require careful framing to avoid unintended negative outcomes (Carey et al. 2020, Mosleh et al. 2021). While this bears further research, we believe caution should be measured; the chilling effect is significantly offset by the magnitude of the reduction in interest in false headlines. Treatment may be still be seen as resulting in a net improvement in the current equilibrium.

Did treatment cause increased effort in shaping how respondents spent time evaluating these headlines? In Figure 7, we present our evaluations of hypotheses **1d** and **1e**. In fact, we cannot reject the null hypothesis in either case. Without placing undue stock in a null result, this hints at a puzzle as performance improved without having an obvious effect on effort. Reconciling this with the results in Figure 6, we believe this tells a positive story; treated respondents did not require additional time to make better choices.

Figure 7. Effects on time allocation in the discernment exercise.



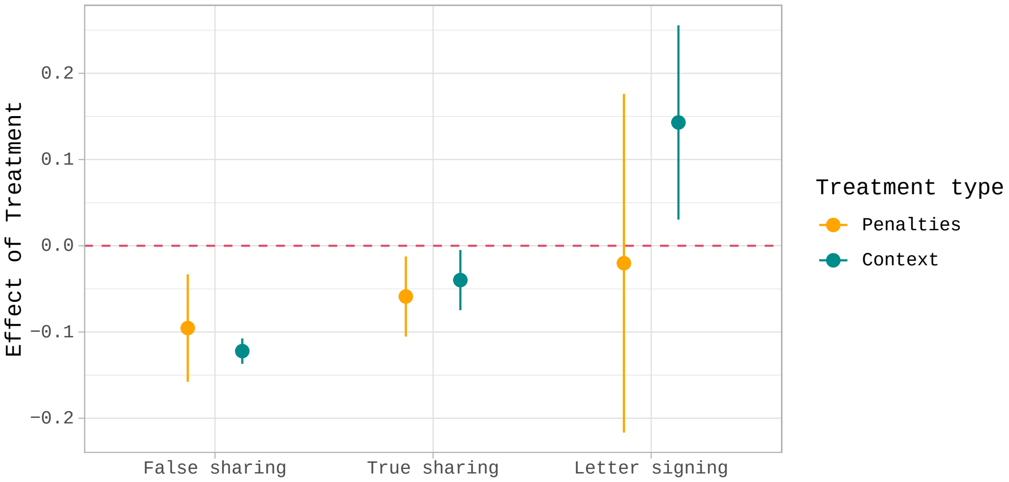
We predicted that treatment would increase scores on a knowledge index of misinformation rules through activating interest and effort that sustains better recall and reasoning. We confirm this hypothesis as treated respondents scored slightly higher ($\beta = 0.16$, $p = 0.04$). Substantively, this impact is small, but does provide basic evidence that there may be latent knowledge of regulations that is activated under the right incentives. Nevertheless, knowledge levels remain generally low and the median score on this index only just favors the average response being correct.[19]

Our prediction that treatment would increase rates of letter signing was not borne out in our analysis. Instead, we find a null result on letter signing that is “robust” to alternative specifications. We argue that this is not uninformative. Only 34% of our sample outright refused to sign the letter. We believe this simply indicates that candidates have stable predispositions (or lack thereof) to participate in costly displays of values. This maps to the findings of Porter, Velez and Wood (2023), which show that accurate beliefs about the world are much more easily shifted than underlying orientations and attitudes.

Turning to our analysis of the impact of the type of message, we present Figure 8.

[19] In Appendix J we show all results described so far are robust to regressing outcomes on only a binary treatment indicator (i.e. with no controls).

Figure 8. Comparing treatment types



For each outcome, we hypothesized that the effect of a *penalty* message would be greater and statistically distinguishable from that of an informational or social norms message. These are rejected. However, we place exceptional value in the combined and false sharing results remaining significant in the *context* conditions, which one could expect to be a weaker treatment. In fact, comparing respondents in context conditions to the control is the only case in which letter signing improved as a result of treatment. Still, we cannot say either type of message was more or less effective, but it should be surprising that even just reminding respondents that the much gentler *context* conditions sustained the over all result.

Folded into this null finding is the possibility that in generating penalty salience, the need for more aggressive messaging may be overstated relative to less aggressive approaches. We pair this assessment with the fact that knowledge scores were low across the board; treated or not, approximately half of the sample didn't know that they can be prosecuted for posting misinformation even if they did not know it was misinformation at the time.[20] This indicates that, at most, the legal changes around misinformation have been successful in building deterrence, but that the criteria for avoiding risk remain unclear. We explore the treatment type comparison further and find no evidence that either caused greater issue fatigue.

In our final analysis, we examine whether treatment effects are moderated by pre-treatment hopes of being elected, i.e. pessimism about your electoral chances decreases responsiveness to treatment. This hypothesis is rejected as revealed in Table 2. Instead we find the opposite; higher levels of pessimism produce even greater reductions in interest in sharing false headlines. This holds even when restricting to just the pre-electoral period. We can muster no reasonable explanation for this result.

[20] Even among mayors who answered our survey incidentally (N= 308) and who should possess more sophisticated knowledge, this figure remained at just 61%.

Table 2. Campaign optimism as a moderator

	Sharing false headlines
Treatment	-0.054^{*}
	(0.031)
Campaign Pessimism	-0.009^{**}
	(0.004)
Treatment:Campaign Pessimism	-0.015^{***}
	(0.005)
Observations	873
Adjusted R^2	0.065
Note: this model includes our full battery of controls, as well as state fixed-effects. Std. errors are clustered at the state-level.	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.001$

Robustness and Discussion

We consider several features of our main approach that may limit the implications of our findings and attempt to address them empirically.

First, we address the concern that our main results are driven by the impact of the party president (Gilberto Kassab) including a note to respondents just before treatment. We leverage the placebo condition to explore whether there remains an effect of pushing on the theme of misinformation that is independent from his endorsement/presence. We repeat our main analyses on the discernment exercise (combined score, sharing false, and sharing true) and regress this on whether a respondent was in a treatment condition or in the placebo condition, dropping the pure control. We find that this experimental comparison still produces significant results (of the same sign and similar magnitude), indicating that the presence of an authority figure alone (Kassab) was not the sole driver of our results.[21]

Secondly, we explore the possibility of spillovers between respondents in the same municipality. This could include sharing the content of a treatment message, as well as information on which of the headlines in the discernment task were false (as is revealed in the debriefing). We have no evidence this occurred and note that both would be a laborious effort for a respondent and require very effective recall. To test it empirically, in Appendix G we re-estimate our main discernment results using cases in which spillovers would bias against our results, namely when spillovers would improve the performance of the control group.[22] We find that the main results on combined sharing, false sharing, and true sharing are all consistent.

Thirdly, we disaggregate our results by the type of headline to explore whether results around discernment were driven solely by topics that are obviously political (including partisan congruent and incongruent) or non-political headlines (which are less likely to intersect with partisan biases). We report this analysis in full in Appendix F. We

[21] See Appendix D for further details.

[22] This includes cases when the first respondent is in the control and the next one or two respondents are also in the control group (allowing for learning), as well as cases in which a treatment respondent is followed by one or more control respondents. We also include cases in which there was only one respondent in that municipality.

distinguish no difference in significance or magnitude when it comes to the combined sharing outcome, sharing false stories, and the chilling effect. Whereas one might expect improvements in accuracy to occur most in areas of respondents' expertise (politics), we observe treatment had similar impact across themes.

Next, we explore whether the null result on the open letter outcomes (which were consistent across every pre-registered analysis) is a product of how we defined the outcome. Recall our preferred approach distinguishes between signatures that provide just a first and last name, from those that include additional identifying details. We re-specify the outcome as a binary, where these two categories are both coded as 1.[23] We then re-estimate this analysis and find no substantive changes in the coefficients and their significance. This is presented in Appendix D.

