Learning to Resist Misinformation: A Field Experiment in India∗

Naman Garg† ‡ Monika Yadav‡
(Job Market Paper)

November 29, 2022
Click here for the latest version§

Abstract

Can we inoculate people against misinformation and mitigate its impact on people’s beliefs, attitudes, and behavior? We conduct a large field experiment in India with an intervention providing weekly digests containing a compilation of fact-checks of viral misinformation. In these digests, we also incorporate narrative explainers to give details and context of issues that are politically salient and consistent target of false stories. Specifically, we address misperceptions about Muslims increasingly fuelled by online misinformation. We find that familiarity with fact-checks increases people’s ability to correctly identify misinformation by eleven percentage points. However, belief in true news also decreases by four percentage points. We estimate a structural model to disentangle the two mechanisms of impact—truth discernment, which is the ability to correctly distinguish between false and true news; and skepticism, which changes the overall credulity for both false and true news. The impact is driven by an increase in both truth discernment and skepticism. Whereas skepticism increases immediately, it takes several weeks to become better at discerning truth. Finally, our intervention reduces misperceptions about Muslims, as well as leads to changes in policy attitudes and behavior. Treated individuals are less likely to support discriminatory policies and are more likely to pay for efforts to counter the harassment of inter-faith couples.

∗This experiment was preregistered in the Social Science Registry at https://doi.org/10.1257/rct.7923. We are grateful to Andrea Prat, Michael Best, Peter Bearman, Cristian Pop-Eleches and Suresh Naidu for their inputs and support. We also thank Reetika Khera, Ebonya Washington, Eric Verhoogen, Laura Boudreau, Krzysztof Zaremba, Florian Grosset, Omar Ahsan, BooKang Seol and seminar participants at Columbia University for their helpful comments and suggestions. We gratefully acknowledge financial support by the American Assembly, Columbia University; and the Program for Economic Research, Columbia University.
†Department of Economics, Columbia University
‡Department of Sociology, Columbia University
§https://namangarg.net/jmp
1 Introduction

A well-informed electorate is vital to the functioning of democracies.\textsuperscript{1} However, in recent years, the information environment of voters has transformed drastically due to the structural changes in news media markets brought by the advent of the internet.\textsuperscript{2} Even though it has been critical in expanding access to information, it has also led to a decrease in the quality of news supplied (Chen and Suen, 2022; Acemoglu et al., 2021). Due to these changes, social media has become a major source of news where people are exposed to a high amount of false and misleading content (Allcott and Gentzkow, 2017), and there is increasing concern about the difficulty in discerning the veracity of information online.\textsuperscript{3} Such high exposure to misinformation combined with people’s inability to identify it is believed to be driving widespread misperceptions about various political issues, consequently leading to changes in policy attitudes and behavior. This is seen as an important factor in not only election outcomes,\textsuperscript{4} but also in driving prejudice against minorities, thereby fueling hate crimes (Müller and Schwarz, 2020, 2021) and ethnic violence.\textsuperscript{5} Hence, it is crucial to evaluate measures that can help people to identify misinformation and thus mitigate its impact on policy attitudes and behavior.

In this paper, we conduct a large pre-registered field experiment (N = 1301) to test an intervention aimed at improving people’s ability to discern the veracity of the information they encounter online, and reduce their misperceptions about minorities. The experiment is done in the state of Uttar Pradesh in India. The region has recently seen high levels of engagement with online misinformation on social media, especially WhatsApp, which is considered to be playing a major role in election outcomes.\textsuperscript{6} A significant portion of this misinformation targets Muslims (Banaji et al., 2019), the largest religious minority in the country, and is circulated in a well-organized manner on a large scale by the Hindu-nationalist

\textsuperscript{1}See Prat and Strömberg (2013) for a review of the effects of news media on political outcomes such as voter turnout, corruption, and accountability.
\textsuperscript{2}See Zhuravskaya et al. (2020) for a review of these structural changes and other political effects of internet and social media.
\textsuperscript{3}A recent representative survey across 46 countries finds that more than half of the people (54%) “worry about identifying the difference between what is real and fake on the internet when it comes to news” (Newman et al., 2022).
\textsuperscript{4}Fake News especially became center of attention after the 2016 US election (Parkinson, 2016; Read, 2016). Facebook has placed the USA, India, and Brazil in its “Tier Zero” i.e., they are considered at the most risk in terms of misinformation impact on elections and have dedicated “war rooms” to monitor it during elections (Newton, 2021).
\textsuperscript{5}Misinformation on Facebook played a major role in anti-Rohingya violence in Myanmar (Whitten-Woodring et al., 2020). Also see Banaji et al. (2019) for an analysis of the role of anti-muslim misinformation on WhatsApp in India in mob violence.
\textsuperscript{6}See Murgia et al. (2019); Goel (2018); Ponniah (2019) for news commentary on the role of WhatsApp in Indian elections.
political party, BJP.\textsuperscript{7} This is considered responsible for driving widespread misperceptions and prejudice against Muslims (Sircar, 2021), which, in extreme cases, even leads to violence and mob lynchings.\textsuperscript{8} A major theme of such misinformation revolves around stoking fears of the Muslim population overtaking the Hindu majority by spreading falsehoods around relative trends in fertility rates of religious communities and inter-faith marriages. The incumbent BJP government in the state has even enacted and proposed various laws to address the alleged rapid changes in demographic composition. Hence, in our experiment, we focus on the effect of the intervention on factual beliefs, policy attitudes, and behavior related to these issues.

The intervention provides weekly digests with summaries of fact-checks of viral misinformation by certified fact-checkers.\textsuperscript{9} Since most misinformation revolves around certain issues, following predictable patterns and techniques of manipulation,\textsuperscript{10} familiarity with these fact-checks can help people internalize heuristics to identify false news. This can also be understood with a machine learning analogy: the intervention provides training data to individuals, so they can learn to better predict the veracity of the information they encounter on social media in the future.

The evaluation of such an intervention is crucial for several reasons. First, such interventions that inoculate against the effects of false news before its exposure (van der Linden et al., 2017) are essential in low-income countries, where encrypted messaging apps are the most popular form of social media; since the unobservability of information flows precludes interventions done during or after the exposure, such as removal or flagging of unreliable posts.\textsuperscript{11} Second, this is more robust and scalable than other inoculation interventions providing digital media literacy trainings in specific strategies to identify misinformation (Guess et al., 2020; Lewandowsky and van der Linden, 2021; Roozenbeek et al., 2022; Badrinathan, 2021). These generally require a high level of effort from individuals and the evidence on their effectiveness is mixed.\textsuperscript{12} Our intervention is more robust as a high level of cognitive effort is not required.

\textsuperscript{7}See Perrigo (2019); Arnimesh (2020) for news reports on the role of BJP in spreading misinformation. Also see Chaturvedi (2016)


\textsuperscript{9}These fact-checkers are certified by the International Fact-Checking Network (IFCN). This certification is also the criteria that Facebook uses for its third-party Fact-Checking program (https://www.facebook.com/formedia/mjip/programs/third-party-fact-checking)

\textsuperscript{10}Banaji et al. (2019) analyze a large amount of such misinformation circulating on WhatsApp in India and categorize it into a concise typology.

\textsuperscript{11}In 2017, Facebook started tagging all URLs that were fact-checked by IFCN-certified fact-checkers with an appropriate warning label. Twitter also experimented with such labels during the 2020 US elections. For impact evaluation of such measures, see Sanderson et al. (2021); Aslett et al. (2022); Clayton et al. (2020).

\textsuperscript{12}Guess et al. (2020) train people in strategies like checking the original source of information or comparing other
for it to be effective: it enables people to gain tacit knowledge of the typology and patterns of false news that they might be exposed to, allowing them to develop effective heuristics to identify it. It is also easier to scale across contexts as it does not require the design of these trainings that need to be tailored to the patterns and techniques of manipulation prevalent in a particular region.\textsuperscript{13} Third, this intervention provides a rigorous evaluation of similar efforts by news outlets and fact-checking organizations providing newsletters and dedicated sections on their website debunking viral online misinformation.\textsuperscript{14} Hence, this impact evaluation provides valuable insight into the effectiveness of such efforts in mitigating the effects of misinformation.\textsuperscript{15}

We used Facebook ads to recruit participants for the experiment. The intervention was delivered through a mobile app that was custom-built by us for this experiment. In addition to the small screening survey during recruitment, participants completed four surveys in total—one baseline survey, two follow-up surveys, and one endline survey—each at roughly an interval of three weeks. The intervention lasted ten weeks (from mid-August to October 2021), during which treated individuals received nine digests.

To measure the ability of truth assessment, we ask study participants about their belief in the truthfulness of various statements, some of which are true headlines picked from mainstream news sources, and others are viral misinformation not included in the digests. This allows us to observe the true positive rate (TPR) and false positive rate (FPR), which we use as reduced form measures of truth assessment ability. We find that the intervention substantially increases the ability to identify misinformation as it reduces FPR by eleven percentage points. However, there is also a slight reduction in TPR of about four percentage points.

To gain further insight into the mechanisms of the reduced form impacts on truth as-

\textsuperscript{13}Recall the machine learning analogy here. Training in specific strategies can be thought of as a rule based algorithm of predicting veracity, for which these rules need to be explicitly designed by the programmer. Whereas our intervention is analogous to a machine learning approach, where the algorithm acquired tacit knowledge of these patterns using training data.

\textsuperscript{14}For instance, during 2020 US elections, The New York Times had a dedicated series of articles aimed at debunking and providing information about viral online misinformation (https://www.nytimes.com/live/2020/2020-election-misinformationdistortions). It also now have a dedicated section tracking viral misinformation (https://www.ny-times.com/spotlight/disinformation)

\textsuperscript{15}Even if major news organizations supply such fact-checking summaries, there is a separate question of how much demand there is for such products. See Chopra et al. (2022) for such a demand estimation. They find that demand for newsletters increases when it includes fact-checking.
assessment ability, we develop and estimate a structural model that formalizes the two major aspects of an individual’s truth assessment: truth-discernment and skepticism. One possibility is that the intervention increases people’s ability to discern true versus false information as they become more familiar with the patterns of misinformation on social media. If this mechanism drives results, it would increase TPR and decrease FPR. Another possibility is that the intervention increases skepticism: reading these digests might lead people to update their priors about the prevalence of misinformation on social media and become less credulous overall. This mechanism would result in a decrease in both TPR and FPR. The estimation of the model allows us to disentangle the impact into these underlying mechanisms.

