Practical Solutions for Addressing Digital Misinformation in India

Authors: Reed Cohen, Danielle Dougall, Kendal Gee, Charlie Plumb, Ashwini Sivaganesh, Yahui Sun, Jacqueline Viteznik, and Gunnar Wainwright

Advisor: Dr. Robert McKenzie

April 2024

TABLE OF CONTENTS

Charlie Plumb, MIA Candidate, Columbia SIPA

CROSS-DOMAIN COLLABORATION FOR AI-POWERED CONTEXT ANALYSIS

BLUF: This proposal focuses on facilitating collaborative innovation in order to build the capacity of peacebuilding- and media-focused nonprofits to leverage AI in their context analyses. The Global Engagement Center (GEC) should organize a design thinking workshop, bringing together immediate stakeholders in the nonprofit sector and tech companies, as well as relevant experts from local universities and international NGOs.

Needs	**Activity**	**Outputs**	**Intended Outcomes**	**Goal**
Understand fundamentals and potential uses of AI	Design thinking workshop	Stakeholders propose refined design ideas for appropriate tools	Understanding of AI basics, and technical feasibility of potential applications in the NGO space; professional and interpersonal lines of contact developed among participants	Efficient context analysis of digital information ecosystems in India
Funding for the development of AI-powered tool(s)	Funding process	Promising innovations proposed by stakeholders are funded for development	Efficient, scalable software that meets individualized organizational needs	
Actors to be engaged: . Nonprofit practitioners from peacebuilding and mediation, and media and fact-checking nonprofit organizations . Engineers and representatives from Indian and international tech companies and platforms . Relevant experts from think tanks and universities . Practitioners from international NGOs that have begun to integrate AI tools in programming				

Opportunity: The potential for AI-powered tools in understanding information ecosystems is, so far, largely untapped. Interviewed NGO representatives expressed significant interest to better understand and leverage AI and social media in their work. However, domain knowledge in peacebuilding and technological fields are largely siloed from one another; personnel in the nonprofit sector tend to not have backgrounds in computer science. Nonprofit organizations in India also face significant financial challenges and shrinking space in which to operate. Fostering cross-domain understanding between technologists and NGO practitioners, and funding the development of appropriate tools that emerge from their collaboration, will effectively fill the knowledge and financial gaps that restrict NGOs' capacity to leverage emerging technologies.

Recommendations:
Activity 1: Bring together key actors from different NGOs and from AI companies through a design thinking workshop, a 2-day immersive experience to facilitate cross-domain learning and collaborative innovation.

Devote the first day to knowledge sharing. AI experts and tech company representatives provide short, informative presentations on AI fundamentals, including machine learning, neural networks, and social media analysis. Then, nonprofit workers give an overview of their various work, including peacebuilding, community-outreach, and media literacy and fact-checking. They should focus on their own organizations' objectives and challenges to help technologists understand the real-world issues NGOs face, particularly in context analysis. Traditional context analysis methods, such as needs assessment,

stakeholder analysis, and conflict mapping, often fail to capture the complexity of inter-communal perspectives, grievances among hard-to-reach individuals, or the relative importance of diverse social divisions, and almost never do so dynamically. Then, peacetech experts (actor groups 3 and 4 in the table) share insights from the field, presenting stories and case studies where technology has successfully aided NGO efforts or where there has been a significant need for improvement.

Devote the second day to collaborative brainstorming. All participants contribute ideas for AI-powered tools that could address the identified challenges. Tech experts provide feedback and guidance on technological feasibility and practical considerations (such as model choice, and required computing power, server space and training data). NGO actors should also communicate ethical considerations, as well as feedback on how their organizations could or would practically use the tools. By the end of the event, participants should have developed a few refined ideas that are understood by all involved.

Activity 2: Fund the development of promising proposals for tools that come out of the design thinking workshop. Grants could be allocated directly to the tech companies building the tools, so as to avoid bureaucratic complications associated with foreign funding of nonprofits.

Intended Outcomes: Developing and implementing ideated tools will facilitate efficient, scaleable, context analysis, enhancing both immediate targeted programming and future resource allocation. The event will also foster understanding of AI within the nonprofit space, and facilitate interpersonal and professional relationships across sectors. Knowledge and connections will help foster further innovations in NGO application of emerging technologies.