We also explore the potential for heterogeneous treatment effects which may indicate sub-populations for whom our intervention was not effective. In Appendix I, we report analyses of two pre-registered hypotheses exploring this across gender and education. We find that experimental effects remain in each sub-sample. In addition, we find no evidence of divergent or disparate effects by gender. However, respondents with lower levels of formal education appear to have been more responsive to treatment as indicated by lower rates of sharing false headlines. We hesitate to identify this result as indicative of a direct effect of formal education on plasticity in incentives in approaching online behavior. Instead, it more likely indicates a role of socio-economic factors that have moderated how informed of the risks of posting misinformation candidates were, prior to treatment. When regressing scores on the rules knowledge index (using our controls and state-fixed effects) among just the control group, we find more formal education is significantly associated with higher knowledge.

Finally, we consider who is in our sample and how this might impact the reach of our results. Non- and partial response rates in elite surveys is a persistent issue (Matteos and Corral, 2022). In Appendix H we explore factors associated with dropping out of the survey (i.e. not completing our main outcomes) assuming these same characteristics drive non-response as well. Level of education is strongly associated with this– with more educated individuals being more likely to finish the survey. We argue this indicates the need for inclusive messaging in anti-misinformation messaging, but also suggests our results represent a floor for the potential impact of anti-misinformation efforts.

5. Conclusion

Misinformation on digital media has threatened voters' abilities to acquire accurate information and much effort and intellect has been expended in the search for solutions to this crisis confronting democratic systems. Most of this effort has focused squarely on solutions that empower information consumers (voters) to value, seek out, and learn from accurate information, while ignoring or avoiding misinformation. The checkered results of those studies, issues relating to scalability, and now our findings suggest efforts may be better focused on upstream solutions. Correctly structuring political elites' incentives may reduce the overall supply of misinformation.

We study this supply-side approach in the context of Brazil's 2024 municipal elections.

[23] We code signatures that include just a first name or last name as not having signed.

With a diverse sample of active candidates for municipal council, we were able to test the effects of communicating the practical risks of posting misinformation, as well as softer approaches on several classes of outcomes. Our findings should be taken as wide encouragement for the targeting of misinformation closer to its source.

Candidates that received a message outlining the risks or context of misinformation were less likely to favor misinformation. This was not contingent on whether the message focused on penalties versus the context, nor does it appear to be moderated by gender, nor education. However, we highlight a concerning secondary ramification in the significant reduction in interest in accurate headlines as well. This finding is robust to several respecifications. It suggests that raising the salience of penalties in the absence of knowledge about the particulars of regulations may encourage risk avoidant behavior instead of empowered online engagement.

We conduct a multitude of tests to mitigate the possibility that these effects are a product of our design and empirical approach. These tests do not suggest that our pattern of results are limited in this way. Instead, we highlight that effects are consistent when accounting for source effects and spillovers within municipalities, as well as alternative definitions of our outcomes (including disaggregating discernment by political and non-political headlines). Furthermore, we show results remain when considering levels of formal education and gender of the respondent.

This study is just the third known experimental evaluation of an attempt to change elites' relationship to misinformation (following Ma et al. (2023) and Nyhan and Reifler (2015)). We find the intervention effective and provide rich new evidence on potential mechanisms behind that efficacy. Moreover, this is the first known study on this topic in the Global South, where media ecosystems and campaigns differ significantly from the United States (where both previous studies focused). Lastly, we note that our sample makes the study more similar to Ma et al. (2023), who attempted to shift the behavior of *state* legislators (as opposed to federal politicians). They attribute null results primarily to fact-checking lacking teeth because of the paucity of media coverage garnered by state legislators. Our sample of candidates for local council rarely feature in the news, but our intervention is, in theory, backed by a social, regulatory, and prosecutorial environment that does not require viewership/readership for penalties to be assessed.

We argue these findings can be extended somewhat liberally. Our sample contained a remarkable balance on gender and education. While we consider potential selection into our sample, it appears those for whom treatment was most effective (less formal education) were more likely to not complete the survey. Thus, we anticipate treatment effects would be likely to grow in magnitude with less selection. Lastly, Brazil's response following the hammer blow of the January 2022 coup attempt has set a new standard for legal responses to the threat of misinformation and general social salience of the issue. Yet crucially, we show that our main results are consistent when examining two of our intervention messages that more gently appeal to social norms against dishonesty and the need to remain informed of campaign rules. We cannot separate any of our treatments from pre-experimental associations of misinformation with legal penalties. That is, there is likely some cognitive equivalence between all treatment conditions despite their wide differences. Still, we demonstrate there is ample play with regard to the kinds of messaging that can be used to make misinformation's costs salient when attempting to address information at its source. Solving issues caused by misinformation need not rely on demand-side interventions alone. Instead, upstream interventions can go beyond dependency on individual conviction or intelligence, to affect behavioural change around misinformation through simple reminders to political actors of the stakes at hand.

References

Alós-Ferrer, Carlos and Johannes Buckenmaier. 2021. "Cognitive sophistication and deliberation times." *Experimental Economics* 24(2):558–592.

Bateman, Jon and Dean Jackson. 2024. Countering Disinformation Effectively: An Evidence-Based Policy Guide. Technical report Carnegie International Endowment for Peace.

Batista Pereira, Frederico, Natália S Bueno, Felipe Nunes and Nara Pavão. 2022. "Fake news, fact checking, and partisanship: the resilience of rumors in the 2018 Brazilian elections." *The Journal of Politics* 84(4):2188–2201.

Bayer, Joseph B, Penny Trieu and Nicole B Ellison. 2020. "Social media elements, ecologies, and effects." *Annual review of psychology* 71(1):471–497.

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

Blair, Robert A., Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote and Charlene J. Stainfield. 2024. "Interventions to Counter Misinformation: Lessons from the Global North and Applications to the Global South." *Current Opinion in Psychology* 55:101732.

Bordalo, Pedro, Nicola Gennaioli and Andrei Shleifer. 2022. "Salience." *Annual Review of Economics* 14(1):521–544.

Bräuninger, Thomas and Nikolay Marinov. 2022. "Political elites and the "War on Truth"." *Journal of Public Economics* 206:104585.

Bullock, John G, Alan S Gerber, Seth J Hill and Gregory A Huber. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10(4):519–578.

Carey, John M, Victoria Chi, DJ Flynn, Brendan Nyhan and Thomas Zeitzoff. 2020. "The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil." *Science advances* 6(5):eaaw7449.

Castro, Lucio and Carlos Scartascini. 2015. "Tax compliance and enforcement in the pampas evidence from a field experiment." *Journal of Economic Behavior & Organization* 116:65–82.

Cusumano, Michael A, Annabelle Gawer and David B Yoffie. 2021. "Social media companies should self-regulate. Now." *Harvard Business Review* 15.

Druckman, James N. 2001. "On the limits of framing effects: Who can frame?" *The journal of politics* 63(4):1041–1066.

Ecker, Ullrich KH, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga and Michelle A Amazeen. 2022. "The psychological drivers of misinformation belief and its resistance to correction." *Nature Reviews Psychology* 1(1):13–29.

Farhall, Kate, Andrea Carson, Scott Wright, Andrew Gibbons and William Lukamto. 2019. "Political elites' use of fake news discourse across communications platforms." *International Journal of Communication* 13:23.

Flynn, Daniel J, Brendan Nyhan and Jason Reifler. 2017. "The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics." *Political Psychology* 38:127–150.

Frenkel, M. 2023. "The pro-Bolsonaro riot and Jan. 6 attack followed a similar digital playbook, experts say." *The New York Times* 10.

Gimpel, Henner, Sebastian Heger, Christian Olenberger and Lena Utz. 2021. "The effectiveness of social norms in fighting fake news on social media." *Journal of Management Information Systems* 38(1):196–221.