We find that the results are driven by both, a 19 percent increase in truth discernment ability and a 6 percent increase in skepticism. The two intermediate survey rounds also allow us to estimate the time trend of these effects. We find that the intervention immediately increases skepticism, whereas it takes much more time for individuals to learn the patterns to be able to correctly distinguish between true and false statements.

Next, we estimate the impact of the intervention on factual beliefs, policy attitudes, and behavior related to issues with a lot of misinformation around them—trends in demographic composition and the conspiracy theory of “Love Jihad”. Two of our digests incorporated narrative explainers on these issues, giving more background and context, including stories of individuals affected by the laws pertaining to them, and summarizing findings from investigations done by law enforcement agencies and independent news media organizations. Giving this information in a narrative form is crucial, as it has been shown to be more effective in changing policy attitudes and behavior than simple quantitative information and statistical facts (Bursztyn and Yang, 2022; Alesina et al., 2018a; Barrera et al., 2020).

We find that the intervention improves people’s factual beliefs: treated individuals are 13 percentage points more likely to give the correct answer about relative trends in fertility rates of Hindus versus Muslims, and are around 7 percentage points more likely to say that most stories of Love Jihad are false or misleading. The intervention also changes policy attitudes and behavior. Treated individuals are four percentage points less likely to support discriminatory policies against Muslims. We measure the impact on behavior by asking people to donate to an organization that helps inter-faith couples falsely accused of “Love Jihad”. We find that the average donation in treatment group is higher than that in control group.
We contribute to two main strands of literature. First, we contribute to the burgeoning literature testing interventions to counter the effects of fake news. Nyhan (2020) classifies these interventions based on their timing relative to the exposure to misinformation: those done after exposure, such as providing fact-checks of misinformation people had been exposed to;\textsuperscript{16} those done during the exposure, such as appropriate tagging of false/misleading information;\textsuperscript{17} and those done before exposure to inoculate against misinformation by enabling people to identify false or misleading content.\textsuperscript{18} Of these, inoculation interventions are especially important in low-income countries where unobservability of content on encrypted messaging apps precludes interventions done during or after the exposure to misinformation. We contribute to this literature by showing that a novel yet simple intervention of providing narrative summaries of fact checks can help inoculate people against misinformation. As argued above, this is more robust and scalable than existing inoculation interventions providing digital media literacy trainings, which require the design of context-specific strategies to identify misinformation and a high-level of effort from individuals. This robustness and generalizability across contexts is corroborated by findings from a parallel experiment by Bowles et al. (2022) evaluating a similar intervention in South Africa.

Our structural estimation also makes a significant methodological contribution to this literature. Guay et al. (2022) discuss the challenge of disentangling truth discernment from skepticism when measuring the impact of interventions. Most studies report changes in the difference between TPR and FPR as truth discernment. Guay et al. (2022) propose measuring changes in the ratio of TPR and FPR—or equivalently differences in log values—and show that the findings of some studies reverse using this measure. However, it is not clear why log differences are better than absolute differences, or any other arbitrary functional form. Our structural estimation provides consistent estimates of parameters of truth-discernment and skepticism that are interpretable primitives in a micro-founded model of truth assessment.

The truth discernment parameter captures the precision of the latent signal used by individuals to predict the veracity of the statements, and can be interpreted as the accuracy of their subjective assessment about statements’ veracity. The skepticism parameter is proportional to the prior odds of a statement being false, and hence can be interpreted as people’s beliefs about overall prevalence of false statements among information that they are exposed to.

\textsuperscript{16}Walter et al. (2020); Pennycook and Rand (2021) \\
\textsuperscript{17}Aslett et al. (2022); Clayton et al. (2020); Sanderson et al. (2021) \\
\textsuperscript{18}van der Linden et al. (2017); Lewandowsky and van der Linden (2021); Guess et al. (2020); Badrinathan (2021)
The second strand of literature to which we contribute focuses on measuring the extent of people’s misperceptions about various policy issues and analyzes the effect of correcting these misperceptions on policy attitudes and behavior.\textsuperscript{19} Amongst these, our study is more specifically related to the interventions aimed at improving misperceptions and policy attitudes about outgroups and minorities (see Bursztyn and Yang (2022) for a meta-analysis). One common finding from these studies is that even though corrective information is effective in changing people’s factual beliefs, this does not lead to changes in policy attitudes.\textsuperscript{20} These findings led many to posit that information delivered in a more narrative form might be more effective in moving policy attitudes; very few studies tested this conjecture and this has been identified as a fruitful avenue to explore further.\textsuperscript{21} Our experiment contributes to this literature by giving corrective information in a narrative form as part of a sustained complementary intervention aimed at training people to better identify false news. We find that such an intervention is indeed effective in changing people’s policy attitudes and behavior in a setting characterized by high levels of affective polarization.

The remainder of the paper proceeds as follows. We provide the details of experiment in Section-2, where background about the empirical context is in Section-2.1 and the experiment design is described in Section-2.2. Section-3 provides details of the data and some baseline descriptives. The reduced form results are presented in Section-4. Section-5 presents the model of truth assessment and structural estimates to disentangle the mechanisms of truth discernment and skepticism. Section-6 concludes.

2 Experiment Details

2.1 Empirical Context

Our study examines the impact of a novel intervention — regular digests with summary and fact-checks of viral misinformation — on people’s factual beliefs, policy attitudes, and behav-

\textsuperscript{19}See Haaland et al. (2022) for a recent review. This literature has explored beliefs, attitudes, and behavior related to several issues such taxation and redistribution (Alesina et al., 2018a, b; Cruces et al., 2013; Ashok et al., 2015; Hoy and Mager, 2018; Karadja et al., 2017; Kuziemko et al., 2015), immigration (Alesina et al., 2018a; Bansak et al., 2016; Barrera et al., 2020; Haaland and Roth, 2020; Hopkins et al., 2019; Jørgensen and Osmundsen, 22ed), affirmative action (Settele, 2022; Haaland and Roth, 2021), and climate change (Dechezleprêtre et al., 2022).

\textsuperscript{20}For instance, Barrera et al. (2020) conduct a survey experiment in the context of French presidential election focusing on misinformation spread by extreme-right candidate Marine Le Pen. They find that providing information about unemployment rate and gender ratio of immigrants leads to people updating their beliefs about these facts, but fails to reverse the effect of original misinformation on people’s favorability towards immigration.

\textsuperscript{21}Haaland et al. (2022) note that “Experiments systematically studying the role of stories, anecdotal evidence and narratives are still very scarce, and we believe a fruitful area for future research”. Similarly, Bursztyn and Yang (2022) note that “A small number of studies feature treatments that are qualitative and narrative in nature ... Hence, in future research, there is ample scope for incorporating qualitative treatments into the experimental design”
ior. We study these questions in the context of India, a country categorized as “Tier Zero” by Facebook — the company’s top priority group along with the US and Brazil regarding the circulation of false or misleading content. We also focus on India because of the popularity of private messaging networks and their role in disseminating misinformation in the country.

Recently, interest has emerged in exploring how private messaging services contribute to the spread of online misinformation. While Facebook and Twitter have come under intense public scrutiny regarding their efforts to check the flow of misinformation on their platforms, private messaging networks have escaped attention due to the private nature of communication on these platforms, which precludes usual interventions such as removal or tagging of misleading/false information to prevent its circulation. However, it is crucial to study these platforms for several reasons. First, they are the dominant social media platform in low-income economies. As of 2021, Whatsapp had two billion monthly active users, most of them from low-to-middle income countries in Asia, Africa, and Latin America. In India alone, Whatsapp has about 530 million active monthly users and is the main channel of communication for a large section of the population.22 Second, many predict a move in the future of social media from public networking sites to private messaging networks. But the absence of large-scale mechanisms to check the spread of misinformation on these networks provides ideal conditions for false content to flow unhindered. Thus, focusing on how Whatsapp and other private communication channels contribute to online misinformation will offer valuable insights that are more relevant for future research and policy discussions.

There are several reasons why online misinformation has become so pervasive in India and other low-income countries. One of the primary reasons revolves around the recent digitalization of these societies. The availability of cheap smartphones has enabled millions of people to go online for the first time. However, many users in these contexts lack adequate digital media literacy skills that prevent them from critically evaluating the information they get exposed to on the internet and social media (Guess et al., 2020). Second, since Whatsapp and other private messaging apps are some of the primary services these newly digitized users utilize, the way communication is structured on these messaging platforms also makes them far more consequential to the circulation of misinformation. Generally, people are less likely to doubt a piece of content when it originates from a source they trust or their in-

group network — friends, family, or other people in their close connections (Serra-Garcia and Gneezy, 2021; Pennycook and Rand, 2021). Even in cases where an individual is doubtful of the veracity of a given piece of information and can verify it, they may not want to share this new evidence for fear of being perceived as questioning the group. Further, in India particularly (but the trend is being seen worldwide including in countries in Africa and Latin America), Whatsapp is also being aggressively used by political parties to reach the electorate. Dedicated IT cells have been established by nearly all political parties in India to spread party material and messages on Whatsapp.

Online misinformation has also significantly impacted social and political discourse in India, contributing to affective polarization by pitting members of the majority group against religious and ethnic minorities. Islamophobic misinformation forms a large portion of false content in circulation at any point, especially around elections when political elites exploit false information for strategic purposes. For example, the leaders of the ruling Bhartiya Janata Party (BJP), a right-wing Hindu nationalist party, are routinely peddling emotionally charged misinformation about minorities. More recently, several leaders have organized their entire election campaigns around false and unsupported information against religious minorities. Most of this content centers around Muslims (who comprise 15% of the Indian population) overtaking the Hindu majority (who are 80% of the population) and the issue of “Love Jihad”, a term used to describe the cases of inter-faith marriages as cases of Muslim men luring and marrying Hindu women for the sole purpose to converting them to Islam. Consequently, these misinformed beliefs do not remain beliefs of the few on the periphery but come to hold significant weight in much of the larger society.

India also presents a great example to show how misinformed beliefs can distort policy opinions and change the content of legislation itself. The concern about ensuring a demographic balance among different religious communities has become a legislative agenda in many states. Some of the ruling party states have already successfully enacted laws to prevent and counter cases of “Love Jihad,” and many others are following suit.

