Learning to Resist Misinformation:
A Field Experiment in India*

Naman Garg‡
Monika Yadav‡
(Job Market Paper)

October 19, 2022
Click here for the latest version§

Abstract

Can people learn to identify misinformation? In this paper, we conduct a large field experiment in India aimed at improving people’s ability to discern the veracity of information they encounter on social media. The intervention is to provide weekly digests with a compilation of fact-checks of viral misinformation. We find that the intervention increases the ability to identify misinformation as treated individuals are eleven percentage points more likely to correctly identify viral misinformation as false. However, it also decreases belief in true news 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. Structural estimates show that the impact is driven by both an increase in truth discernment and skepticism. A significant portion of misinformation in this region targets Muslims, the largest religious minority in India, fuelling fears about rapid changes in the demographic composition by spreading falsehoods about fertility rates and inter-faith marriages. We incorporate narrative explainers in the weekly digests to counter such misinformation. The intervention results in more accurate factual beliefs around these issues, which also leads to changes in policy attitudes and behavior.

*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 important for the functioning of democracies\(^1\). In recent years, the information environment of voters has seen drastic changes due to structural changes in news media markets brought by the advent of the internet. On the one hand, it has been critical in expanding access to information, as voters can cheaply consume various kinds of content on the internet and social media. On the other hand, the lower fixed costs of publishing has led to the entry of news providers who are unencumbered by reputational incentives for credibility, resulting in a higher supply of low quality news online (Zhuravskaya et al., 2020; Chen and Suen, 2022). As a result of these changes, social media has become an important source of news where people get exposed to a high amount of false and misleading content (Allcott and Gentzkow, 2017), and online misinformation has emerged as a huge problem.

Such high exposure to online misinformation combined with people’s inability to discern the veracity of information\(^2\) is believed to be driving widespread misperceptions about various political issues, consequently leading to changes in policy attitudes and behavior. Not only is this seen as an important factor in election outcomes\(^3\), it is also deemed responsible for spreading misperceptions and prejudice against minorities thereby fueling hate crimes (Müller and Schwarz, 2020, 2021) and ethnic violence\(^4\). Given such increased exposure and suspected adverse effects of online misinformation, it is crucial to evaluate measures that can increase people’s ability to identify misinformation, and hence mitigate its impact on policy attitudes and behavior.

In this paper, we conduct a large 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 reducing their misperceptions about minorities. The experiment is done in the state of Uttar Pradesh in India, a region that has recently seen high levels of engagement with online misinformation on social media, especially WhatsApp, which is considered to be playing a crucial role in election outcomes\(^5\). A large portion of this misinformation targets Muslims (Banaji et al., 2019), which

---

\(^1\)See Prat and Strömberg (2013) for a review of the effects of news media on political outcomes such as voter turnout, corruption, and accountability.

\(^2\)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).

\(^3\)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).

\(^4\)See Banaji et al. (2019) for a detailed analysis of anti-muslim misinformation on WhatsApp in India. Misinformation on Facebook is also believed to have played a major role in anti-Rohingya violence in Myanmar (Whitten-Woodring et al., 2020).

\(^5\)See Murgia et al. (2019); Goel (2018); Ponniah (2019) for news commentary on the role of WhatsApp in Indian elections.
is the largest religious minority in the country, and is circulated in a well-organized manner at a large scale by the Hindu nationalist party, BJP. A major theme of such misinformation revolves around creating fears of a rapid change in the country’s demographic composition 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 laws to address this alleged sudden demographic change. Hence, in our experiment, we focus on the effect of the intervention on factual beliefs around these issues, and the consequent impact on support for related policies.

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

The evaluation of such an intervention is crucial for several reasons. First, such interventions that inoculate against the effects of fake news before its exposure are important in low-income countries where encrypted messaging apps are the most popular form of social media, which precludes other popular solutions such as tagging or removal of false/misleading posts. Second, this is more robust and scalable than the other major inoculation intervention – digital media literacy training. Such training interventions require a high level of effort from social media users as they encourage trainees to check the original sources of any information, compare it with other news reports on the same issue (Guess et al., 2020), or even teach people to identify fake images by doing reverse image search (Badrinathan, 2021); and the evidence on the effectiveness of these is mixed. Our intervention is more robust as high level of cognitive effort is not required for it to be effective: it makes people familiar with the typology and patterns of fake news that they might be exposed to, allowing them to implicitly develop effective heuristics to identify it. Third, this intervention provides a rigorous evaluation of similar efforts by some news organizations and fact-checking organizations that have newsletters and dedicated sections.
on their website summarizing and debunking viral online misinformation. Hence, this impact evaluation provides useful insight into effectiveness of such efforts in mitigating the effects of fake news.