Green, Donald P, Bradley Palmquist and Eric Schickler. 2004. *Partisan hearts and minds: Political parties and the social identities of voters.* Yale University Press.

Grossman, Gene M and Elhanan Helpman. 2023. "Electoral competition with fake news." *European Journal of Political Economy* 77:102315.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler and Neelanjan Sircar. 2020. "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* 117(27):15536–15545.

Gurr, Gwendolin and Julia Metag. 2021. "Examining avoidance of ongoing political issues in the news: A longitudinal study of the impact of audience issue fatigue." *International Journal of Communication* 15:21.

Hameleers, Michael. 2020. "Populist disinformation: Exploring intersections between online populism and disinformation in the US and the Netherlands." *Politics and Governance* 8(1):146–157.

Hochschild, Jennifer L and Katherine Levine Einstein. 2015. "Do facts matter? Information and misinformation in American politics." *Political Science Quarterly* 130(4):585–624.

Jerit, Jennifer and Yangzi Zhao. 2020. "Political misinformation." *Annual Review of Political Science* 23(1):77–94.

Lasser, Jana, Segun Taofeek Aroyehun, Almog Simchon, Fabio Carrella, David Garcia and Stephan Lewandowsky. 2022. "Social media sharing of low-quality news sources by political elites." *PNAS nexus* 1(4):pgac186.

Lenz, Gabriel S. 2012. *Follow the leader?: how voters respond to politicians' policies and performance.* University of Chicago Press.

Lewandowsky, Stephan and Sander Van Der Linden. 2021. "Countering misinformation and fake news through inoculation and prebunking." *European Review of Social Psychology* 32(2):348–384.

Ma, Siyuan, Daniel Bergan, Suhwoo Ahn, Dustin Carnahan, Nate Gimby, Johnny McGraw and Isabel Virtue. 2023. "Fact-Checking as a Deterrent? A Conceptual Replication of the Influence of Fact-Checking on the Sharing of Misinformation by Political Elites." *Human Communication Research* 49(3):321–338.

Margolis, Mac and Robert Muggah. 2023. "Brazil breaks new ground in the global fight against fake news.".

URL: *https://www.opendemocracy.net/en/democraciaabierta/brazil-crack-down-fake-news-disinformation-lula-restore-trust-internet/*

Mateos, Araceli and Margarita Corral. 2022. "Partial non-response in political elite studies: an approach to parliamentary elites in Latin America." *Quality & Quantity* 56(6):4089–4106.

Morrison, Mark, Kevin Parton and Donald W Hine. 2018. "Increasing belief but issue fatigue: Changes in Australian household climate change segments between 2011 and 2016." *PloS one* 13(6):e0197988.

Mosleh, Mohsen, Cameron Martel, Dean Eckles and David Rand. 2021. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a Twitter field experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. pp. 1–13.

Mosleh, Mohsen and David G Rand. 2022. "Measuring exposure to misinformation from political elites on Twitter." *Nature Communications* 13(1):7144.

Nyhan, Brendan. 2021. "Why the backfire effect does not explain the durability of political misperceptions." *Proceedings of the National Academy of Sciences* 118(15):e1912440117.

Nyhan, Brendan and Jason Reifler. 2015. "The Effect of Fact-Checking on Elites: A Field Experiment on U.S. State Legislators." *American Journal of Political Science* 59(3):628–640.

Pasek, Josh, Gaurav Sood and Jon A Krosnick. 2015. "Misinformed about the affordable care act? Leveraging certainty to assess the prevalence of misperceptions." *Journal of Communication* 65(4):660–673.

Pennycook, Gordon and David G Rand. 2019. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition* 188:39–50.

Pennycook, Gordon and David G Rand. 2022. "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation." *Nature communications* 13(1):2333.

Pereira, Frederico Batista, Natália S Bueno, Felipe Nunes and Nara Pavão. 2023. "Inoculation Reduces Misinformation: Experimental Evidence from Multidimensional Interventions in Brazil." *Journal of Experimental Political Science* pp. 1–12.

Pérez, Efrén O. 2015. "Ricochet: How elite discourse politicizes racial and ethnic identities." *Political Behavior* 37:155–180.

Persily, Nathaniel, Joshua A Tucker and Joshua Aaron Tucker. 2020. "Social media and democracy: The state of the field, prospects for reform.".

Porter, Ethan, Yamil Velez and Thomas J Wood. 2023. "Correcting COVID-19 vaccine misinformation in 10 countries." *Royal Society open science* 10(3):221097.

Prike, Toby, Lucy H Butler and Ullrich KH Ecker. 2024. "Source-credibility information and social norms improve truth discernment and reduce engagement with misinformation online." *Scientific Reports* 14(1):6900.

Samuels, David, Fernando Mello and Cesar Zucco. 2024. “Partisan stereotyping and polarization in Brazil.” *Latin American Politics and Society* 66(2):47–71.

Samuels, David J and Karine Belarmino. 2024. “Partisan Dehumanization in Brazil’s Asymmetrically Polarized Party System.” *Journal of Politics in Latin America* .

Shao, Chengcheng, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer and Giovanni Luca Ciampaglia. 2018. “Anatomy of an online misinformation network.” *Plos one* 13(4):e0196087.

Slemrod, Joel, Marsha Blumenthal and Charles Christian. 2001. “Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota.” *Journal of Public Economics* 79(3):455–483.

Traberg, Cecilie S, Trisha Harjani, Jon Roozenbeek and Sander van der Linden. 2024. “The persuasive effects of social cues and source effects on misinformation susceptibility.” *Scientific Reports* 14(1):4205.

Vraga, Emily K. and Leticia Bode. 2020. “Defining Misinformation and Understanding Its Bounded Nature: Using Expertise and Evidence for Describing Misinformation.” *Political Communication* 37(1):136–144.

Yildirim, Mustafa Mikdat, Jonathan Nagler, Richard Bonneau and Joshua A Tucker. 2023. “Short of suspension: How suspension warnings can reduce hate speech on twitter.” *Perspectives on Politics* 21(2):651–663.

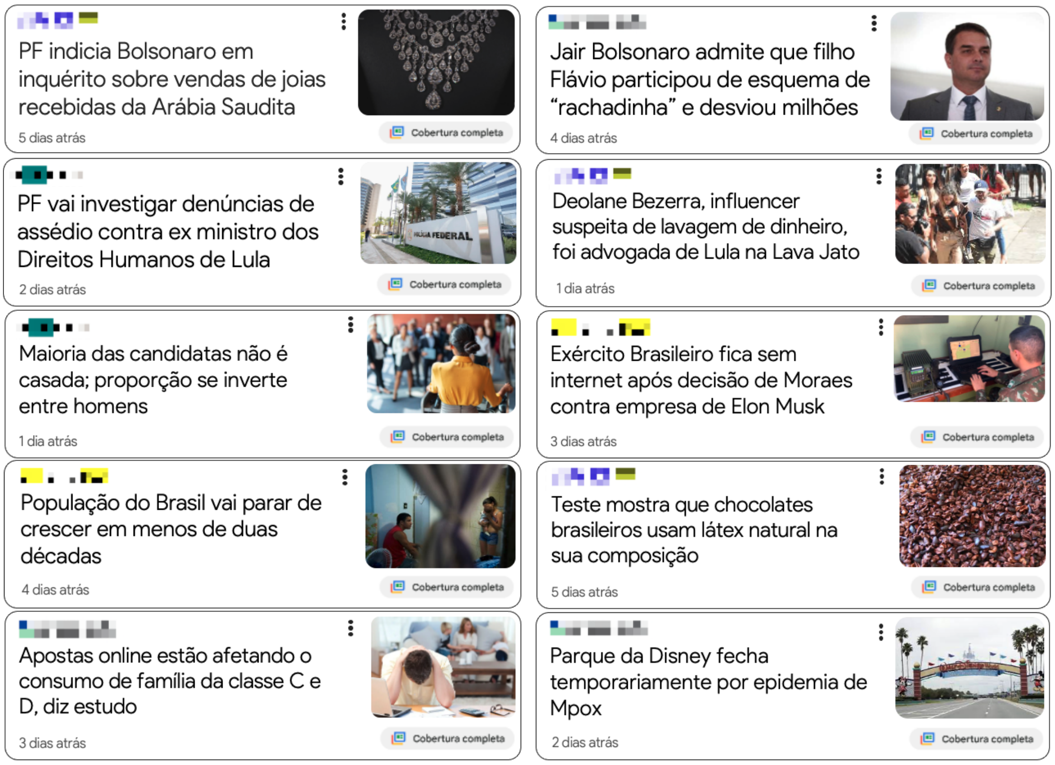
Zaller, John R. 1992. *The Nature and Origins of Mass Opinion.* Cambridge Studies in Public Opinion and Political Psychology Cambridge: Cambridge University Press.