---

23“How WhatsApp is used and misused in Africa.” The Economist, 20 Jul. 2019

24“Disinformation Spreads on WhatsApp Ahead of Brazilian Election” The New York Times, October 19, 2018

25“How BJP’s IT Cell Waged War And Won In UP” NewsLaunder, March 17, 2017

26“Kairana’ exodus’, love jihad key issues for BJP in UP polls: Adityanath” The Indian Express, February 4 2017
https://indianexpress.com/article/india/kairana-exodus-love-jihad-key-issues-for-bjp-yogi-adityanath/

27“In India, a debate over population control turns explosive” The Washington Post, August 29 2021

28“Love jihad: The Indian law threatening interfaith love.” BBC News, 8 Dec. 2020
Finally, although our study focuses on the Indian context, and thus, has some particular features that set it apart from the rest of the world, the questions we are addressing have broader implications about underlying factors that drive people’s factual beliefs and policy attitudes. We discuss these in more detail in the concluding section of the paper.

2.2 Experiment Design

Overview

Our experiment is aimed at improving how people process and evaluate information. The experiment ran from July to October 2021 in India’s northern state of Uttar Pradesh (UP). The experiment details are shown in figure 1.

We recruited participants using Facebook ads. The ads targeted research participants only from the state of UP and were shown in both English and the native language, Hindi. To prevent initial selection bias, we did not include any information in the ads about the purpose of the study as being related to misinformation or fact-checking. The ads were viewed by 2,306,157 unique individuals, of whom 23,541 clicked on them.29

Interested individuals who clicked on the ad were redirected to a short screening survey where we described the study and sought consent for participation. The respondents were told that they would need to install a mobile application on their phones that would provide them with information about fake news circulating on social media and their fact-checks. They were also informed that they would need to complete four surveys over a period of two to three months, which they can fill online on their phones as well. Additionally, we asked a few background questions about demographics, social media usage, and partisanship. To be eligible, a participant had to be a current resident in the state of UP, between the age of 18 to 60 years, primarily an Android phone user, and using WhatsApp most frequently as their private messaging app. A total of 2718 individuals who satisfied these criteria completed the screening survey and consented to participate in the study by providing their contact information.

Next, we called these 2718 individuals to verify their phone numbers, ensure their understanding of the study, and obtain a final confirmation for their participation. We were able to contact and confirm the participation of 1806 respondents at this stage. The remaining 912

29 This click-through rate of around one percent is similar to other experiments that used Facebook ads for recruitment (Allcott et al., 2020, 2022)
individuals who had initially consented to participate in the screening survey either remained unreachable (wrong number/did not take the call) or declined to continue when contacted by phone.

Subsequently, the individuals who confirmed their participation on the phone were sent links to install a placebo version of our mobile application. When opened, the placebo version of the app simply showed a message that more features would be implemented soon. We asked the study participants to download the placebo app before treatment assignment to prevent differential attrition in treatment group, as we expected downloading of the app to be a significant friction that might reduce study participation; hence even the control group had to download the app. Indeed this was the case, and of the 1806 individuals who had agreed to participate, about 465 did not install the app. In total, 1341 participants installed the placebo version of the app.

After this, the participants who installed the app, completed a comprehensive baseline survey. In this survey, we measured our respondents’ cognitive processing capacity — ability to override a quick, intuitive response to engage in further reflection — through a fairly standard test consisting of 3 questions (Frederick, 2005). We also asked our participants various questions about their perceptions regarding misinformation around them. Finally, we captured their attitudes towards minorities using different measures, and their factual knowledge and policy attitudes about certain issues relating to minority groups. 1301 individuals completed the baseline survey. We fixed the study sample at this stage and this is our final sample size.

These 1301 participants were randomized into two equal sized control and treatment groups. The randomization was stratified on the basis of baseline partisanship, religion, caste and the ability to identify concordant misinformation (misinformation that aligns with your prior beliefs and ideology). Finally, we remotely enabled the digest feature for participants assigned to the treatment group and sent a demo video explaining the new feature in their app (August 10).

After the baseline survey, participants in the treatment group started receiving regular digests on their apps. A total of nine digests were published over a period of around ten weeks.

30One more treatment was cross randomized within these groups, but due to technical issues in the app, it did not work as intended. As a result, it did not produce any first stage and the current treatment results are not affected by it. Complete details of the other intervention are in the pre-analysis plan.
The 1st digest was posted on August 16, and the last digest was published on October 22.\textsuperscript{31} Each new digest publication generated a push notification alerting the user of its availability on the app which they could read. As the experiment progressed, participants completed three surveys – two short follow-up surveys and an endline survey – each roughly at an interval of three weeks. The survey link was sent to participants on their phones via WhatsApp along with two reminders at a gap of two to three days each.

Figure-(2) shows the compliance rates with the treatment. We saw a modest amount of attrition in our sample; 1103 participants (85\% of the sample) completed the first follow-up survey, 963 participants (74\%) completed the second follow-up survey, and 1022 (79\%) participants completed the endline survey.\textsuperscript{32} We offered INR 100 (1.5 USD) to all the participants for completing the endline survey.

**Digest Treatment**

Digests were pushed to the app regularly. These digests consisted of a compilation of misinformation that went viral in the past few days with explanations about how the content was identified as false or inaccurate. We also include graphics, links to the original source of the fact-check, and other contextual information that are shown in literature to be more effective than simply providing factual information.\textsuperscript{33} The objective of this is to guide the participants through the landscape of online misinformation — the issues that saw false news floating around them, the context of these issues, and the typology and patterns of various types of viral misinformation — so they become better equipped to judge the veracity of content they come across in future. In short, through this particular form of digest, individuals are taught effective heuristics to sort and filter accurate information from false stories (Guess et al., 2020).

In two of the digests, we also incorporated short explainers around two issues that see a lot of misinformation around them and have been a major part of political discourse and even proposed legislation – trends in fertility rates of Muslims as compared to that of Hindus, and religious conversion of Hindu women by Muslim men. Besides collating the different strands of misinformation on these issues, we provided a detailed background to these issues, included not only numbers and statistics but also anecdotes and stories of individuals who

\textsuperscript{31}See Appendix A.2 for more details.

\textsuperscript{32}There is no differential attrition across treatment and control groups. See Appendix A.1 for details

\textsuperscript{33}See Appendix A.2 for a sample of digests and screenshots of the app.
were impacted by these issues and laws pertaining to them, listed instances of cases when mainstream media reporting on these issues was later discovered to be inaccurate or outright deceptive, and included findings from investigations conducted both by law enforcement agencies as well as independent media organizations.\textsuperscript{34}

The basic idea of the qualitative counter-narratives is to offer a more comprehensive picture of these issues and the misinformation surrounding them. Existing literature has found that one-off corrections are unlikely to work in such cases and has emphasized the potential of information presented in a narrative form generating higher impacts, especially on policy attitudes (Haaland and Roth, 2021).

\textbf{Survey Outcome Variables}

The follow-up and endline surveys asked questions designed to measure respondents’ ability to correctly assess the veracity of statements related to ongoing events in the socio-political discourse. To this end, we showed some statements to our study participants and asked them the following questions: whether they are familiar with the statement i.e. they have seen or heard about it somewhere; whether they think the statement is true; and the confidence in their assessment of the statement’s veracity.

In each of the three surveys, 8 headlines were presented that varied across two dimensions: accuracy (mainstream/true versus false) and political valence (right versus left). All the true headlines were published by mainstream news sources within one month of the respective survey. All the false headlines circulated on social media within one month of the respective survey and were fact-checked and labeled as false by at least one third-party fact-checking website. Also, it is important to note that we did not use any of the false news headlines that we had covered in the weekly digests in our follow-up surveys, hence the estimated impacts of intervention are about individuals’ ability to assess veracity of information in general, and not just their willingness to believe or ability to recall information provided in digests.

In our endline survey, we also sought to measure the change in factual beliefs about Muslims, a religious minority out-group, and how it consequently leads to changes in policy attitudes and behavior. To capture this we asked two questions concerning factual beliefs about Muslims. These relate to Muslim population growth and the issue of coerced religious conversion which form the bulk of misperceptions about the religious community and have

\textsuperscript{34}See Appendix A.2 for details.
dominated political discourse in UP in recent years. First, we asked our respondents what they thought about the fertility gap between Hindus and Muslims — if it is increasing, narrowing or is constant; and second, what they thought about the stories/cases of “Love Jihad” — if all/most of them are true or if all/most of them are false.

We also elicited policy attitude questions that directly relate to these issues. We asked our respondents whether they thought there is any need for a population control bill to maintain demographic balance between religious communities; and second, whether they thought there is a need to bring laws to prevent instances of “Love Jihad” in different states across the country.

Finally, we were also interested in measuring behavior change. At the end of the final follow-up survey, we gave our study participants a chance to donate some part of their survey earnings to a non-profit organization that assists interfaith couples who are harassed under the provisions of the “Love Jihad” law. These donations were deducted from the earnings that study participants received for completing the survey.

3 Baseline Descriptives

Demographic characteristics of our sample are shown in Table-1. Column-(1) shows the distribution in our sample, column-(2) shows that for Uttar Pradesh, and column-(3) for the population of social media users in India.

There are a few dimensions on which certain sections of the population are over-represented — our sample is younger, much more educated and overwhelmingly male as compared to the general population in Uttar Pradesh. However, the sample age and the education distribution is reasonably balanced compared to the population of interest of social media users. The gender imbalance in the sample is also less concerning when compared against the population of active social media users. Further we note that studies have consistently found that women lag on several determinants of agency and political participation, such as voice in making important household decisions, mobility outside home, their self-assessed leadership skills, and knowledge about how civic and political institutions work (Iyer and Mani, 2019), which might explain their lower willingness to participate in such a study.

Next, we describe some other measures collected in the baseline data.
Partisanship

We collected three measures of political partisanship. These are presented in Figure-3. First is the satisfaction with the central government, which was ruled by right-wing Bhartiya Janata party (BJP) at the time of the experiment. This was captured through a Likert scale from 1 (extremely dissatisfied) to 6 (extremely satisfied). Second is the feeling thermometer towards the Prime Minister Modi, which is captured through a Likert scale going from 1 (extremely unfavorable) to 6 (extremely favorable). Finally, we asked five policy opinion questions, the score of each of which was elicited on a Likert scale as well. Figure-3 shows the 1st principal component of these policy attitudes.

These figures show that our sample has people over the entire range of ideological distribution, allowing us to explore heterogeneity in treatment effects by partisan ideology. In all subsequent analyses, we use the satisfaction with central government to split our sample into left-leaning (anti-BJP) and right-leaning (pro-BJP).35

Truth assessment ability

We tracked the ability to discern veracity of information in all survey rounds. Figure-4 depicts the ability at baseline. The distribution of beliefs about each type of statement is shown in the figure. Two observations can be made from this figure. First, there is higher belief in veracity of true statements compared to false statements, which indicates that individuals have some ability to discern which statements are false versus true. Second, ideological concordance of statements plays an important role in beliefs. Individuals tend to have higher beliefs in statements that are concordant to their ideology, which is evidence for existence of motivated reasoning (Epley and Gilovich, 2016).