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 filled out 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 for ten weeks—from mid-August to October 2021—during which treated individuals received nine digests.

To measure the ability for 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 that was not included in the digests. This allows us to observe 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 small reduction in TPR of about four percentage points.

Next, we estimate the impact of the intervention on factual beliefs, policy attitudes, and behavior related to issues that see 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 around them, 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, as compared to simple quantitative information and hard facts (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 have shown to be fake or misleading. More accurate factual beliefs also lead to changes in 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

---


11 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. (2021) for such a demand estimation. They find that demand for newsletters increases when it includes fact-checking.
asking people to donate to an organization that helps inter-faith couples falsely accused of “Love Jihad”, and the average donation in treatment group is 15 percent higher than that in control group.

To gain further insight into the mechanisms of the reduced form impacts on truth assessment ability, we develop and estimate a structural model that formalizes the two major aspects of an individual’s truth assessment, viz. 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 result in an increase in TPR and a decrease in FPR. Another possibility is that the intervention increases overall skepticism: the intervention might lead people to update their priors about overall 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 overall 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.

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 on the basis of their timing relative to the exposure to misinformation: those done after exposure such as providing specific fact-checks (Walter et al., 2020; Pennycook and Rand, 2021); those done during exposure such as appropriate tagging of false/misleading information (Aslett et al., 2022; Clayton et al., 2020; Sanderson et al., 2021); and those done before exposure to inoculate against its effect (van der Linden et al., 2017; Lewandowsky and van der Linden, 2021), such as digital media literacy training (Guess et al., 2020; Badrinathan, 2021). Of these, inoculation interventions are especially important in regions where private messaging apps are the most popular form of social media, as the encrypted nature of communication precludes interventions such as labeling or removing false posts. 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 fake news. Bowles et al. (2022) evaluate a similar intervention in South Africa at the same time as us. They find similar results, which corroborates the potential and generalizability of this approach in low-income countries.

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 impact of interventions. Most studies report changes in difference between TPR and FPR as truth discernment. Guay et al. (2022) propose measuring changes in ratio of TPR and FPR—or equivalently
differences in log values—and show that findings of some studies reverse using this measure. However, it is not clear why log differences are better than absolute differences, or any other such functional form. We provide a micro-founded model that maps the space of FPR and TPR to the space of truth discernment and skepticism, where higher truth discernment correspond to a higher area under ROC curve\(^{12}\). 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. Hence our structural estimation provides consistent estimates of parameters of truth-discernment and skepticism that are interpretable primitives in a micro-founded model.

Second strand of literature that we contribute to focuses on measuring the extent of people’s misperceptions about facts regarding various policy issues and analyzing the effect of correcting these factual beliefs on policy attitudes and behavior\(^{13}\). This literature has explored beliefs, attitudes, and behavior related to several issues such as 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). Amongst these, our study is more specifically related to the interventions aimed at improving misperceptions and policy attitudes about outgroups or minorities (see Bursztyn and Yang (2021) for a meta-analysis). One consistent 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. 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 reduced favorability towards immigration (Barrera et al., 2020). 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\(^{14}\). Our experiment con-

---

\(^{12}\)Receiver Operator Characteristic (ROC) curves are used in machine learning literature to measure the performance of binary classifiers (Bradley, 1997).

\(^{13}\)See Haaland et al. (2022) for a recent review.

\(^{14}\)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,
tributes 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 fake 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 behavior. 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. Second, many predict a move in the future of

Bursztyn and Yang (2021) 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”

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 Africa16 and Latin America17), 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 Whatsapp18.

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 fake 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\textsuperscript{19}. Most of this content centers around Muslims (who comprise 15\% of the Indian population) overtaking the Hindu majority\textsuperscript{20} (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\textsuperscript{21}.

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\textsuperscript{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\textsuperscript{22}.

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

\textsuperscript{19}“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/
\textsuperscript{20}“In India, a debate over population control turns explosive” The Washington Post, August 29 2021
\textsuperscript{21}“Love jihad: The Indian law threatening interfaith love.” BBC News, 8 Dec. 2020
\textsuperscript{22}This click-through rate of around one percent is similar to other experiments that used Facebook ads for recruitment (Allcott et al., 2020, 2022)
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 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. One 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.
ship, 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. 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. 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. The objective of this is to guide the participants through the landscape of online misinformation — the issues that saw fake 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

---

24 A total of nine digests were published over a period of around ten weeks. The 1st digest was posted on August 16, and the last digest was published on October 22. See Appendix for more details.