Zhang, Yini, Fan Chen and Josephine Lukito. 2023. “Network amplification of politicized information and misinformation about COVID-19 by conservative media and partisan influencers on Twitter.” *Political Communication* 40(1):24–47.

Zucco, Cesar and Timothy J Power. 2024. “The Ideology of Brazilian Parties and Presidents: A Research Note on Coalitional Presidentialism Under Stress.” *Latin American Politics and Society* 66(1):178–188.

Appendix A. Discernment Task Details

The discernment task revolves around the presentation of the same 10 headlines to respondents in a subject-level randomization of their order. The headlines selected were the following:



PF indicia Bolsonaro em inquérito sobre vendas de joias recebidas da Arábia Saudita

5 dias atrás

Cobertura completa

PF vai investigar denúncias de assédio contra ex ministro dos Direitos Humanos de Lula

2 dias atrás

Cobertura completa

Maioria das candidatas não é casada; proporção se inverte entre homens

1 dia atrás

Cobertura completa

População do Brasil vai parar de crescer em menos de duas décadas

4 dias atrás

Cobertura completa

Apostas online estão afetando o consumo de família da classe C e D, diz estudo

3 dias atrás

Cobertura completa

Jair Bolsonaro admite que filho Flávio participou de esquema de "rachadinha" e desviou milhões

4 dias atrás

Cobertura completa

Deolane Bezerra, influencer suspeita de lavagem de dinheiro, foi advogada de Lula na Lava Jato

1 dia atrás

Cobertura completa

Exército Brasileiro fica sem internet após decisão de Moraes contra empresa de Elon Musk

3 dias atrás

Cobertura completa

Teste mostra que chocolates brasileiros usam látex natural na sua composição

5 dias atrás

Cobertura completa

Parque da Disney fecha temporariamente por epidemia de Mpox

2 dias atrás

Cobertura completa

Figure A1. Headlines used in the discernment task. The left-hand column contains the true headlines, while the right-hand column contains the false headlines.

In the main text, we present the distributions of sharing rates at the headline level, noting that all headlines appeared to be of reasonable interest and salience. Here, we also check whether the amount of time respondents spent on each item varied significantly. In general, it appears this was not case– respondents' median amount of time spent on each item was relatively similar with no major outliers. From this we argue that the headlines resulted in a very steady series of "trials", and that no one headline appears to have introduced design effects by this measure.

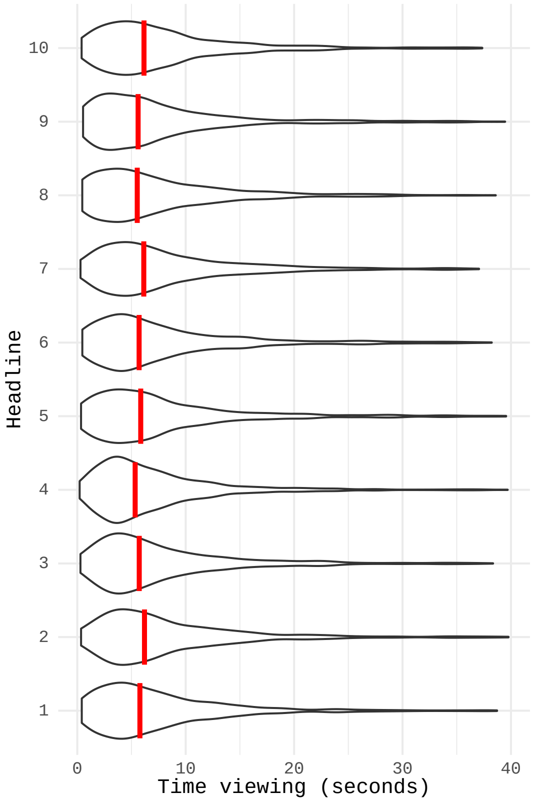


Figure A2. Distribution of time spent evaluating each headline. Vertical lines are the median time spent.

Appendix B. Additional measures

Here we provide additional details on several measures used in our main analyses.

Misinformation rules knowledge index. The fundamental goal of this index was to gauge the extent to which a potential null result was due to respondents already being very well equipped with knowledge around misinformation, its risks, and its costs. At such, we positioned it after treatment because we broadly expect that after successfully raising prior beliefs about the risks and costs of posting misinformation, respondents would on average exert more effort in demonstrating their best knowledge around regulations. In order to construct an index that can meaningfully parse knowledge levels, we sought and included items that vary in difficulty, from the relatively easy to the highly refined knowledge.

Respondents were presented with three factual propositions which they were asked to evaluate according to a 5 point scale, from *definitely false* to *definitely true.* In the end two of the three items were false. Responses were mapped to an integer scale where -2 corresponds to *definitely* on the incorrect pole, and 2 corresponds to *definitely* on the correct pole. The resulting distribution is presented in Figure B1.

The three items (in order of difficulty) were as follows:

- You cannot be reported to the electoral court for posting misinformation that you didn't know was misinformation (False)
- If a candidate posts some kind of misinformation, they can have their candidacy or eventual victory nullified (True)
- Currently, online platforms are fined R$ 5 mil. for every hour they leave misinformation on their sites (False, R$ 50 mil.)

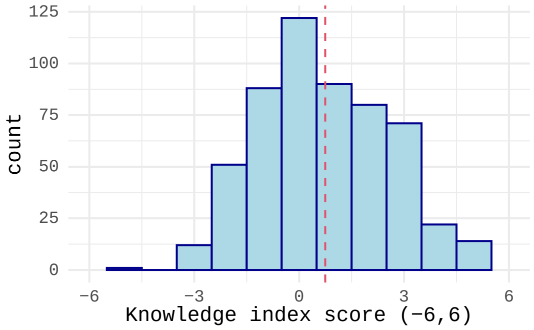


Figure B1. Distribution of Rules Knowledge outcome.

Issue fatigue index. The issue fatigue index attempts to capture the extent to which candidate's messaging environment has already been saturated with PSAs around the topic of misinformation. Again, we have a descriptive interest in analyzing this variable as a potential explanation for potential null results– if a candidate has been "oversensitized around an issue, they are unlikely to respond meaningfully to a brief warning, such as our treatment. As such, placing it after the treatment offers a window into how such a brief message affects this kind of fatigue or apathy around misinformation.

The index consists of the sum of an integer mapping of responses to two items:

- How often have you seen ads or received messages warning about misinformation, either online or in your daily life?
  - Never, Monthly, Weekly, Daily
- Do you agree or disagree with the following: Politicians would benefit from more information and targeting on misinformation and regulations around it.
  - Strongly Agree - Strongly disagree

The former is scaled 0-3, while the latter is scaled -2 to 2. These are summed to form the index with the distribution depicted in Figure B2 (recall higher score indicates higher levels of fatigue).