Background: Context analysis is vital for effective programming, but it is expensive and time-consuming, presenting significant challenges for nonprofit peacebuilding and media organizations. NGOs require in-depth knowledge of key actors, groups, attitudes, behaviors, grievances, and divisions in order to design and implement interventions. Such analyses are often done ad hoc, or else informally rely on practitioners' individual or collective knowledge.

This gap in programming is compounded by the rise of social media and digital manifestations of polarization and grievance, as online behavior both reflects and impacts perspectives and discourse. AI-powered context analysis tools are tailored, customized models that can analyze diverse datasets. Examples of potential tools include:

- Graph neural networks for topological network analysis, to identify key influencers and/or potential echo chambers online.
- Narrative recognition and modeling through entity-based semantic analysis, prioritizing those narratives linked to social divisions.
- Epidemiological models for diffusion analysis, to examine the pathways, speed, and influence of ideas and content disseminated in a network.
- Pattern analysis to detect anomalous activity, which may indicate a focal point or help predict a trigger event.

India's media environment, including social platforms, news outlets and mainstream media, is vast and complex. Traditional data mining techniques are difficult because so much content is irrelevant to the analysis at hand, known as 'noise' in datasets. Ideas generated in the design thinking workshop should be narrow in scope and scale, and those that require Natural Language Processing (NLP) will likely be limited to English and Hindi.

The overarching goal of this proposal is not to alter India's information environment directly. There's therefore no need to circumvent or come into conflict with entities fueling information manipulation.

Rather, this program aims to increase the capacity of NGOs to understand online contexts, empowering them to develop and scale targeted interventions appropriately.

Reed Cohen, MPA Candidate, Columbia SIPA

SCALING TIPLINES TO UNDERSERVED LINGUISTIC COMMUNITIES

BLUF: The Global Engagement Center should use grants to fund the creation and sustainment of digital fact-checking departments at local print news organizations, helping expand the fact-checking ecosystem to include unserved, vulnerable language groups, combating the proliferation of online misinformation within these groups.

Opportunity: Online misinformation in India often inflames already tense divisions between different ethnic, religious, and tribal groups, which has led to and exacerbated communal violence. In 2018, for example, a Pakistani video PSA edited to look like a real kidnapping went viral on WhatsApp in India, fueling false rumors about kidnapping and ultimately galvanizing deadly mobs against suspected "child lifters" across the country.[1] Thankfully, an ecosystem of fact-checking organizations has emerged in India since 2018, which works to debunk viral misinformation like that at the center of these cases, helping stem the further spread of communal violence in these regions. However, gaps continue to exist in the scope of delivery of these services, namely concerning the number of languages they cover. Between the 23 International Fact-checking Network (IFCN) certified organizations operating in India, fact-checks are only published in 17 of the more than 120 languages spoken in the country.[2] These gaps have real life consequences. In Manipur (a state in northeast India that borders Myanmar) for example, the circulation of online misinformation ignited ethnic violence between the Kukis and Meiteis last spring and continues through today.[3] Neither group speak languages currently covered by Indian fact-checking organizations. Without fact-checking resources in these more regionally isolated and smaller linguistic communities, the target groups of misinformation have little hope of pushing back against false, harmful narratives online.

Recommendation: A GEC grant program for local news publishers will help scale tiplines to combat misinformation spreading in the more than 86% of Indian languages that current fact-checking organizations do not service. The specific cultural contexts of local linguistic groups across India make it difficult for non-native speakers and AI tools to catch misinformation, which is often coded in slang and cultural norms, further driving the need for local fact-checking resources. This grant program should provide funds to local news publishers with a demonstrated commitment to combating all genres of misinformation. The GEC should prioritize publishers based on their region's level of vulnerability to communal violence (i.e., whether there is a history of violence in the region, is there a history of violence nationally between religious, ethnic, or other groups in the region, and smartphone ownership rates). Publishers' eligibility for these grants should follow Meta's criteria for WhatsApp's third party fact-checkers (3PFCs); nonpartisanship and fairness, transparency of sources, transparency of funding and organization, transparency of methodology, and open and honest corrections policy; given these publishers should ultimately launch tiplines on the app.[4] Grantees should be required to undergo fact-checking training with DataLEADS, a New Delhi based digital media organization which has trained a majority of India's fact-checking organizations through its Google funded India Training Network initiative (the initiative has also trained more than 2,500 newsrooms and colleges in fact-checking and