Beliefs and attitudes towards out-group

We collected data on misperceptions about Muslims — a religious minority out-group. To do so, we asked two questions measuring factual beliefs about the relative trends in fertility rates of the two religious groups — Hindus and Muslims, and the veracity of “Love Jihad” stories. According to the latest representative National Family Health Survey in India, fertility rates for both groups have declined over the past two decades, but the rate of decline has been much

35These measures are highly correlated with each other, and all results are robust to using alternative measures of partisanship.
higher in the case of Muslims (NFHS 5 Data). However, as the top left panel of Figure 5 shows, most respondents are unaware of this objective fact that feeds into the misperception that Muslims will outgrow Hindus in the near future. Similarly, investigations from independent organizations have revealed that most of the cases of “Love Jihad” are bogus.\textsuperscript{36} However, as the top right panel of Figure 5 shows that a little more than 25% left-leaning and about 75% right-leaning individuals believe these instances to be true.

We also asked study participants about their policy attitudes related to these issues. Bottom two panels of Figure 5 shows that a majority of the sample is in favor of a population bill aimed at maintaining demographic balance between different communities and implementing laws to prevent instances of “Love Jihad” across the country.

4 Results

4.1 Truth Assessment

In this section, we present reduced form results for the effect of intervention on people’s ability to assess the veracity of statements. We estimate the following OLS regression

\[ y_{ij} = \alpha + \beta d_i + \gamma \cdot X_i + \lambda_j + \epsilon_{ij} \] (1)

where \( y_{ij} \in \{0, 1\} \) denotes the assessment of individual-\( i \) about statement-\( j \), where \( y_{ij} = 1 \) denotes that the statement is assessed as being true; \( d_i \) is a dummy variable indicating the treatment assignment; \( X_i \) is a vector of individual level controls from the baseline; and \( \lambda_j \) denotes statement fixed effects. The coefficient of interest is \( \beta \). The estimation is done separately for False and True statements which gives us the impact of the intervention on FPR and TPR respectively.

The results from equation-(1) estimated for endline data are shown in Table-2. Columns (1) and (2) show the results separately for false and true statements. The intervention increases the ability to identify misinformation as FPR reduces by around 11 percentage points. However, there is also a reduction in TPR of about 4 percentage points.

There is important heterogeneity in these results by individuals’ partisanship which is shown in columns (3)-(6). For right leaning individuals, the effect on FPR is much larger.

\textsuperscript{36}‘Love Jihad’: Myth vs Reality NewsLaundry Project, 2021
at 13.1 percentage points than that for left leaning individuals which is only 7.5 percentage points. The effect on TPR for right leaning individuals is much smaller however at around 2.7 percentage points, compared to that of left leaning individuals which is 4.9 percentage points.

Looking at these results, it is not obvious what the underlying mechanisms of the impact are; how much of these results are driven by increased ability to discern true from false information, and how much is it driven by an overall increase in skepticism due to updating of priors about prevalence of false news? In Section-5 we estimate a structural model to disentangle these underlying mechanisms of impact.

4.2 Misperceptions, Policy Opinion and Behavior

We are interested in people’s ability to assess the veracity of news to see if it results in more accurate beliefs, and consequently to changes in policy attitudes and behavior. Specifically, in the context of India, we are interested in issues related to relative demographic trends of religious communities and inter-faith marriage, as these have dominated the political discourse in Uttar Pradesh and has seen a lot of misinformation around them. The intervention of weakly digests incorporated narrative explainers around these issues which provided more background around these issues, provided an overview of the relevant facts, and also included stories of people affected by these laws.

We run the following OLS regression to estimate the impact of intervention on factual beliefs, policy attitudes and behavior

\[ y_i = \alpha + \beta d_i + \gamma x_i + \varepsilon_i \]  

where \( y_i \) is the endline outcome for individual-\( i \), \( d_i \) is their treatment status and \( x_i \) is a vector of controls which includes age, caste, gender, religion and score on CRT. The outcome for factual beliefs and policy attitudes also includes baseline outcome as controls. The results are reported in Table-3

Figure-6 shows persuasion rates for these outcomes, which is a standardized measure of the causal impact allowing comparison across various studies (DellaVigna and Gentzkow, 2010). These are calculated by dividing the treatment effect with the fraction of persuadable people estimated from the control group i.e. the fraction of people in the control group who
do not have the desired outcome value. Persuasion rates can be interpreted as the percentage of people who are persuaded by being assigned to the treatment, among those that are not already persuaded.\footnote{Note that this interpretation is for persuasion rate of being assigned to treatment (having the option to read digests on the mobile phone), and not of the treatment itself (reading the digests). The persuasion rates we report is a robust lower bound of the persuasion rate of reading the digests (Jun and Lee, 2022). We argue that the persuasion rate of being assigned to the treatment is a more policy relevant metric, as it captures the net effect of both: the likelihood of having complied with the treatment, and the treatment effect on these compliers. The persuasion rate of the treatment would be relevant when there is a possibility of ensuring perfect compliance with the treatment in a real-world context, which is not the case for the treatment considered in this experiment.}

We can see that the intervention results in more accurate factual beliefs. People in the treatment groups are 13 percentage points more likely to give correct answer about fertility gap narrowing between Hindus and Muslims in India, and 7 percentage more likely to say that all or most stories about love jihad are false. This corresponds to a persuasion rate of 17 percent and 13 percent respectively.

The intervention also leads to changes in policy attitudes and behavior. Treated individuals are 4 percentage points less likely to support discriminatory laws against muslims regarding population control and ‘love jihad’, corresponding to a persuasion rate of 5 percent. We measure changes in behavior by asking study participants to donate a portion of their survey earnings of rupees 100 ($\approx 1.5$ USD) to an NGO that help inter-faith couples false accused of ‘love jihad’. The treatment increases the average donation amount given to the NGO.

The finding that the impact on policy attitudes is much smaller than that on beliefs suggests that these beliefs are not the sole factor driving policy attitudes. In fact, it is possible that the narratives directly affect policy attitudes, independent of any effect of factual beliefs (Alesina et al., 2018a; Eliaz and Spiegler, 2020; Bénabou et al., 2020; Schwartzstein and Sunderam, 2021). Future work should try to disentangle the relative importance of beliefs versus these other factors as determinants of policy attitudes.

5 Structural Estimation

The reduced form results in Section-4.1 for impact of intervention on truth assessment showed that it increased the ability to identify misinformation by 11 percentage points, but also reduces beliefs in true statements by 4 percentage points. Since the reduction in FPR is much larger than TPR, this suggest that the intervention does increase the truth discernment ability i.e. the ability to distinguish between true and false statements. However, it also seems that
the intervention increases overall skepticism as the belief in both kind of statement decreases.

In this section we present a micro-founded model that formalizes these two possible mechanisms: truth discernment, and skepticism. The parameters corresponding to these two mechanisms are primitives of the underlying model and have intuitive interpretations. The truth discernment parameter in the model captures the precision of the latent signal used by individuals to predict the veracity of the statements and can be interpreted as the accuracy of their subjective assessment about statements’ veracity. The skepticism parameter is proportional to the prior odds of a statement being false and hence can be interpreted as people’s beliefs about overall prevalence of false statements among information that they get exposed to. We derive a likelihood function and calculate maximum likelihood estimates of the parameters separately for control and treatment groups, and hence disentangle the mechanisms of impact.

5.1 Model of Truth Assessment

We formalize the process of truth assessment of a statement as an optimal decision making problem under uncertainty with state contingent payoffs for various actions (Blackwell, 1953). The state of the world is the truth value of statement and the action is the individual’s prediction about the state. Let \( t_j \in \{0, 1\} \) denote the truth of statement-\( j \), where \( t_j = 1 \) signifies that the statement is true; and let \( a_{ij} \in \{0, 1\} \) denote individual-\( i \)’s prediction. The utility function \( u : \{0, 1\}^2 \to \mathbb{R} \) is given by

\[
    u(t_j, a_{ij}) = \begin{cases}
        b & t_j = 1, a_{ij} = 1 \\
        -c & t_j = 0, a_{ij} = 1
    \end{cases}
\]

i.e. \( b \) is the utility from correctly predicting a true statement and \( c \) is the disutility of predicting a false statement as true.\(^{38} \) Without loss of generality, normalize \( b = 1 \).

We follow Angelucci and Prat (2020) and introduce a latent signal \( s_{ij} \in \mathbb{R} \) that individual-\( i \) observes about \( t_j \) to make their prediction. This signal can be thought of as their subjective assessment about the veracity of the statement. More informative signal means more accu-

---

\(^{38}\) These utility parameters can also be interpreted in terms of value of information. Suppose the true and false statement are signals about some policy relevant state of the world, where true statements are informative about the state, whereas false statements are independent of the state and systematically biased. Then \( b \) can be thought of as value of information of informative signals (Blackwell, 1953), and \( c \) can be thought of loss of utility from updating beliefs from a biased uninformative signal.
rate assessment. Let $f(s_{ij}|t_j)$ denote the conditional probability distribution of signal, and
$F(s_{ij}|t_j)$ denote the conditional CDF. Assume, without loss of generality that the signal is
affiliated with the state i.e $\frac{f(s|t=1)}{f(s|t=0)}$ is an increasing function of $s$. This guarantees that the
posterior probability that statement is true is a monotonic function of signal realization. Also
assume that the signal has support on the real line.