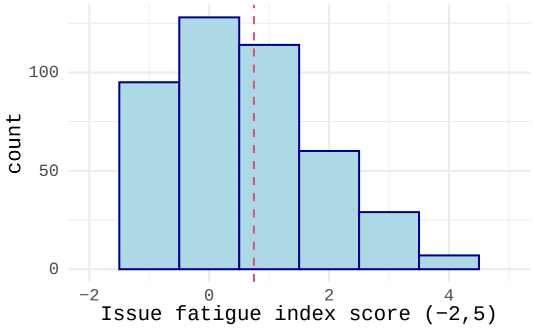


Figure B2. Distribution of Issue Fatigue Outcome

Campaign pessimism. In one of our hypotheses we explore whether treatment effects may be moderated by the pessimism or cynicism a candidate feels about their electoral chances. Briefly summarized, out of spite, anger, or simply lower perceived opportunity costs, we believe that pessimistic candidates are at baseline more likely to share misinformation. In turn, we predicted this would impinge upon our treatments' ability to improve their behavior.

To this end we presented them with a single item:

- In your opinion, what are your chances of getting elected this year?

Respondents evaluated this from 0-9, where 0 is very likely (least pessimistic). The histogram of this item skews heavily to the left end of the scale. We should expect this; active candidates would seem unlikely to admit insecurity, especially in a survey. However, we do observe a fair amount of variation around the mean.

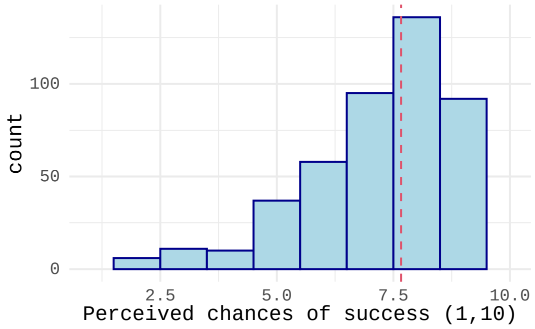


Figure B3. Distribution of Campaign Pessimism Item

Appendix C. Manipulation Checks

We included four items that are meant to map to the underlying dimension each treatment attempted to manipulated. These items were pre-registered, but we did not commit to them as the only indicators of effective treatment as the boundaries between each treatment are not clearly defined and seek to manipulate disparate aspects of candidates' thinking. The four items were as follows:

- (TSE penalty) In your opinion, what risk of being prosecuted from the TSE do candidates that post misinformation face? (0-9)
- (Meta penalty) In your opinion, what risk of losing access to their Facebook and Instagram accounts do candidates that post misinformation face? (0-9)
- (Social norms) In your opinion, what risk of being seen in a negative light do candidates that post misinformation face? (0-9)
- (Informational) In your opinion, how likely is it that candidates that post misinformation do so because they don't know the rules (0-9)

In Figure C1 we present our analysis of the manipulations using these items. Each coefficient corresponds to the results of regressing the given outcome on a treatment indicator that takes on the value 1 if the respondent was in the corresponding treatment. Each model includes our full set of controls.

These checks only corroborate the efficacy of one treatment in manipulating the underlying dimension. Whilst we generally do not place too much stock in this exercise, as each of these tests lack statistical power,– averaging just over 110 treated individuals per treatment group– we interpret the negative sign on the informational treatment check as indicative of the treatment having activated a social desirability bias around how informed candidates should be. In this sense, it is closer to capturing the dynamics we hoped to manipulate in the social norms treatment.

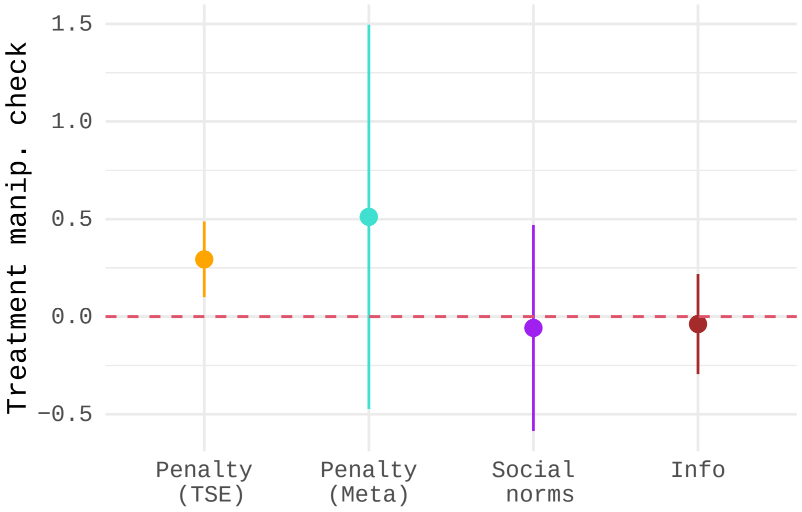


Figure C1. Manipulation checks.

Appendix D. Placebo Analysis

We explore the potential that the main results of treatment are driven largely by the presence of a letter from the party president (Kassab) addressed to respondents and bearing his real signature. This feature of the design was included to increase the strength of what was otherwise a brief treatment, building interest and attention. Simultaneously, this also bears clear ties to many PSAs which leverage the credibility of public figures to increase reach and receptivity of messages, often studied as *source effects* in the literature. In order to be able to recover an independent effect of the President's words, we included a control condition (placebo) that contains a brief message from the President, with no mention of misinformation.

In this analysis we estimate average treatment effects comparing treatment respondents to just the placebo condition. If the main results were fully driven by this endorsement of a trusted and well-known figure, we would expect there to be no discernible effects of having been in a treatment group compared to the placebo. Using the same specification as our first three results on the discernment module (combined performance, false sharing, true sharing), we find that all three results are stable after "netting out" the impact of the letter. In other words, there remains an independent effect of the misinformation treatments. Table D1 presents these models.

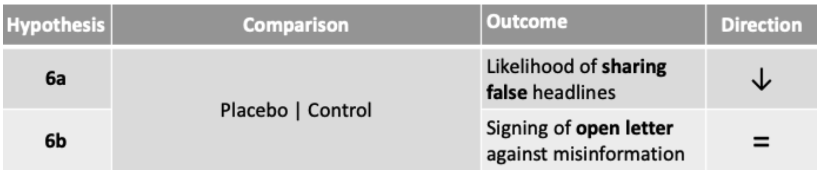
We also pre-registered two hypotheses that compare the placebo and pure control, which allow us to explore how the appearance/intervention of an authority figure shapes behavior around misinformation, in this case the impact of Kassab's intervention (in Table D1). A sense of duty provoked in the placebo condition should induce better performance in identifying misinformation, but we are skeptical that such an endorsement would not sufficiently heighten an individual's willingness to sign an open letter.

Both of these hypotheses are rejected. In fact, in Figure D2 we show that the placebo condition seems to have worked against the aggregate results; relative to the control

Table D1. Netting out potential source effects

	Dependent variable:		
	Combined sharing	False sharing	True sharing
	(1)	(2)	(3)
Treated vs. placebo	0.055^{***} (0.005)	-0.139^{***} (0.011)	-0.061^{***} (0.009)
Observations	630	630	630
Adjusted R^2	0.058	0.066	0.026
Note:			*p<0.1; **p<0.05; ***p<0.01

condition, the letter appears to have *increased* the reported likelihood of sharing false headlines. We view this result as an endorsement of the potential role of source effects and, again, indicative of the need for care around the potential to generate countervailing or potentially negative effects. While we lack any empirical evidence as to what may drive this result, it may indicate that Kassab's endorsement of the study in this way mitigated potential demand effects generated by though participation in an "international research initiative" as it was described to respondents. Future interventions would do well to consider how sources effects may vary when comparing a familiar voice, versus an external or independent actor.