[1] Vindu Goel, Suhasini Raj, and Priyadarshini Ravichandran, "How WhatsApp Leads Mobs to Murder in India," The New York Times, July 18, 2018, https://www.nytimes.com/interactive/2018/07/18/technology/whatsapp-india-killings.html.
[2] Spandana Singh, rep., *The False Information Ecosystem in India* (Tokyo: Toda Peace Institute, 2019), p 10.
[3] "How Fake News and Videos Inciting Violence in Manipur," The Economic Times, July 24, 2023, https://economictimes.indiatimes.com/news/india/how-fake-news-and-videos-inciting-violence-in-manipur/articleshow/102065845.cms?from=mdr#.
[4] "How Meta's Third-Party Fact-Checking Program Works," facebook.com, June 1, 2021, https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works.

online verification).[5] India's increasingly zealous enforcement of its Foreign Contributions Regulation Act could make it difficult for the GEC to disburse grants directly to local news publishers.[6] Instead, given its proven track record of funding independent fact-checking organizations, the GEC should partner with Meta to run this program, select grantees, and disburse grants.

Intended Outcomes:

1) Decrease the spread of online misinformation within underserved language groups
2) Decrease the occurrence of communal violence within target communities
3) Reduce the negative sentiment of outgroups within these communities

Metrics for Success:

1) Adoption of tipline
    - Number of fact-check requests made by users
    - Request fulfillment rate
    - Reach of corrective content (via forwards) on WhatsApp
2) Survey data: positive sentiment shift (compared to baseline survey results ascertained at start of program) toward ethnic, religious, etc. outgroup within target region
3) Behavioral changes: decline in communal violence occurrence rate in target region

Background: Fact-checking organizations are a critical pillar of combating online misinformation in India given the limited visibility platform owners have over the mediums where false narratives proliferate. Direct messaging apps like WhatsApp are a significant vector of misinformation in this landscape and the nature of end-to-end encryption of individual and group chats limits Meta's ability to exercise oversight and content moderation. This dynamic has motivated Meta to partner with a variety of 3PFCs, which the company approves to operate tiplines for its users. 3PFCs verify or debunk content (i.e., text, images, videos) forwarded to them by WhatsApp users, many of whom then forward this fact-checked content to other individual or group chats to share the veracity of the information. WhatsApp users demonstrated strong interest in tipline services ahead of India's last general election; a majority of viral content in public WhatsApp groups was actually first shared on tiplines before appearing in these groups.[7] Not only can tiplines debunk misinformation before it goes viral, they may also be more effective than other interventions at stemming its spread since corrective messages from peers online have been shown to substantially reduce the believability of misinformation.[8]

While Indian fact-checking organizations currently publish in just 17 languages, its print media landscape is far more linguistically diverse, publishing 5,638 daily newspapers and 348 weeklies in 101 languages.[9] Print news publishers have been able to scale to far more language groups because their products are directly purchased by their consumer bases, whereas most IFCN certified fact-checking organizations in India are funded by Meta, limiting the scope of their services to what the company is willing to underwrite. This gap represents a market failure which the GEC could help remedy. The GEC should establish a grant program for local news organizations publishing in under-served languages so they can fund online fact-checking departments.

---

[5] "GNI Network - OW Dataleads Private Limited," DataLEADS, April 1, 2024, https://dataleads.co.in/GNINetwork/.

[6] "India Should Stop Using Abusive Foreign Funding Law," Human Rights Watch, February 10, 2022, https://www.hrw.org/news/2022/01/18/india-should-stop-using-abusive-foreign-funding-law.

[7] Ashkan Kazemi et al., "Research Note: Tiplines to Uncover Misinformation on Encrypted Platforms: A Case Study of the 2019 Indian General Election on WhatsApp," *Harvard Kennedy School Misinformation Review*, January 31, 2022, 1–17, https://doi.org/10.37016/mr-2020-91.

[8] Sumitra Badrinathan and Simon Chauchard, "'I Don't Think That's True, Bro!' Social Corrections of Misinformation in India," *The International Journal of Press/Politics* 29, no. 2 (February 27, 2023): 394–416, https://doi.org/10.1177/19401612231158770.