Then the optimization problem for the individual is to choose a prediction policy $a : \mathbb{R} \rightarrow \{0,1\}$ that gives prediction as a function of signal realization. Let $A \subset \mathbb{R} \times \{0,1\}$ be the set
of all such prediction policies. Then for any $a \in A$, we can define True Positive Rate (TPR) and False Positive Rate (FPR) as

$$TPR(a) = Pr(a(s_{ij}) = 1|t_j = 1) \quad (3)$$
$$FPR(a) = Pr(a(s_{ij}) = 1|t_j = 0) \quad (4)$$

Let $\pi$ denote the prior probability that statement-$j$ is true. Then the ex-ante expected utility
is given by

$$U(a) = \pi \cdot TPR - (1 - \pi) \cdot c \cdot FPR \quad (5)$$

Hence the optimal prediction policy is a solution of maximization problem $\max_{a \in A} U(a)$
subject to constraints (3), (4). It is easy to see that the optimal policy can be completely
characterized\textsuperscript{39} by a threshold value $k \in \mathbb{R}$ such that

$$a(s) = \begin{cases} 
1 & s \geq k \\
0 & s < k 
\end{cases}$$

Then FPR and TPR can be written as

$$TPR(k) = 1 - F(k|1) \quad (6)$$
$$FPR(k) = 1 - F(k|0) \quad (7)$$

Plugging these in the utility function we get the FOC for the optimization problem

$$\frac{d(TPR)}{d(FPR)} = \frac{f(k|1)}{f(k|0)} = c \left( \frac{1 - \pi}{\pi} \right) \quad (8)$$

\textsuperscript{39}This follows from posterior porbablity being an increasing function of signal realizat
The optimization problem can also be visualized in terms of Receiver Operator Characteristic (ROC) curves. Eliminating $k$ from equations-(6),(7), we obtain the equation for ROC curve given by

$$TPR = 1 - F \left( F^{-1} (1 - FPR\mid 0) \mid 1 \right)$$  \hspace{1cm} (9)

It is easy to verify that equation-9 is a ROC curve i.e. (0, 0) and (1, 1) lie on the curve corresponding to $k = \infty$ and $k = -\infty$ respectively; and $\frac{d(TPR)}{d(FPR)} = \frac{f(k \mid 1)}{f(k \mid 0)} > 0$. Thus, a given signal structure implies a ROC curve, and hence characterizes the feasibility set for the prediction problem. A more informative signal will have a better ROC curve and larger feasibility set.\footnote{This follows directly from payoff richness characterization of information (Blackwell, 1953).}

Figure-7 shows a typical ROC curve.

The threshold value $k$ paremetrizes various points on the ROC curve. When $k = \infty$ the individual always predicts the statement to be false, and hence TPR=FPR=0. This person can be thought of as radical skeptic. As $k$ decreases, statements become more likely to be predicted as true. The optimal value of $k$ is chosen to optimize the trade-off between correctly predicting true and false statements. This tradeoff is captured by the slope of the indifference curves given by expected utility function in equation-5 which are straight lines with slope $c \left( 1 - \pi \right)$. The optimum $k$ is chosen such that the indifference curve is tangent to ROC curve which is captured in FOC condition-(8).

The slope of indifference curve $c \left( 1 - \pi \right)$ captures the ratio of marginal benefit from correctly predicting false statements to that of true statements. As $\pi$, the prior probability of statement being true, increases, the individual has higher ex-ante incentive to correctly predict true statements. Indeed this decreases slope of indifference curve and the optimal point is higher up the ROC curve corresponding to higher ex-ante probability of any given statement being predicted as true. Similarly a decrease in $c$ signifies higher incentive to correctly predict true statements, which results in optimal point being higher up the ROC curve.

Hereafter, I’ll refer to $c \left( 1 - \pi \right)$ as the skepticism parameter denoted by $r$; higher values of skepticism parameter leads to lower probability of a statement being predicted as true.
5.2 Maximum Likelihood Estimation

For estimation I’ll assume that the signals are normally distributed with precision $\rho$, i.e.

$$s \sim N \left( t, \frac{1}{\rho^2} \right)$$  \hspace{1cm} (10)

Then we have that $f(s|t) = \phi(\rho(s-t))$ and $F(s|t) = \Phi(\rho(s-t))$ where $\phi(.)$ and $\Phi(.)$ are pdf and cdf respectively of standard normal distribution. The TPR and FPR is given by

$$TPR = 1 - \Phi(\rho(k-1))$$
$$FPR = 1 - \Phi(\rho k)$$

The ROC curve is

$$TPR = 1 - \Phi(\Phi^{-1}(1-FPR) - \rho)$$ \hspace{1cm} (11)

The optimal threshold value $k$ is given by FOC-(8) is

$$k = \frac{2 \log (r)}{\rho^2} + \frac{1}{2}$$ \hspace{1cm} (12)

Now to derive the likelihood function, the probability of a statement-$j$ being predicted as true by individual-$i$ can be written as

$$Pr(a_{ij} = 1) = 1 - \Phi(\rho(k-t_j))$$

where $a_{ij}$ is the prediction and $t_j$ is the actual value. Then assuming that signals are conditionally independent across individuals and statements, the log-likelihood function is:

$$\mathcal{L}(a, t; \rho, k) = \sum_{i,j} a_{ij} \log (1 - \Phi(\rho(k-t_j))) + (1 - a_{ij}) \log (\Phi(\rho(k-t_j)))$$ \hspace{1cm} (13)

where $a, t$ denote stacked vectors of observations. Hence $\rho, k$ can be estimated using MLE, and plugging these values in equation-(12) we can get estimate of skepticism parameter $r$.

Before we present the MLE estimates, a few remarks about interpretation of these parameters are in order. Note that $\rho$ captures the informativeness of the latent signal and hence is a measure of how good a person is at discerning whether a statement is true. Hereafter,
I’ll call $\rho$ the *truth-d discernment* parameter. The skepticism parameter $r$ is a combination of two primitives of the model: the prior probability $\pi$ of a statement being true, and the utility function. Note that even though the model identifies $r$, the components $\pi$ and utility function parameters are not identified separately in a given sample. I assume that the intervention does not change the utility function. It can however change the prior of statements being true. Hence, differences in skepticism parameter across control and treatment groups can be attributed to differences in prior probability of statement being true i.e. changes in skepticism due to the interventions are driven by changes in beliefs about prevalence of true statements.

This structural estimation is a significant improvement over existing reduced form approaches to estimating impact of interventions on truth discernment. Most prior studies report use the difference in TPR and FPR as the measure of discernment (Guess et al., 2020; Roozenbeek et al., 2022). Guay et al. (2022) propose an alternative measure of using ratio of TPR and FPR and show that it gives different results, sometimes even reversing the sign of impact of interventions. It is not clear why either of these measure, or any other such arbitrary function form, should be preferred over other. Our method is better as it provides consistent estimates for parameters of truth-d discernment and skepticism that are interpretable primitives in a micro-founded model of truth assessment.\(^{41}\)

### 5.3 Estimation Results

Table-4 shows the estimates of parameters of the models by maximizing the likelihood function in equation-(13). To get the treatment effect on these parameters, the parameters are estimated separately for control and treatment groups, and the difference is reported in the Table. The treatment increases the truth discernment by 19 percent and increases the skepticism by 6 percent.

To see heterogeneity of impact by ideology, we also show the estimates separately for left leaning and right leaning individuals. For right-leaning individuals, the effect is driven by an increase in both truth discernment and skepticism, whereas for left-leaning individuals there is no increase in truth discernment and effect is only driven by increased skepticism.

These results show that the treatment is quite effective in increasing the ability to accurately assess the veracity of the headlines, especially for right-leaning individuals. This

\(^{41}\)See Apendix-C for a detailed comparison of these reduced form measures and the structural estimation.
means that reading these digests increases people's ability to discern truth of news that they come across. However, there is an important trade-off to consider of higher skepticism, as people update their priors about prevalence of false news after reading these digests and become less credulous overall. It is difficult to say whether this increase in skepticism is desirable as it would require some notion of a correct level of skepticism.\textsuperscript{42} We currently do not take a stand on this and leave it for future research, but the fact that any such treatment aimed at increasing truth discernment ability also produces an effect on higher skepticism in noteworthy.

In addition to the baseline and endline surveys, we also did two intermediate follow-up surveys in which we tracked truth assessment abilities. This allows us to observe the time trends in the treatment effects. These are shown in Figure-8. It can be seen that the treatment increases skepticism almost immediately. As people start reading these digests, they update their prior about the prevalence of misinformation and become more skeptical as a result. However, it takes more time to actually learn the patterns to be able to correctly distinguish between true and false statements. It takes full two months for them to become better at truth discernment as the estimate of truth discernment is positive and significant only in the endline survey.

6 Conclusion

We show that familiarity with fact-checks can help people learn effective heuristics to discern the veracity of the information on social media. Our intervention is more robust than existing interventions which provide digital media literacy trainings in specific strategies to identify misinformation and require a high level of cognitive effort from individuals. It is also more scalable as it does not require the design of these trainings that need to be tailored to the patterns and techniques of manipulation prevalent in a specific context. Indeed, this generalizability across various contexts is demonstrated by similar findings from a parallel experiment by Bowles et al. (2022) in South Africa. Some news outlets have already been providing similar summaries and newsletters tracking viral misinformation on social media.\textsuperscript{14}

\textsuperscript{42}In principal, if one could observe the universe of information that individuals are exposed to, the ratio of true information among these would constitute the correct prior. Notwithstanding the onerous data requirements for such an exercise, there would still be an additional challenge of disentangling how the prior probability and utility function parameters from estimates skepticism parameter $r$. Even though the changes in prior probability are identified by the model (assuming utility in unchanged), its levels cannot be estimated. This problem of priors and utility function parameters producing observationally equivalent variation in posteriors has been noted more generally in other contexts as well (Little, 2021)
and our experiment provides a rigorous evaluation of such efforts. As these become more
common, and to the extent there is sufficient demand for such fact-checking (Chopra et al.,
2022), this is a potentially promising approach to inoculate the electorate against the effects
of misinformation on factual beliefs.

Interventions aimed at improving truth discernment ability also tend to simultaneously
affect overall skepticism. We develop and estimate a micro-founded model of truth assess-
ment to disentangle these two mechanisms of impact, which is a significant improvement over
existing reduced form approaches of measuring discernment. We find that as the individu-
als start reading these digests their skepticism rises immediately, whereas it takes much more
time to become better at truth discernment. Future research should look at the welfare impli-
cations and tradeoffs of these two mechanisms, and investigate aspects of the intervention
that determine their relative impact.