Hypothesis	Comparison	Outcome	Direction
6a	Placebo | Control	Likelihood of **sharing false** headlines	↓
6b		Signing of **open letter** against misinformation	=

Figure D1. Pre-registered placebo hypotheses

Taken together, we argue with a high degree of confidence that the misinformation messages themselves drive much or most of our effects. Instead, viewed from two different angles, it appears not only that the endorsement of a party figure not only does not impact our results much, but that it may have counter-intuitively limited the impact of the misinformation messaging.

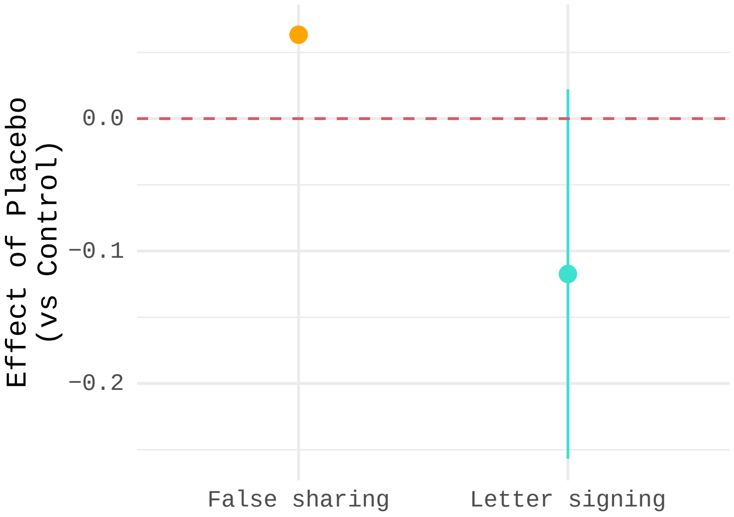


Figure D2. Effect of placebo vs control

Appendix E. Open Letter Outcome

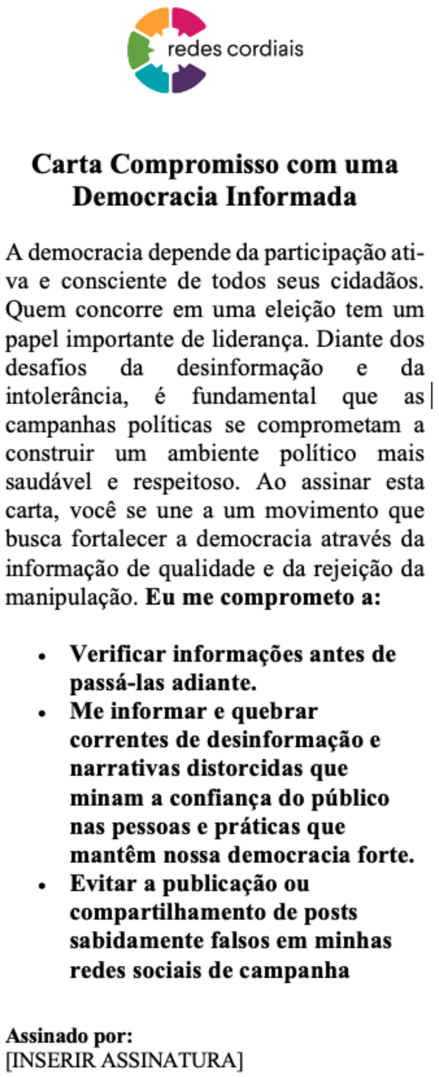
One of our main outcomes in the paper is how respondents behave when presented with the opportunity to sign and open letter. This open letter was real and the survey prompt made a directed effort to make that clear (even noting the url where respondents could later see their signature).

Specifically, it said (in Portuguese):

> Below, we present an open letter calling for an election free of fake news, produced by the organization Redes Cordiais, a non-profit and non-partisan group that fights to reduce misinformation in Brazil. Please read the letter and decide if you wish to sign it.
>
> The letter will be published at this location before the first round: redescordiais.org.br

If respondents responded they wish to sign, on the following page they were asked how they wish to their name to appear. For ethical reasons, we allowed respondents to still opt out by making this response optional. Doing so allowed us to generate the richer mapping of outcomes that appears in the main text. Here we test whether our results are consistent to using a binary signing threshold; if respondents gave a first and last name or a ballot name that with some effort might be used to identify them, this is considered a viable signature (indeed these are the only names we supplied to Redes Cordiais for the actual letter). All other responses, non-signers and non-viable signatures are coded as zero.

Swapping this binary outcome into an identical specification, we find this is roundly a null result and interpret it the same as our preferred approach in the main text; treatment shifted incentives, but perhaps did not move attitudes in a way that caused respondents to make the substantively large jump from being letter skeptics to signing.



redes cordiais

Carta Compromisso com uma Democracia Informada

A democracia depende da participação ativa e consciente de todos seus cidadãos. Quem concorre em uma eleição tem um papel importante de liderança. Diante dos desafios da desinformação e da intolerância, é fundamental que as campanhas políticas se comprometam a construir um ambiente político mais saudável e respeitoso. Ao assinar esta carta, você se une a um movimento que busca fortalecer a democracia através da informação de qualidade e da rejeição da manipulação. **Eu me comprometo a:**

- **Verificar informações antes de passá-las adiante.**
- **Me informar e quebrar correntes de desinformação e narrativas distorcidas que minam a confiança do público nas pessoas e práticas que mantêm nossa democracia forte.**
- **Evitar a publicação ou compartilhamento de posts sabidamente falsos em minhas redes sociais de campanha**

Assinado por:
[INSERIR ASSINATURA]

Figure E1. The open letter

Did respondents take the letter seriously? In the back-end of the survey instrument we recorded how long respondents spent on the letter page. The median time was 33 seconds which we find fully sufficient to assess the letter and its contents. We also recorded the number of times the respondent clicked or tapped on the page (a behavior very closely associated with interest). The median number of clicks was 9, while significant proportion clicked more than 20 times.

Though it was always implicitly the case that this is genuine real-world behavioral measure of a costly expression of values around misinformation, these data points corroborate that the letter was considered seriously by respondents.

Appendix F. Discernment by Type of Headline

We consider the possibility that the effects on our discernment outcomes are specific to the type of headline. As described in the main text, we included two true and two false headlines that are not obviously political and distinctly non-partisan, while the other six were political (four of which were plainly partisan). Both types of (mis)information of practical interest as some of the most harmful consequences of misinformation have occurred in and outside the political realm (e.g. COVID denialism and the storming of the US capitol in 2021).

We explore potential divergent effects by generating new discernment outcomes that parallel those in the main text, but calculated from each type of headline (political or non-political). We regress these using the same specification and estimation method and report results in Figure F1.