[9] "Indian Languages," education.gov.in, accessed March 25, 2024, https://www.education.gov.in/sites/upload_files/mhrd/files/upload_document/languagebr.pdf.

Jacqueline Viteznik, MIA Candidate, Columbia SIPA

BOTTOM-UP APPROACHES FOR WHATSAPP: MANIPULATED INFORMATION IN INDIA

BLUF: The GEC should organize a design sprint, leveraging LLMs to identify manipulated information on closed, private group chats on WhatsApp, which feature close social ties that affect trust. In partnership with academia, CSOs, fact-checkers, and AI developers, this LLM will create visibility of this critical spreader of disinformation and can tailor bottom-up interventions.

Opportunity: Closed, private group chats on WhatsApp (e.g., family/diaspora groups, neighborhoods, friends, hobbies, larger communities, etc.) have a higher degree of perceived trust due to close social ties, which makes disinformation more dangerous. WhatsApp is a particular challenge for tracking misinformation due to the lack of visibility from the end-to-end encryption feature, as well as the ability for users to forward and send messages across multiple groups. Particular narratives can be repeated easily across group chats, reinforcing confirmation bias. Certain mediums, like audio and images (e.g., memes) tend to be more trusted than simple text.[10] These group chats, especially in non-English languages, have little data collection and represent an extremely vulnerable area. Simply capturing this data, organizing it, and creating a means of identification will be valuable for relevant stakeholders.

Recommendations: The GEC should fund and coordinate a design sprint in the form of a workshop/working group to leverage Large Language Models (LLM) in Hindi to identify, expose, and categorize misinformation and cater interventions/corrections. Stakeholders should include American academics who have already conducted similar LLM projects on misinformation, as well as on the ground, third-party fact-checkers. Local digital rights groups should also be consulted. These groups should include academic researchers (Dr. Garimella, Rutgers), fact-checkers (The Quint, Mythos Labs, DataLeads, Logically Facts), and digital rights groups (Internet Freedom Foundation).

This design sprint should begin with identifying communities that are accessible to researchers and are proven to be vulnerable to disinformation, as decided by local stakeholders. These stakeholders should utilize a data donation approach, in which users would consent to share WhatsApp group messages with researchers. The design of this collection methodology should include an incentive for users, such as temporary SIM cards, in addition to an informed consent waiver. One academic interview is currently piloting this methodology on this issue of misinformation in India and is soliciting donations through QR codes.

In labeling these chats and information, developers should categorize the type of social ties, as well as social hierarchy. The effectiveness/trust in manipulated information can be measured by in-app interactions (e.g. likes, replies, forwarding) and the message's appearance across multiple groups (network analysis could also be leveraged). In addition, the degree of manipulation and medium should also be categorized. Fact-checking worthiness (or the degree of possible violence) of disinformation should also be identified in the design sprint.

This identification and exploration of the above variables will help identify and inform much-needed bottom-up interventions. Previous successful interventions have used micro-influencers and comedians to counter false narratives and create content users can share organically. For example, fact-check labs and digital rights groups interviewed in India underscored the importance of humor-based interventions, such

[10] Paris, Britt & Pasquetto, Irene. "Hidden Virality and the Everyday Burden of Correcting WhatsApp Mis- and Disinformation," in *Governing Everyday Misinformation*, ed. (Cambridge University Press, 2024); Garimella, Kiran & Eckles, Dean. "Images and Misinformation in Political Groups: Evidence from WhatsApp in India," Harvard Kennedy School Misinformation Review. https://misinforeview.hks.harvard.edu/article/images-and-misinformation-in-political-groups-evidence-from-whatsapp-in-india.

as comedy sketches and memes as a keyway to engage users and help sustain focus on correcting/challenging manipulated information. Digital literacy and engagement programs are also ways to encourage organic corrections. Providing certifications has emerged as an effective method to incentivize participation in younger demographics.

Intended Outcome: There needs to be visibility of these private group chats, creating a comprehensive and quantitative understanding of the role of close social ties and messages' medium on manipulated information. This design sprint aims to leverage and empower stakeholders and community members to engage in effective interventions. If successful, this design sprint can be replicated in other communities in India or in other countries with similar WhatsApp usage (e.g., Brazil).