The interest in the effect of misinformation on beliefs stems from concerns about its con-
sequent impact on attitudes and behavior. Our intervention, which incorporates narrative
explainers about the relevant issues, changes attitudes and behavior as well. However, the
effect on policy attitudes is much smaller than that on beliefs, suggesting that the beliefs are
not the sole determinant of attitudes. It is even possible that the narratives affect attitudes
directly, completely independent of any impact of beliefs (Alesina et al., 2018a). Recent work
formalizes various such notions of ‘content free’ narratives that can affect attitudes with-
out containing any information about the state of the world: for instance by changing the
worldview of people, through moral reasoning, by trigerring certain emotions such as em-
pathy, or by changing the salience of personal identities. Experiments cross-randomizing

43See Appendix-C for a detailed discussion.
44Even though higher truth discernment ability is unambigiously better, that is not the case for skepticism. A
very high level of skepticism can result in rejection of too many true statements as false, whereas a very low level of
skepticism can result in too many false statement being accepted as true. The optimal level of skepticism corresponds
to the correct prior about the overall prevalence of false statements that one is exposed to. Future work can benchmark
estimated priors with relevant empirical data to gain insight into these welfare implications.
45Eliaz and Spiegler (2020) formalize narratives as causal models of the world (represented by directed acyclic
graphs), and analyze how competing narratives can ause political polarization. Schwartzstein and Sunderam (2021)
formalize narratives as the likelihood function from which the observed signals are drawn, and show how narratives can
persuade by changing the likelihood function used to update the priors for given signal realizations.
46Bénabou et al. (2020) formalize narratives as arguments about moral significance of one’s actions which can affect
attitudes and behavior by changing the perceived private and/or social payoffs in the utility function.
47Haidt (2001, 2007, 2012) argues that intuition is the main driving force for many attitudes, especially about
charged topics in politics and religion. These intuitions are mainly driven by emotions and rational reasoning is a
post-hoc process undertaken to justify these intuitive attitudes.
48Gerber et al. (2010) conduct a field experiment to show that strengthening party identity causes shift in political
attitudes and behavior. Mutz and Kim (2017); Mutz et al. (2021) show how racial identity affects preferences for trade
policy and displays in-group favoritism. Broockman and Kalla (2016); Kalla and Broockman (2020) show that non-
judgmental exchange of personal narratives that does not threaten people’s self-image is effective in durably reducing
transphobia.
various features of narratives can provide useful insight into these potential mechanisms of impact of narratives, and this is a fruitful area for future research.

Insights into the relative importance of factual beliefs versus these other factors as determinants of policy attitudes also have critical implications for our understanding of misinformation more broadly. Most research on online misinformation has been based on the premise of misinformed policy attitudes: that certain policy attitudes are merely misinformed and correcting factual beliefs can change these attitudes. Indeed, if beliefs are the major determinant of policy attitudes, more accurate factual beliefs will result in higher political accountability as people’s attitudes more closely reflect the state of the world. However, if misinformation affects policy attitudes by triggering certain emotions or by changing the salience of personal identities, then merely correcting beliefs will not be effective. Perhaps online misinformation is more of a side effect, and not the root cause, of the processes causing political polarization: as social media diminishes the gatekeeping role of mainstream media, politicians use their direct access to voters to trigger certain emotions and change the salience of various personal identities by spreading biased and false information, thus polarizing their attitudes. In that case more attention should be paid to the higher agenda setting power of politicians derived from direct access to voters, as the underlying structural factor contributing to political polarization in the age of social media.

---

49 A large class of theoretical models of electoral accountability show that more accurate beliefs of the electorate about state of the world results in more effort exerted by politicians in equilibrium (Duggan and Martinelli, 2017).

50 Bonomi et al. (2021); Grossman and Helpman (2021) formalize social identity theory and show how switches in identity due to dominance of cultural or ethnic issues in political discourse can lead to polarization of policy preferences on trade and redistribution.
References


### Tables

<table>
<thead>
<tr>
<th></th>
<th>Sample (UP)</th>
<th>Population (UP)</th>
<th>Social Media Users (India)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 - 25</td>
<td>17.93</td>
<td>22.71</td>
<td>20</td>
</tr>
<tr>
<td>26 - 35</td>
<td>51.47</td>
<td>28.91</td>
<td>31</td>
</tr>
<tr>
<td>36 - 45</td>
<td>20.09</td>
<td>22.69</td>
<td>16</td>
</tr>
<tr>
<td>46 - 55</td>
<td>8.50</td>
<td>14.32</td>
<td>7</td>
</tr>
<tr>
<td>56 - 65</td>
<td>2.01</td>
<td>11.32</td>
<td>4</td>
</tr>
<tr>
<td><strong>Religion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hindu</td>
<td>85.01</td>
<td>79.73</td>
<td>80</td>
</tr>
<tr>
<td>Muslim</td>
<td>9.66</td>
<td>19.26</td>
<td>12</td>
</tr>
<tr>
<td>Other</td>
<td>5.33</td>
<td>1.01</td>
<td>8</td>
</tr>
<tr>
<td><strong>Caste</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General</td>
<td>54.71</td>
<td>19</td>
<td>29</td>
</tr>
<tr>
<td>OBC</td>
<td>32.61</td>
<td>39</td>
<td>38</td>
</tr>
<tr>
<td>SC</td>
<td>11.05</td>
<td>20.8</td>
<td>13</td>
</tr>
<tr>
<td>ST</td>
<td>0.77</td>
<td>0.57</td>
<td>5</td>
</tr>
<tr>
<td>Other</td>
<td>0.85</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-literate</td>
<td>0</td>
<td>32.3</td>
<td>2</td>
</tr>
<tr>
<td>Primary School</td>
<td>0.77</td>
<td>37.7</td>
<td>7</td>
</tr>
<tr>
<td>Secondary School</td>
<td>15.84</td>
<td>24.7</td>
<td>22</td>
</tr>
<tr>
<td>College Degree</td>
<td>83.38</td>
<td>4.96</td>
<td>70</td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td>93.43</td>
<td>52.28</td>
<td>66</td>
</tr>
</tbody>
</table>

| N                        | 1301        | 200 million     | 460 million               |

*Notes: Table show the composition (in percentage terms) of study sample, Uttar Pradesh Population and Population of Social Media users in India. Population estimates are from Census 2011. The data on social media users come from a representative nation-wide survey conducted by CSDS (2019) conducted in 2018-2019. OBC - Other Backward Castes; SC - Scheduled Castes, ST - Scheduled Tribes. Indian Govt does not enumerate caste-wise population other than SCs and STs. The proportion of general castes and OBCs is not accurate and is based on various estimates.*
<table>
<thead>
<tr>
<th></th>
<th>False Statements</th>
<th>True Statements</th>
<th>False Statements</th>
<th>True Statements</th>
<th>False Statements</th>
<th>True Statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>-0.107***</td>
<td>-0.038**</td>
<td>-0.131***</td>
<td>-0.027</td>
<td>-0.075***</td>
<td>-0.049</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.015)</td>
<td>(0.021)</td>
<td>(0.018)</td>
<td>(0.022)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>Control Mean</td>
<td>0.480</td>
<td>0.758</td>
<td>0.449</td>
<td>0.745</td>
<td>0.500</td>
<td>0.766</td>
</tr>
<tr>
<td>Observations</td>
<td>4088</td>
<td>4088</td>
<td>2504</td>
<td>2504</td>
<td>1584</td>
<td>1584</td>
</tr>
</tbody>
</table>

**Notes:** Table shows the impact of intervention on truth assessment ability. Regression coefficients are from estimation of equation (1) on endline data, where outcome variable is a dummy indicating whether individual believes the statement to be true. Demographic controls include age, caste, gender, religion and, score on Cognitive Reflection Test (CRT). All specification include statement fixed effects. Results for full sample are shown in column (1) and (2) for false and true statements respectively. We can see that the intervention decreases FPR by 10.7 percentage points and decreases TPR by 3.8 percentage points. Columns (3), (4) show these results for only right leaning individuals, and columns (5), (6) shows these for left-leaning individuals. Robust standard errors clustered at the level of individuals are in parentheses.

*** p < 0.001; ** p < 0.01; * p < 0.05.
Table 3: Treatment Effect on Beliefs, Attitudes and Behavior

<table>
<thead>
<tr>
<th>Panel</th>
<th>Treatment Effect</th>
<th>Control Mean</th>
<th>Persuasion Rate</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A. Factual Beliefs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fertility Gap Narrowing</td>
<td>0.131*** (0.027)</td>
<td>0.215</td>
<td>17%</td>
<td>1022</td>
</tr>
<tr>
<td>Is Love Jihad Fake</td>
<td>0.068** (0.026)</td>
<td>0.465</td>
<td>13%</td>
<td>1022</td>
</tr>
<tr>
<td>Beliefs Index</td>
<td>0.098*** (0.020)</td>
<td>0.340</td>
<td></td>
<td>1022</td>
</tr>
<tr>
<td>Panel B. Policy Attitudes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population Control Law Need</td>
<td>-0.038* (0.020)</td>
<td>0.842</td>
<td>5%</td>
<td>1022</td>
</tr>
<tr>
<td>Love Jihad Law Need</td>
<td>-0.040* (0.024)</td>
<td>0.769</td>
<td>5%</td>
<td>1022</td>
</tr>
<tr>
<td>Attitudes Index</td>
<td>-0.039** (0.017)</td>
<td>0.806</td>
<td></td>
<td>1022</td>
</tr>
<tr>
<td>Panel C. Behavior</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Donation</td>
<td>6.310** (2.595)</td>
<td>38.519</td>
<td></td>
<td>1022</td>
</tr>
</tbody>
</table>

*** p < 0.01; ** p < 0.05; * p < 0.1. Robust standard errors in parentheses.

Notes: Table shows treatment effects by regressing endline outcomes regarding beliefs, attitudes and behavior on treatment status and controls. Control variables included in all regressions are age, caste, gender, religion and score on CRT. Regressions for factual beliefs and policy attitudes also include corresponding baseline outcomes as control.

Panel A shows results for factual beliefs regarding policy issues. Outcome Fertility Gap Narrowing as a dummy that denotes whether the respondent correctly says that the fertility gap between Hindus and Muslims is narrowing. Outcome Is Love Jihad Fake denotes whether the respondent thinks that all or most of the stories alleging love jihad are false. Beliefs Index is an average of the the two beliefs outcomes.

Panel B shows results regarding support for discriminatory policies against Muslims. Population Control Law Need denotes support for draft law to control population which included maintanence of balance between population to religious communities amongst its explicitly stated objectives. Love Jihad Law Need measures support for the “Love Jihad” law that was passed earlier and has been used by vigilantes and police to harass inter-faith couples in India. Attitudes Index is an average of the above two measures.

Panel C. shows results for behavior. Donation outcome measures donation given by respondents out of survey earnings (Rs 100) to NGO helping interfaith couples harassed by police and vigilantes under love jihad laws.