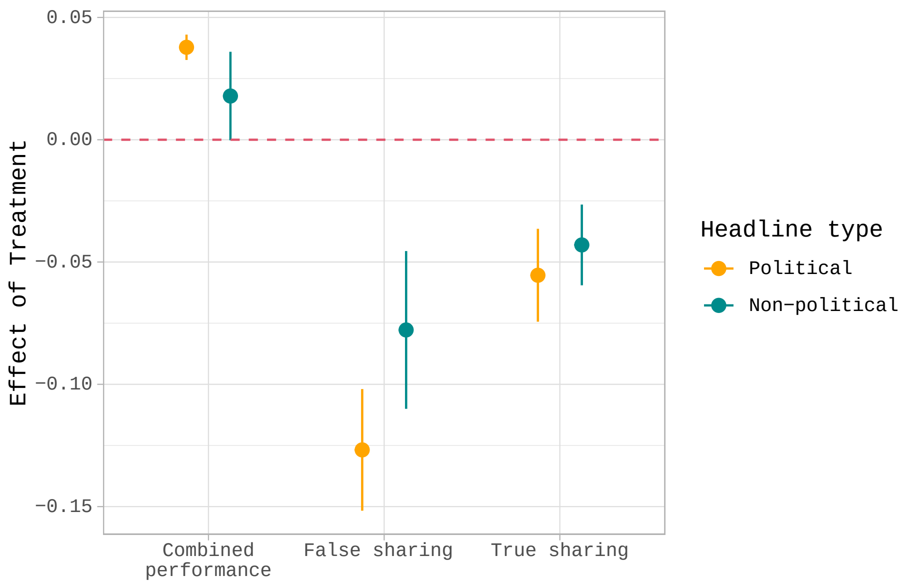


Figure F1. Discernment effects by type of headline (Political/non-political)

The magnitude of effects on the combined, false, and true sharing measures are not statistically distinguishable (referencing the 95% confidence intervals). We cautiously conclude that treatment was affective across domains and was not limited to areas where candidates should have the most expertise. This may surprising given many aspects of the intervention and survey should have favored greater social desirability bias around political themes.

Appendix G. Potential Spillovers within Municipalities

One of the few threats to inference in our experiment is the possibility that the first respondent(s) in a municipality took the survey and immediately reported its contents to other respondents before they open the survey themselves. Facially, we do not find this to be of great concern. Firstly, randomization was done at the individual level, meaning that what a "first responder" views and shares with other respondents has only a 1/6 chance of matching the version of another individual, and a 50/50 chance of matching their binary treatment-control status. Thus, while they could alert other respondents to thematic elements of the survey (misinformation and the campaign), this is true of every version of the survey.

Secondly, we consider that the first-responders could have faithfully viewed the debrief at the end of the survey, confirmed which headlines were false and shared this with subsequent respondents. We again find this to be an unlikely cause of our results– we lost many observations when respondents found the discernment task so laborious that they closed the survey at that point. Many others never clicked the final submit button on the discernment exercise, indicating a similar dynamic; attention and effort were relatively low and appear unlikely to support "passing the answer key" to other respondents.

Thirdly, we explore this empirically. We initially identify four types of spillovers where R_1 is the first respondent, R_2 is the subsequent one(s), and $T(x)$ is the treatment status of R_1 or R_2:

- Type 1: $T(R_1) = 0$ and $T(R_2) = 0$
- Type 2: $T(R_1) = 0$ and $T(R_2) = 1$
- Type 3: $T(R_1) = 1$ and $T(R_2) = 0$
- Type 4: $T(R_1) = 1$ and $T(R_2) = 1$

In the case of Type 1 and Type 3, spillovers revealing the treatment or the veracity of headlines would in theory bias against us finding effects, as it would in both cases improve the performance of the control group. In the case of Type 2, there is only the possibility of informational spillovers around the veracity of headlines. In the case of Type 4, there is the possibility of revealing (a) treatment and information around the veracity of headlines. Such spillovers have the potential to improve performance of subsequent respondents.

To show that our results are robust to accounting for spillovers, we repeat our main analysis of discernment performance among only the first few respondents in each municipality using only observations in which $T(R_1) = 0$ and $T(R_2) = 0$ and $T(R_3) = 0$, $T(R_1) = 1$ and $T(R_2) = 0$ and $T(R_3) = 0$, or there only was one respondent. Figure G1 presents these results and shows that our first two results are stable when removing those cases in which spillovers are most likely to accumulate. The chilling effect, while maintaining the same sign is no longer significant. Please, note that these estimates are likely statistically under-powered and should not alone be taken as proof positive that our spillovers play no role in our main results.

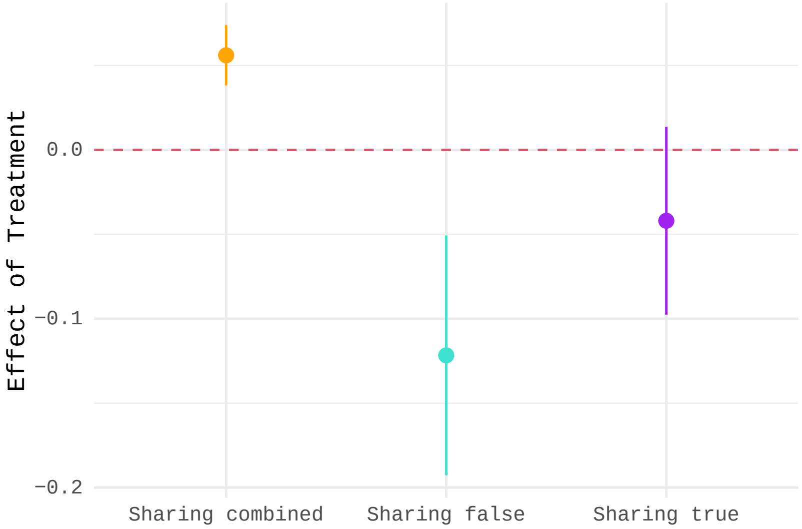


Figure G1. Estimating discernment results without spillovers

Appendix H. Balance and selection into the Sample

We asked active candidates in a heated part of their electoral campaign to spend 8-12 minutes on a survey which posed no obvious benefits to them. We have no means of knowing how many individuals received the link to our survey, though this was at least several thousand. Our best records show that around 1,900 people actually opened the survey. Roughly 66% of them completed our set of basic covariates. Advancing to the discernment task, 73% of those completed all 10 headlines (i.e. did not drop out before the the main outcomes).

This is not an exceptionally low response/attrition rate among political elites, but does raise the possibility that our findings are conditioned on some pre-treatment characteristic, therein limiting the external validity of our results. In table H1 we present our main covariates and their prevalence among those that dropped out and those that remained.

Table H1. Covariate balance of those that dropped out

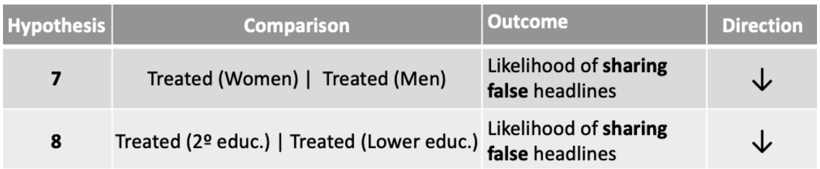
Trait	Sample Mean	Dropouts Mean	!Dropouts mean	Range
Gender (woman)	0.37	0.38	0.43	0,1
Completed Secondary	0.45	0.21	0.79	0,1
Elections competed in	2.27	2.27	2.03	1- 10
São Paolo state	0.6	0.62	0.65	0,1

For three of four variables we see relative balance between the dropout and non-dropout sub-samples. However, a key factor in dropping out was education– respondents with lower than secondary education abandoned the survey with very high frequency.

We draw two conclusions form this. It potentially signals that these individuals have a much lower tolerance/appetite for learning about and discussing electoral rules, meaning that in future efforts to induce compliance, additional measures or incentives may be required in order to ensure message reach. Secondly, we believe this makes our findings more remarkable, as people with higher education are typically more likely to be better informed. Instead, these potential ceiling effects did not manifest.

Appendix I. Two pre-registered HTE hypotheses

We pre-registered two hypotheses testing for potential heterogeneous treatment effects listed in Table I1. In each we split the sample in two, by secondary education and above (versus below) and gender. We predicted that women would respond more strongly to treatment compared to men, being less likely to share false headlines. Similarly, we expected more educated respondents to do the same relative to respondents that did not complete secondary. Our criteria for confirming both was that the 95% percent confidence intervals of each subgroup not overlap and we fail to reject the null in each case. We present these in Figure I2. We do note that our results hold within each subgroup. In this sense, it would appear that our main results are broadly applicable and not conditional on these important demographic characteristics.