Metrics for Success:

- Effectiveness and reach of data donation to understand the information landscape
- Amount of disinformation tagged and identified and the correlation with medium and social tie; engagement and circulation across group chats
- Participation rate in intervention programs and percent increase in organic corrections
- Feasibility of replication and adaptability to local contexts

Background: WhatsApp has quickly emerged as one of the most widely used messaging services around the world and within India. Estimates from 2023 list the app as having 490 million users within India. WhatsApp has also evolved from simply peer-to-peer messaging to a news, marketing, and political platform. Often, these functions merge into one familial meso-news space. The danger of misinformation on the app became global news in 2018 when a violent mob targeted an outsider community after a false story was spread through WhatsApp groups about child kidnappings. Meta, in response to this violent attack, limited the size of group chats and the ability to forward (forwarded content also now has a label). However, given the vast infrastructure and previous penetration on WhatsApp by motivated actors, these interventions have done little to curb those efforts.

Private, closed messaging apps are absent of algorithms pushing the content. Users trust that encountered content is being sent from a trusted person; motivated actors can easily target specific audiences and measure how effective content is within chats, which has helped fuel the WhatsApp marketing industry. Because of close social ties (many members of chats are close offline), the incentive to correct false claims is low, and users are often aware of the content's potential to be false but share anyway.[11] Similarly, users also have reported that they are aware of exposure to misinformation but do not challenge the falsehoods.[12] Corrections are especially hard to encourage among young people exclusively due to cultural factors and the increased social status of elders sharing misinformation. Programs and intervening content should also target elderly populations, who also spend large amounts of time on the app.

Fact-checking labs interviewed in India listed deep-fakes and cheap-fakes (low-tech versions of the former) as being particularly harmful with the upcoming elections. Academics and CSOs asserted that narratives and content are often being recycled across multiple years. The concern is in limiting the spread of disinformation, not necessarily stopping its generation. Resilience to misinformation that is human-made will also likely produce resistance to AI generation.

[11] Paris, Britt & Pasquetto, Irene. "Hidden Virality and the Everyday Burden of Correcting WhatsApp Mis- and Disinformation," in *Governing Everyday Misinformation*, ed. (Cambridge University Press, 2024).

[12] Neyazi, T. A., Kalogeropoulos, A., & Nielsen, R. K. (2021). Misinformation Concerns and Online News Participation among internet Users in India. Social Media + Society, 7(2). https://doi.org/10.1177/20563051211009013

Danielle Dougall, MPA Candidate, Columbia SIPA

ESTABLISHING AN EARLY WARNING SYSTEM FOR MISINFORMATION ON TELEGRAM

BLUF: The Global Engagement Center should fund and coordinate a partnership with Indian fact-checking coalitions and anti-disinformation research firms to establish a proactive early warning system that detects emerging election falsehoods and hate speech on Telegram.

Opportunity: Telegram, with its vast user base and lax content moderation, has become a breeding ground for misinformation, hate speech, extremism, and terrorist activities. Notably, it has facilitated terrorist recruitment by the Islamic State in Khorasan Province (ISKP) in India.[13] These challenges pose direct threats to India's elections and public safety. By partnering with leading anti-disinformation research firms and established Indian fact-checking organizations, the Global Engagement Center can enhance intelligence sharing and local verification of information. These local stakeholders are invaluable, providing essential insights and lending credibility to significantly boost the effectiveness of the early warning system. Fostering partnerships between the U.S. embassy[14] and Indian fact-checking organizations will amplify local verification efforts. This collaborative approach will enable detailed investigations into previously overlooked security risks and help establish a comprehensive early warning system designed to proactively counteract threats on Telegram and mitigate related offline harm.

Strategic Partner/Initiative	**Role and Key Initiatives**	**Metrics for Success**
Graphika (or other anti-disinformation research and analysis firms)	Graphika's AI-powered tools and information network will enhance threat identification on Telegram	Broad threat identification capabilities, including improved threat detection and response times Enhanced fact-check opportunities sharing through networks
Misinformation Combat Alliance (MCA)	MCA's cross-industry coalition of 16 organizations specializes in disseminating fact-checks and reports to counter election misinformation and address hate speech	Increased public awareness of accurate election information and rapid dissemination of verified fact-checking content
DataLEADS (a digital media and tech company training Indian fact-checking organizations)	DataLEADS enhances local fact-checking through its Google-funded Fact-Shala program	Increased local fact-checking capacity to ensure precise and timely identification of misinformation

Recommendation: The GEC should establish a robust early warning system at its Mission Center through a public-private partnership with anti-disinformation research firms and Indian fact-checking organizations. This system would proactively detect and counteract false information on Telegram by analyzing emerging misinformation and its consequences. Insights gained would be shared with approved Indian organizations responsible for publishing fact-checks and reports.