25 There is no differential attrition across treatment and control groups. See Appendix for lee bounds for the main results.

26 See Appendix for a sample of digests and screenshots of the app.
a detailed background to these issues, included not only numbers and statistics but also anecdotes and stories of individuals who 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{27}.

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 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

\textsuperscript{27}See Appendix for details.
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.

A. Partisanship

We collected three measures of political partisanship. These are presented in Figure (3). First is the satisfaction with the central government, which is currently ruled by right-wing Bhartiya Janata party (BJP). This was also 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).28

B. 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.29 Figure 4 below shows the total CRT score in our sample.

C. 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 5 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 5 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.

D. Misperceptions and Prejudice towards Out-group

We collected data on misperceptions about Muslims — a religious minority outgroup. 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 higher in the case of Muslims (NFHS 5 Data). However, as the top left panel of Figure 6 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

28These measures are highly correlated with each other, and all results are robust to other measurements of partisanship
29See Appendix for detailed survey instruments used.
bogus\textsuperscript{30}. However, as the top right panel of Figure 6 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 6 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.

Finally, 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 7 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\textsuperscript{31}.

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

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

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.

\textsuperscript{30}“Love Jihad”: Myth vs Reality” NewsLaundry Project, 2021

\textsuperscript{31}Details of the list experiment and its estimation are in Appendix
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 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 fake 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 interention 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

We also report persuasion rates for each outcome, which is a standardized measure of the causal impact allowing comparison across various studies (DellaVigna and Gentzkow, 2010). Persuasion rates can be interpreted as the percentage of
people who are persuaded by the treatment, among those that are not already per-
suaded. These are calculated by dividng 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.

We can see that the intervention results in more accurate factual beliefs. Peo-
ple 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.

More accurate factual beliefs also translate to changes in policy attitudes and
behavior. Intervention results in a 4 percentage point reduction in support for
discriminatory laws against muslims regarding population control and ‘love jihad’.
We also measure changes in behavior by asking study participants to donate a
portion of their survey earnings of rupees 100 (≈ 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.

5 Structural Estimation

The reduced form results in Section-4.1 for impact of intervention on truth assess-
ment 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 possi-
ble mechanisms impact: truth discernment, and skepticism. The parameters corre-
spanding 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 seperately 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 \( u : \{0, 1\}^2 \rightarrow \mathbb{R} \) 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\(^{32}\). 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 accurate 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. \( f(s_{ij}|t=1) / f(s_{ij}|t=0) \) 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

\[
 TPR(a) = Pr (a(s_{ij}) = 1 | t_j = 1) \\
 FPR(a) = Pr (a(s_{ij}) = 1 | t_j = 0)
\]

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

\[
 U(a) = \pi \cdot TPR - (1 - \pi) \cdot c \cdot FPR
\]

Hence the optimal prediction policy is a solution of maximization problem \( \max_{a \in A} U(a) \)

\(^{32}\)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.
subject to constraints (3), (4). It is easy to see that the optimal policy can be completely characterized\footnote{This follows from posterior probability being an increasing function of signal realization} 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

\[
\begin{align*}
\text{TPR}(k) &= 1 - F(k|1) \\
\text{FPR}(k) &= 1 - F(k|0)
\end{align*}
\]

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

\[
\frac{d(\text{TPR})}{d(\text{FPR})} = \frac{f(k|1)}{f(k|0)} = c \left( \frac{1 - \pi}{\pi} \right)
\]

The optimization problem can also be visualized in terms of Receiver Operator Characteristic (ROC) curves. Eliminating \( k \) from equations-(5),(6), we obtain the equation for ROC curve given by

\[
\text{TPR} = 1 - F\left(F^{-1}(1 - \text{FPR}|0) |1\right)
\]

It is easy to verify that equation-8 is a ROC curve i.e. \((0,0)\) and \((1,1)\) lie on the curve corresponding to \( k = \infty \) and \( k = -\infty \) respectively; and \( \frac{d(\text{TPR})}{d(\text{FPR})} = \frac{f(k|1)}{f(k|0)} > 0 \). 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). See Appendix for a formal proof.}. Figure-8 shows a typical ROC curve.