Persuasion rates are calculated according to DellaVigna and Gentzkow (2010) by dividing the treatment effect with fraction of persuadable people as estimated from the control group. These can be interpreted as the percentage of people who are persuaded by being assigned to the treatment (to change beliefs and attitudes), among those that are not already persuaded.
Table 4: Estimates from the model

<table>
<thead>
<tr>
<th></th>
<th>Truth-Discernment</th>
<th></th>
<th>Skepticism</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control $\rho_c$</td>
<td>Treatment $\rho_t$</td>
<td>Difference $\log(\rho_t/\rho_c)$</td>
<td>Control $r_c$</td>
<td>Treatment $r_t$</td>
</tr>
<tr>
<td>Full Sample</td>
<td>0.748</td>
<td>0.904</td>
<td>0.189**</td>
<td>0.886</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.041)</td>
<td>(0.071)</td>
<td>(0.009)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Left Leaning</td>
<td>0.786</td>
<td>0.811</td>
<td>0.031**</td>
<td>0.901</td>
<td>0.962</td>
</tr>
<tr>
<td>Individuals</td>
<td>(0.065)</td>
<td>(0.066)</td>
<td>(0.116)</td>
<td>(0.014)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Right Leaning</td>
<td>0.725</td>
<td>0.965</td>
<td>0.286**</td>
<td>0.877</td>
<td>0.929</td>
</tr>
<tr>
<td>Individuals</td>
<td>(0.052)</td>
<td>(0.053)</td>
<td>(0.091)</td>
<td>(0.012)</td>
<td>(0.013)</td>
</tr>
</tbody>
</table>

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ (stars reported only for “Difference” column).

Notes: Table shows MLE estimates for the truth discernment parameter ($\rho$) and the skepticism parameter ($r$). The estimation is done separately for control and treatment groups, and the differences in parameter values are reported. $\rho$ denotes the accuracy of subjective assessment about veracity of the statements (equation-10), and hence interpreted as truth discernment. $r$ is proportional to prior odds of a statement being false, and hence changes in skepticism parameter are driven by changes in beliefs about overall prevalence of false statements.

First row shows estimates for full samples. Subsequent rows present estimates separately for left leaning and right leaning individuals. Standard errors in parentheses.
Figure 1: Experiment Design

*Notes:* Figure shows a schematic of the experiment design. The recruitment was done through FB Ads. Interested participants clicked on the ad and fill out a short screening survey. We then contacted them on phone to verify contact details, explain the study and confirm participation. All these individuals were asked to download a placebo version of the mobile app built by us for the purpose of this study. The placebo version did not have any features implemented in it. We then did our baseline survey with everyone who successfully downloaded the placebo app. The 1301 individuals who downloaded the app and filled the baseline survey constitute our study sample. These were then randomly assigned to treatment and control groups and the digest feature was rolled out for those assigned to treatment groups, and they started receiving weekly digests. We did two additional small follow-up surveys in between and a comprehensive endline survey at the end.
Figure 2: Compliance Rates with the treatment

Notes: The distribution of total number of digest read by individuals is reported among the treatment group i.e. in the treatment group what proportion of individuals read a given number of digests.
Figure 3: Political Partisanship

Notes: Figure shows baseline distribution of political ideology of the sample using different measures. Figure-(a) shows satisfaction with central government that was ruled by BJP at the time of the experiment. Figure-(b) shows feeling thermometer towards PM Modi. Figure-(c) shows distribution of 1st principal component of five policy attitudes in the baseline survey.
Figure 4: Baseline Truth Assessment Ability

Notes: Figure shows truth assessment at baseline. For each statement we asked respondents to report whether they think the statement is true or not and how confident they are in their assessment. The distribution is reported separately for true versus false statements, and for statements that concordant versus discordant to person’s. Two important observations can be made. First, there is higher belief in true statements compared to false statements, which means people are able to distinguish between false and true statements. Second, individuals tend to have higher beliefs in statements that are concordant to their ideology for both false and true statements.
Figure 5: Factual beliefs and policy attitudes about issues related to Muslims

Notes: Figure shows factual beliefs and policy attitudes at baseline for issues with a lot of misinformation around them. Each measure is reported separately for left-leaning and right-leaning individuals. Figure-(a) shows distribution of beliefs about relative decline of fertility rates. The correct answer is that fertility rates declined more for Muslims than Muslims. Figure-(b) shows beliefs about veracity of stories of “Love Jihad” in news and social media on a likert scale. Figure-(c) shows whether people feel the need of a law to maintain population balance between different communities. The Uttar Pradesh government had recently proposed a law with this as one of the stated goals. Figure-(d) shows whether people feel the need of a law to prevent love jihad. Such a law had been recently enacted by the Uttar Pradesh government.
Figure 6: Persuasion Rates

Notes: Figure shows persuasion rates (DellaVigna and Gentzkow, 2010) of the treatment assignment on factual beliefs and policy attitudes. 90% and 95% confidence intervals are shown. **Fertility Gap Narrowing** denotes whether the respondent correctly believes that the fertility gap between Hindus and Muslims is narrowing. **Love Jihad is Fake** denotes whether the respondent believes that all or most of the stories alleging love jihad are false. **Population Control Law Not Needed** denotes (lack of) support for draft law to control population which included maintenance of balance between population to religious communities amongst its explicitly stated objectives. **Love Jihad Law Not Needed** denotes (lack of) support for the “Love Jihad” law that was passed earlier and has been used by vigilantes and police to harass inter-faith couples in India.
Figure 7: ROC Curve

Figure shows a typical ROC curve. (0,0) corresponds to $k = \infty$ and (1,1) corresponds to $k = -\infty$. The region below the ROC curve characterises the feasibility set. The blue line represents the indifference curve which is tangent to ROC curve at the optimum.
Figure 8: Time Trends in Treatment Effects

Notes: Figure shows the time trends in treatment effects for skepticism and truth-discernment from the structural model. The truth discernment parameter captures accuracy of the subjective assessment about the veracity of statements (equation-10). Skepticism parameter is proportional to prior odds of a statement being false, and hence changes in skepticism parameter are driven by changes in beliefs about prevalence of false statements. We can see that the treatment results in an immediate increase in skepticism, however it take more time to become better truth discernment.
APPENDIX

A Further Experiment Details

A.1 Recruitment and Attrition

Table-A.1 shows the statistics for FB advertisement used for recruitment. Impressions is the total number of times the ad was seen in the news feed. Reach is the unique number of users who saw the ad. Link clicks captured the number of people who clicked on the ad which directed them to the screening survey. Survey Completion denotes the number of people who completed the screening survey for participating in the study.

Table A.1: Facebook Recruitment

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Impressions</td>
<td>2,306,157</td>
</tr>
<tr>
<td>Reach</td>
<td>744,046</td>
</tr>
<tr>
<td>Link Clicks</td>
<td>23,541</td>
</tr>
<tr>
<td>Survey Completion</td>
<td>2718</td>
</tr>
</tbody>
</table>

Table-A.2 shows the attrition across survey rounds. There is no differential attrition across control and treatment groups.

Table A.2: Response Rates

<table>
<thead>
<tr>
<th></th>
<th>Total Participants</th>
<th>Follow-up 1</th>
<th>Follow-up 2</th>
<th>Endline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>648</td>
<td>539 (83%)</td>
<td>469 (72%)</td>
<td>502 (77%)</td>
</tr>
<tr>
<td>Control Group</td>
<td>653</td>
<td>564 (86%)</td>
<td>494 (76%)</td>
<td>520 (80%)</td>
</tr>
<tr>
<td>Total</td>
<td>1301</td>
<td>1103 (85%)</td>
<td>963 (74%)</td>
<td>1022 (79%)</td>
</tr>
</tbody>
</table>

Table shows attrition rates for various survey rounds. There is no differential attrition across treatment and control groups.
A.2 Weekly Digests

The weekly digests contained summary of fact-checks of misinformation. Figure-A.1 shows the timeline of the digests. Participants received a total of nine digests during the course of the experiment.

To prepare these digests, we maintained a comprehensive database of all fact-checks from 16 IFCN (International Fact-Checking Network) certified websites in India. For inclusion in the digests, we selected fact-checks that related to major news events in that week and were related to misinformation that was going viral. We track the number of websites that fact-check a particular false news to use a proxy of how widespread particular misinformation is: viral misinformation is more likely to be fact-checked by multiple websites than that which is less widespread.

Figure-A.2 shows a sample digest. Figure-A.3 shows how misinformation tends to revolve around certain issues, following predictable patterns and techniques of manipulation. Our intervention can help people learn these patterns and hence develop effective heuristics to identify misinformation in future.

Figure-A.6 shows the distribution of partisan leaning of original misinformation for fact-checks that were included in the digests. Note that the categories are not mutually exclusive. For instance a particular false news can be both pro-government and anti-muslim. For assigning partisan slant, we chose the most salient topic that the false news was referring to. For instance if it relates to some action by the government, then it will be coded as pro or anti government.

In the last two digests, we also included narrative explainers on the issues of fertility rates and “love jihad”. Figure-A.5 shows a sample of the digest that included narrative explainer on the issue of fertility rates.
Weekly Digest 6
September 29th, 2021

By Research Team
Good morning. Last few days we saw a lot of fake news about PM Modi’s US visit. In this digest, we review some of that and other fake news.

US Newspaper, New York Times did not call Modi, the “Last, Best Hope of Earth”
In the backdrop of Modi’s US visit, several fake and misleading posts have surfaced on social media. A photoshopped image is being shared recently that shows a fake New York Times front page applauding PM Modi. The headline states that Modi is the most loved and powerful leader. Other than making obvious errors like spalling September wrong right at top, the creator of the fake image did not even use diplomatic language to make it look convincing. NYT did not publish any such image or headline about Modi.

Old pictures of demonstrations against Modi shared as recent anti Modi protests in US
While it is true that a dozen protestors demonstrated recently in front of White House US, a lot of the images showing anti Modi protests which are being shared on social media are old. Some of these images can be traced back to 2014 and one of the most recent ones is from PM Modi’s visit to Houston, USA in 2019. These are not images from Modi’s recent visit.

Navjot Singh Sidhu raised several religious slogans and not just “Allahu Akbar”
Celebrating the appointment of new chief minister of Punjab, Charanjit Singh Channi, a short video of Sidhu has gone viral on social media where he is seen raising

has gone viral on social media where he is seen raising the slogan of “Allahu Akbar”. This video has been clipped to only share the portion where Sidhu chants Allahu Akbar. Actually, Sidhu and other congress party supporters present at the venue raised several religious slogans like “Bhole So Nihal... Sat Sri Akal”, “Jai Jai Bajrangl Har Har Mahadev” and “Naaz-e Takbeer Allahu Akbar”.