Hypothesis	Comparison	Outcome	Direction
7	Treated (Women) | Treated (Men)	Likelihood of **sharing false** headlines	↓
8	Treated (2º educ.) | Treated (Lower educ.)	Likelihood of **sharing false** headlines	↓

Figure I1. Enter Caption

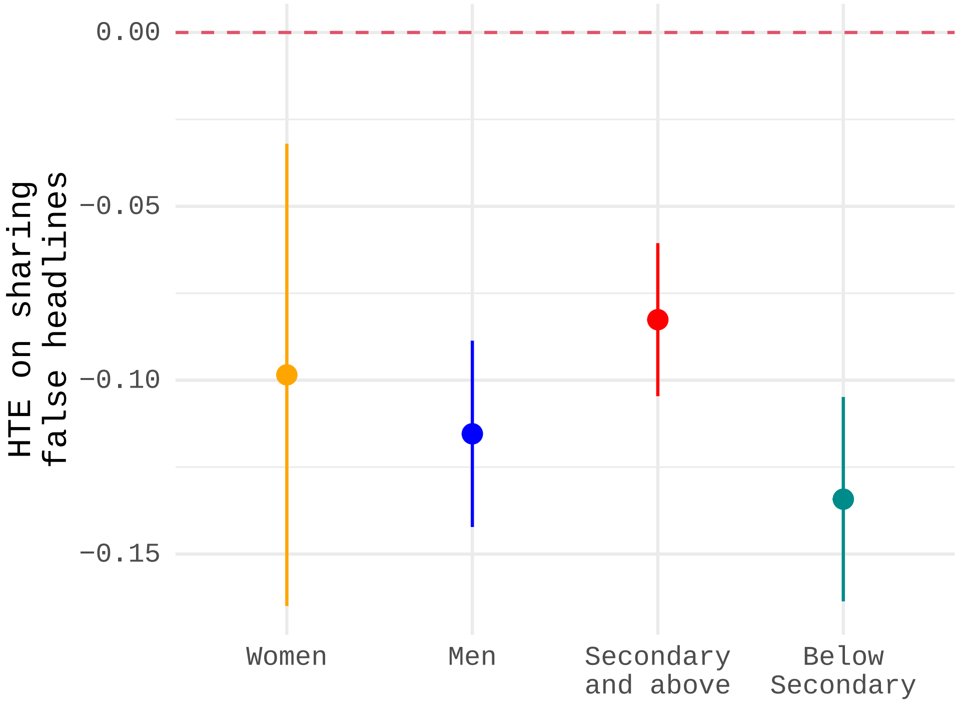


Figure I2. Testing for HTEs on sharing false

Appendix J. Main results without controls

While we believe that our pre-registered covariates are effect additions to our empirical models, in this section we re-run our first five analyses without controls, dropping gender, state, number of elections ran, and education. That is, we estimate bi-variate OLS models (outcome regressed on a binary treatment indicator) with standard errors clustered at the state-level.

With no exceptions these are all replicated, presenting identical significance at the 95%-level on coefficients of very similar magnitude.

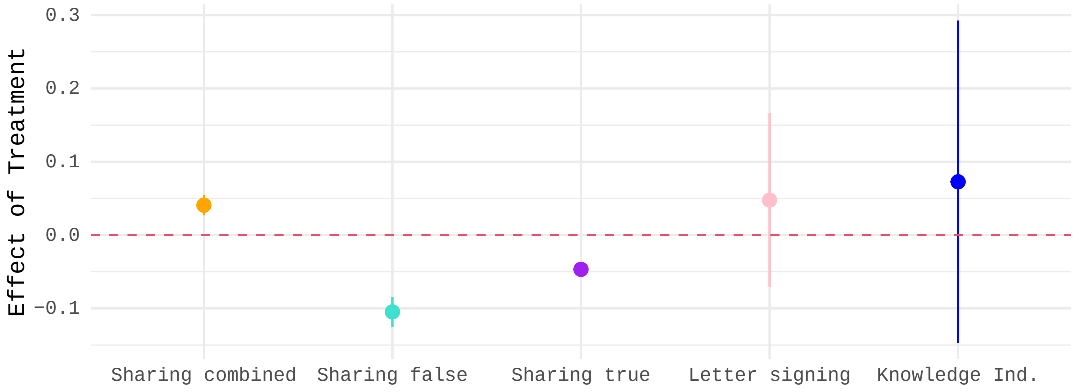


Figure J1. Re-estimating some analyses without controls.

Appendix K. Examining effectiveness of treatment by message

We provide here an exploratory analysis of experimental effects on our feature outcome (interest in sharing false headlines) by the specific message treated candidates viewed. Using the same empirical specification we estimate the false headline sharing score on a binary treatment indicator. These results are presented in Figure K1, where each coefficient represents the effect of an individual treatment.

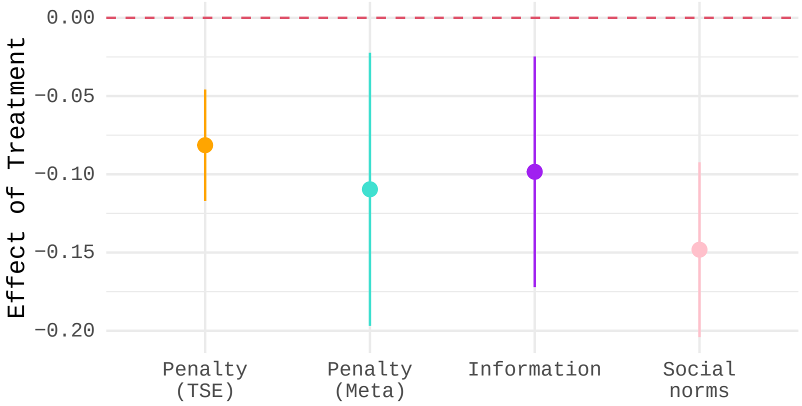


Figure K1. Effects of individual treatments

We firstly have no statistical evidence that any treatment was more effective than any other. Secondly, we cannot confirm an effect of the informational treatment by itself. This may suggest that in our pre-registered comparisons of penalty-focused messages versus context-giving messages, pooling the latter caused the the informational treatment to pull down the efficacy of the social norms treatment. Yet, we urge readers to place only limited stock in these results as statistical power drops significantly when dividing our sample to this degree.

Appendix L. Robustness to using different samples

We define our sample in two different ways in this section to test for the robustness of our main outcomes. In the first, presented in Table L1, we include respondents who had previously been excluded from experimental analyses for having spent too long on items in the discernment task, which we judged to be indicative of having left the survey. We effectively replicate these results with no exceptions. We are confident this decision was not essential to our main estimates.

Table L1. Main estimates including "Googlers"

	Dependent variable:			
	Combined sharing	False sharing	True sharing	Letter score
Assigned to Treatment	0.040*** (0.007)	−0.107*** (0.013)	−0.050*** (0.005)	0.045 (0.067)
Observations	873	873	873	937
Adjusted R^2	0.050	0.053	0.025	0.001

Note: Estimates use pre-registered controls and standard errors are clustered at the state level *p<0.1; **p<0.05; ***p<0.01

Secondly, we restrict our sample to just those who responded prior to the election, which may have changed the stakes of participation in unpredictable ways. In Table L2 we show that does not appear to be the case. Our results are fully consistent in this restricted sample.

Table L2. Main estimates excluding post-electoral submissions

	Dependent variable:			
	Combined sharing	False sharing	True sharing	Letter score
Assigned to treatment	0.036*** (0.003)	−0.106*** (0.018)	−0.055** (0.014)	0.014 (0.088)
Observations	474	474	474	474
Adjusted R^2	0.030	0.039	0.010	0.002

Note: Estimates use pre-registered controls and standard errors are clustered at the state level *p<0.1; **p<0.05; ***p<0.01