The threshold value \( k \) parametrizes 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 though 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-(??) which are straight lines with slope \( c \left( \frac{1 - \pi}{\pi} \right) \). The optimum \( k \) is chosen such that the indifference curve is tangent to ROC curve which is captured in FOC condition-(7).

The slope of indifference curve \( c \left( \frac{1 - \pi}{\pi} \right) \) 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 $c \left( \frac{1 - \pi}{\pi} \right)$ 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.

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

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

$$\begin{align*}
TPR &= 1 - \Phi(\rho(k-1)) \\
FPR &= 1 - \Phi(\rho k)
\end{align*}$$

The ROC curve is

$$ TPR = 1 - \Phi \left( \Phi^{-1} (1 - FPR) - \rho \right) $$

The optimal threshold value $k$ is given by FOC-(7) is

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

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:

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

where $a, t$ denote stacked vectors of observations. Hence $\rho, k$ can be estimated using
MLE, and plugging these values in equation-(10) 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-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.

In what follows I’ll estimate the parameters separately for control and treatment group and compare them. I’ll 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 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-9a 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-9b 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-9c 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.

5.3 Estimation Results

Table-4 shows the estimates of parameters of the models by maximizing the likelihood function in equation-(11). To get the treatment effect on these parameters,
the parameters are estimated separately for control and treatment groups. The treatment increases the truth discernment parameter by 19 percent and increases the skepticism parameter 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.

Overall, 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 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 as it also increases overall skepticism, as people update their priors about prevalence of fake 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. 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-10, where Figure-10a shows the reduced form effects and Figure-10b shows structural estimates for each survey round.

We can see that from these figures that the treatment almost immediately results in a reduction in beliefs about both true and false statements, which is also captured by an immediate increase in the skepticism parameter. As people start reading these digests, they update their prior about overall prevalence of misinformation and become more skeptical overall 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.

---

35 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)
6 Conclusion

We find that voters’ ability to assess the veracity of information can be improved by a simple intervention – providing them with regular summaries of fact-checks of viral misinformation. Compared to other interventions aimed at inoculation against misinformation, such as digital literacy, this is much simpler and more effective as it need not be tailored to baseline digital proficiency or sophistication of the user, and does not require a high level of cognitive effort during news consumption. This also provides a rigorous evaluation potential implication of recent trends where some news organizations provide similar summaries and newsletters to their readers.

More accurate factual beliefs also translate to changes in policy attitudes and behavior when information is provided in a narrative form, even in a context characterized by high level of affective polarization. This finding adds to the emerging evidence highlighting the importance of narratives in shaping people’s attitudes and behavior regarding outgroups and minorities (Alesina et al., 2018a; Broockman and Kalla, 2016; Kalla and Broockman, 2020).

None of the existing studies finding effect of narratives on attitudes, including ours, is able to provide direct evidence on mechanisms of the impact of narratives. Future research should focus on understanding these mechanisms. Better insights into mechanisms can provide a more holistic understanding of determinants of policy attitudes and help in generalizing interventions to reduce prejudicial or exclusionary attitudes against outgroups across different contexts.
References


Table 1: Sample Characteristics

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Sample (UP)</th>
<th>Population (India)</th>
<th>Social Media Users (India)</th>
</tr>
</thead>
<tbody>
<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>
<tr>
<td>N</td>
<td>1301</td>
<td>200 million</td>
<td>460 million</td>
</tr>
</tbody>
</table>

*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>Full Sample</th>
<th>Right-Leaning Individuals</th>
<th>Left-Leaning Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>False</td>
<td>True</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td>Statements</td>
<td>Statements</td>
<td>Statements</td>
</tr>
<tr>
<td>Treatment</td>
<td>-0.107***</td>
<td>-0.038**</td>
<td>-0.131***</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>(0.015)</td>
<td>(0.021)</td>
</tr>
<tr>
<td>Control Mean</td>
<td>0.480</td>
<td>0.758</td>
<td>0.449</td>
</tr>
<tr>
<td>Observations</td>
<td>4088</td>
<td>4088</td>
<td>2504</td>
</tr>
</tbody>
</table>

*** p < 0.001; ** p < 0.01; * p < 0.05.

Notes: Tables shows results from estimation of equation-(1) on endline data. Demographic controls include age, caste, gender, religion and, score on CRT. All specification include statement fixed effects. Column (1) and (2) show results for only false and true statements respectively for the full study sample. 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.
Table 3: Treatment Effect on Beliefs, Attitudes and Behavior