In Other Fake News
• Viral video is from Bangladesh, has no relation with Pulwama Attack: A video showing a man disguised in a burqa is being shared with the claim that this is how explosives reached Pulwama in 2019 where the attack claimed the lives of 40 CRPF personnel. However, the video is actually from Bangladesh and the man was dressed in a woman’s clothes in order to smuggle alcohol. The incident happened in March this year and he was captured by police.
• PM Modi’s letter appealing people to buy made in India products this Diwali is fake: A fake letter in PM’s name has been doing rounds on social media in which he is appealing to Indian citizens to buy only products made in India during the festive season. The PM has not written any such letter and in fact had previously warned about such fake letters himself. The wide availability of Modi’s signature on the internet has made it easier for anyone to forge such documents.
• The viral video of a couple singing a song from film “Pyasa” is not that of former Raymond Group Chairman Vijaypat Singania and his wife. The two individuals in the video are from Pakistan and are cousins.
• Nobel prize winner in medicine, William Campbell did not claim that the drug Ivermectin cures Covid-19, viral post is fake. Dr. Campbell also denounced such fake quotes attributed to him.
• 5-year-old image from Patna shared to claim that cities are not standing in solidarity with Bharat Bandh called by the Samyukt Kisan Morcha (SKM).
• Video of a dispute between two families in Jodhpur, Rajasthan shared with false communal spin. The accused beating some people was not a Muslim. Both groups were Hindus, and the argument broke out over a recently organized jagraan. Such fake posts have become very common to create animosity between different religious groups, please stay vigilant.

Figure A.2: Sample Weekly Digest
Examples of repeating patterns regarding issues and techniques of manipulation

Typology of misinformation circulating on WhatsApp constructed by Banaji et al. (2019)

Figure A.3: Repeating patterns of misinformation

Notes: Figure shows how misinformation tends to revolve around certain issues following predictable patterns and techniques of manipulation. Our intervention can help people learn these patterns to become better at identifying misinformation. Figure-(a) shows a specific example that was also included in our digests. Figure-(b) shows typology of misinformation constructed by Banaji et al. (2019) after analyzing large amount of content circulating on WhatsApp.
Figure A.4: Screenshots of NYT tracking viral misinformation

Notes: Many other news organizations have recently started providing similar summaries and newsletters tracking viral misinformation on social media. Screenshots from New York Times of such summaries are shown regarding Covid and during the 2020 US election.
Weekly Digest 8
October 16th, 2021

By Research Team
Good morning. In this digest, we separate the facts from fiction concerning the demographic imbalance between religious communities in India. Read the digest to know more about this and other fake news.

Facts about demographic imbalance between communities
Recently, there has been a lot of public discussion about the possibility of rapid demographic change, especially a demographic imbalance between communities. This has also brought up debates about a population policy that can correct such an imbalance. We want to highlight some important facts regarding this issue.

To understand the issue of population growth properly, we need to look at trends in fertility rates. Fertility rate is the average number of children a woman will have in her lifetime. Fertility rates in India have been declining rapidly for everyone. All over India, it declined from 3.4 in 1993 to 2.2 in 2015. In the state of Uttar Pradesh, it declined from 4.8 in 1993 to 2.7 in 2016 and is expected to touch 2.1 by 2025. According to the UN Population Division, the population will stop growing much once fertility rates reach 2.0, which is the replacement rate.

Now, some people have raised concerns about demographic imbalance between different communities. But fertility rates in India have been declining rapidly for all religions. In fact, if we compare the fertility rates of Hindus and Muslims, even though the fertility rate for Muslims is higher than that of Hindus, it is declining at a much faster pace than those of Hindus. As per data from National Family Health Survey (NFHS), fertility rate for Muslims declined from 4.4 in 1992 to 2.6 in 2015, a decline of 1.8 whereas the fertility rate for Hindus declined from 3.3 in 1992 to 2.1 in 2015, a decline of 1.2. Hence, the gap in Hindu-Muslim fertility rate narrowed down from 1.1 to 0.5, and it is expected to vanish soon.

Even if we look at long term trends, there have not been drastic changes in religious composition of India. According to Census data, the population of Hindus was 85% in 1951 and was around 80% in 2011. Population of Muslims increased slightly from around 10% to 15% in this period. Given rapidly declining fertility rates, it is unlikely that there will be any rapid changes in the religious composition of India in future. Veteran demographer P.N. Mari Bhat has predicted that at current levels, Muslim population in India will stabilize at around 18%.

Hence, the problem of population growth has been exaggerated out of proportion. The population will stabilize even without any population control measures. Similarly, the claims and fears concerning a particular community’s population ballooning out of control are also unfounded.

Below is some other fake news that circulated this week:

- Several old and unrelated images and videos are being shared as evidence of recent violence on Muslims in Kashmir. For instance, one image is from an anti-CAA protest in Maharashtra in 2020, another is from a clash that broke out between DU’s Ramjas students and ABVP four years ago, still another one is from Delhi Riots in 2020.
- The video of an incident which took place in Chhattisgarh is being shared with the false claim that people of the Muslim community attacked the Durga Puja pandal in the Raigarh town of Pratapgarh district in UP and a group of Hindus retaliated against them. The video going viral is from the riot-hit Kawardha district of Chhattisgarh; no such incident took place in UP.
- BJP’s Amit Malviya falsely gave communal angle to an incident in Chhattisgarh where a car ran over a Durga Puja procession. Several reports have confirmed this is not an incident of communal violence and the accused, Rahul Vishwakarma and Shishupal Sahu, both from MP have been arrested.
- Indian army did not capture more than 150 Chinese soldiers in Arunachal Pradesh. Viral image is a scene from a movie called LAC which can even be found on YouTube.

Vishwakarma and Shishupal Sahu, both from MP have been arrested.

- Indian army did not capture more than 150 Chinese soldiers in Arunachal Pradesh. Viral image is a scene from a movie called LAC which can even be found on YouTube.

- Delhi CM Kejriwal did not put up an ad appealing to people to donate coal. The newspaper ad is edited and fake. This false claim comes at a time when Delhi could face an electricity crisis due to coal shortage.

- To give a message of communal harmony, prayers from all major religions of India were played at the Kisan Nyay Rally in Varanasi and not just Azaan. Clipped video is being shared to claim that Congress and Priyanka Gandhi were promoting Islam in the rally.

- Man talking about Hinduism and criticizing Nehru is not a Muslim minister in the UK government. He is director of an organization called Hindu Academy.

- The viral story of Shahnuk Khan’s son being slapped by NCB officer Sameer Wankhede is baseless. The fake story also adds that Wankhede told Khan that if he had slapped his son earlier, he would not have become a drug addict. This story is completely false.

Figure A.5: Digest with narrative explainer on fertility rates
(a) Overall

(b) For each digest

Figure A.6: Distribution of topics of fact-checks in Digests

Figure show the distribution of partisan slant of original misinformation that was included in the digests. Figure-(a) shows the distribution across all digests. Figure-(b) distribution for each digest seperately.
B  Further Baseline Descriptives

Cognitive Reflection Test (CRT)

Literature on dual-process theory in cognitive psychology has shown that deliberate cognition leads to accurate judgments. Across a number of research studies, it has been found that more reflective people are less likely to believe false content even if the information is consistent with their political ideology (Pennycook and Rand, 2019; Pennycook et al., 2021; Bago et al., 2020). We asked our respondents a set of 3 questions to record their cognitive reflection ability. These three questions constitute a standard test for measuring the capacity to ignore intuitive and quick but incorrect responses in favor of more deliberate and effortful reasoning to arrive at an accurate response. Figure B.7 below shows distribution of CRT scores in our sample.

![Figure B.7: Score on Cognitive Reflection Test (CRT)](image)

Perceived Exposure to Misinformation

We also asked respondents questions about their perceived exposure to misinformation on social media. When asked, "How often do you see things on social media that are obviously false?", the left panel of Figure B.8 shows that individuals on both sides of the political spectrum responded overwhelmingly that they encounter misinformation on social media platforms frequently. However, when asked about the partisan slant of the misinformation the participants come across, there appears a clear divide based on political partisanship.
The right panel of Figure B.8 shows that a most individuals on both ends of the political spectrum blame the opposite side for peddling a major share of misinformation on online platforms.

Figure B.8: Perceived exposure to misinformation on social media

Affective Polarization

We also measured affective prejudice towards Muslims using a feeling thermometer, i.e. asking how favorable people felt towards Muslims. For comparison, we also asked people to report how they felt towards Sikhs, which is another religious minority group. Figure B.9 shows how people in our sample feel towards the two religious communities, broken down by partisanship. In comparison to Sikhs, we see a significant unfavorable attitude towards Muslims. Also, right-leaning individuals are more likely to feel negatively towards Muslims than left-leaning individuals. Though our sample seems to respond truthfully to this question, we had included another measure to minimize social desirability concerns surrounding prejudice reporting. We used a double-list experiment to measure sensitive attitudes by offering participants greater secrecy in reporting their responses to capture prejudice towards Muslims in our sample. The estimates from double list experiment also revealed high prejudice towards the Muslim community. Around 22 percent left-leaning individuals and 35 percent right-leaning individuals are prejudiced against Muslims.
Figure B.9: Feeling Thermometers for Muslims and Sikhs
C Comparison with alternative measure of Discernment

This structural estimation is a significant improvement over existing approaches to estimating impact of interventions on truth discernment. Most prior studies report the difference between impact on TPR and FPR as discernment (Guess et al., 2020; Roozenbeek et al., 2022). An alternative of measuring impact on ratio (or equivalently differences in log values) of TPR and FPR is proposed by Guay et al. (2022).

These reduced form approaches of calculating discernment implicitly impose an arbitrary structure on the nature of discernment that is not theoretically grounded. This can be understood by looking at the mapping between the space of FPR and TPR to discernment implied by these approaches. Figure-C.10a shows the iso-discernment lines corresponding for measuring discernment as difference between TPR and FPR i.e points on each line correspond to same value of discernment as per this measure. Figure-C.10b show these for ratio measure of discernment. It is not clear why one of these should be preferred over the other; or even to other similar alternatives such as one shown in Figure-C.10c which corresponds to measure of discernment corresponding to \((1 - FPR)/(1 - TPR)\). The structural model gives a theoretically grounded measure for discernment which is widely used to evaluate the performance of binary classifiers: the position of the ROC curves.
Notes: Figure shows iso-discernment curves for various measures of discernment i.e. points on each line correspond to same value of discernment as per the measure. Panel-(a) shows the most common discernment measure ($TPR - FPR$) used in the literature (Guess et al., 2020; Roozenbeek et al., 2022). Panel-(b) shows the ratio measure ($\frac{TPR}{FPR}$) proposed by Guay et al. (2022). Panel-(c) shows another similar alternative measure ($\frac{TNR}{FNR}$). In general these measures give different results about effect of interventions of discernment; including the sign of impact. Panel-(d) shows ROC curves, which correspond to the measure of discernment in the model estimated in this paper.