<table>
<thead>
<tr>
<th>Panel A. Factual Beliefs</th>
<th>Treatment Effect</th>
<th>Control Mean</th>
<th>Persuasion Rate</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fertility Gap Narrowing</td>
<td>0.131***</td>
<td>0.215</td>
<td>17%</td>
<td>1022</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is Love Jihad Fake</td>
<td>0.068**</td>
<td>0.465</td>
<td>13%</td>
<td>1022</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beliefs Index</td>
<td>0.098***</td>
<td>0.340</td>
<td></td>
<td>1022</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel B. Policy Attitudes

<table>
<thead>
<tr>
<th>Population Control Law Need</th>
<th>Treatment Effect</th>
<th>Control Mean</th>
<th>Persuasion Rate</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.038*</td>
<td>0.842</td>
<td>5%</td>
<td>1022</td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Love Jihad Law Need</td>
<td>-0.040*</td>
<td>0.769</td>
<td>5%</td>
<td>1022</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitudes Index</td>
<td>-0.039**</td>
<td>0.806</td>
<td></td>
<td>1022</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel C. Behavior

<table>
<thead>
<tr>
<th>Donation</th>
<th>Treatment Effect</th>
<th>Control Mean</th>
<th>Persuasion Rate</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.310**</td>
<td>38.519</td>
<td>14%</td>
<td>1022</td>
</tr>
<tr>
<td></td>
<td>(2.595)</td>
<td></td>
<td></td>
<td></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 viz. age, caste, gender, religion and score on CRT. Regressions for Factual Beliefs and Policy Attitudes also include corresponding baseline measure as a control.

Panel A shows outcomes 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 fake. *Beliefs Index* is an average of the two beliefs outcomes.

Panel B shows outcomes measuring support for discriminatory policies against Muslims. *Population Control Law Need* denotes 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 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 behavior outcomes. *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 the treatment (to change beliefs, attitudes or behavior), among those that are not already persuaded. Persuasion rate for donation corresponds to binary measure which is one when amount of donation is more than or equal to Rs 50.
### Table 4: Estimates from the model

<table>
<thead>
<tr>
<th></th>
<th>Truth-Discernment</th>
<th></th>
<th>Skepticism</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>
</tr>
<tr>
<td>Full Sample</td>
<td>0.748</td>
<td>0.904</td>
<td>0.189**</td>
<td>0.886</td>
</tr>
<tr>
<td></td>
<td>(0.041)</td>
<td>(0.041)</td>
<td>(0.071)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>Left Leaning Individuals</td>
<td>0.786</td>
<td>0.811</td>
<td>0.031</td>
<td>0.901</td>
</tr>
<tr>
<td></td>
<td>(0.065)</td>
<td>(0.066)</td>
<td>(0.116)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Right Leaning Individuals</td>
<td>0.725</td>
<td>0.965</td>
<td>0.286**</td>
<td>0.877</td>
</tr>
<tr>
<td></td>
<td>(0.052)</td>
<td>(0.053)</td>
<td>(0.091)</td>
<td>(0.012)</td>
</tr>
</tbody>
</table>

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

**Notes:** Table shows MLE estimates obtained by maximizing likelihood function (11) where the value of $r$ is obtained by plugging in MLE estimates of $\rho$ and $k$ in equation (10). The estimation is done separately for control and treatment groups and the differences in parameter values are reported.
Figures

Figure 1: Experiment Design

- Screening survey completed on FB
  (N = 2718)
- Contact details verified and participation confirmed on phone
  (N = 1806)
- Placebo app installed
  (N = 1341)
- Baseline survey completed
  (N = 1301)
- Block randomization
- Treatment
  (N = 648)
- Control
  (N = 653)

- September 8
  Follow-up Survey 1 - 85% (N = 1103)
- October 1
  Follow-up Survey 2 - 74% (N = 963)
- October 26
  Endline Survey - 79% (N = 1022)
Figure 2: Compliance Rates with the treatment
Figure 3: Political Partisanship
Figure 4: Score on Cognitive Reflection Test (CRT)
Figure 5: Perceived exposure to misinformation on social media
Figure 6: Factual beliefs and policy attitudes about issues related to Muslims
Figure 7: Feeling Thermometers for Muslims and Sikhs
Figure 8: 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.
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.
Figure 10: Time Trends in Treatment Effects