[13] Mona Thakkar and Vineet P, "The State of Play: Islamic State Khorasan Province's Anti-India Propaganda Efforts," GNET, 22 May 2023, https://gnet-research.org/2023/05/22/the-state-of-play-islamic-state-khorasan-provinces-anti-india-propaganda-efforts/.

[14] U.S. embassy support could take the form of capacity-building programs for vetted Indian fact-checking organizations, including analysis and research assistance, training on rigorous monitoring and evaluation, and funding.

The early warning system will feature a dedicated section in its interactive dashboard, enabling Indian fact-checking organizations to access and address misinformation and hate speech in English and Hindi. It will initially monitor 100 pre-identified Telegram accounts known for disseminating harmful content. Integrating local insights and verifications will refine the system's analytics, ensuring a culturally sensitive and accurate response to election misinformation. Alerts for identified threats will be promptly sent to GEC partners, utilizing advanced detection techniques to enhance proactive fact-checking.[15] Key components of the early warning system include: [16]

- Data Collection: Begin by scraping data from 100 pre-selected Telegram channels using its API, while considering ethical implications. This data will be used to build a detailed analysis and database.
- Long-term Strategy: Develop a comprehensive database of misinformation narratives and threat actors on Telegram, continuously monitored for new developments. This database will support an AI-powered dashboard that provides alerts and facilitates deeper analysis.
- Network Analysis: Examine social media connections to uncover coordinated misinformation campaigns. If interconnected patterns are detected, notify other social media platforms for review and potential action.
- Trend Analysis: Identify and address emerging trends in election misinformation and hate speech to preempt their spread.
- Alert System: Implement an automated alert via the early warning system within the partnership network, enabling rapid and proactive response to online falsehoods.
- User Education: Enhance public understanding by empowering users to critically evaluate and verify information through targeted fact-checking and educational initiatives.

Background: Telegram, with over 100 million users in India and 900 million globally, poses unique risks to India's elections due to its limited content moderation. This instant messaging service enables rapid dissemination of extremist content and falsehoods. Telegram channels, which allow one-way broadcasting to unlimited users and groups, where up to 200,000 members can interact and share content, amplify the potential for widespread misinformation and hate speech linked to offline violence.[17] Research for this project identified over 80 Telegram channels promoting misinformation and anti-Muslim hate speech, collectively reaching over 100,000 users.

Telegram also provides a sense of community through reinforced narratives, safety, and impunity from consequences users face on more mainstream platforms, making it popular among global extremist groups. There is a demonstrated link between these groups, election misinformation narratives, and anti-Muslim messages, with narratives framing Muslims as threats and advocating violence against them. Noted election misinformation themes include rigged voting machines, false polling data for the 2024 Lok Sabha election, and unfounded claims about public figures or institutions. Telegram's content moderation is lax, lacking a proper mechanism for reporting election-related misinformation and hate speech. Its reporting options, including spam and violence, do not adequately cover these issues. Users must report problematic channels under a generic 'Other' category, but Telegram often chooses not to remove misinformation, disinformation, and hate speech.

---

[15] While the development of such a system is timely and crucial, it is important to note that its full operational capability may not be achieved in time to influence the upcoming election cycle, but it will be a vital asset for future engagements.

[16] Although this proposal is tailored specifically for India, the fundamental framework could be applied and adapted in other global contexts.

[17] Coats, Dan. "Parliamentary elections in India increase the possibility of communal violence if Indian Prime Minister Narendra Modi's Bharatiya Janata Party (BJP) stresses Hindu nationalist themes." Business Standard, 29 Jan. 2019, https://www.business-standard.com/article/pti-stories/ahead-of-general-election-possibility-of-communal-violence-in-india-us-spymaster-119012901238_1.html.

In the first half of 2023, Hindutva Watch, a U.S. research agency, recorded 255 hate speech instances against Muslims in India on Telegram, mainly during elections and in BJP-led states. A September 2023 study with India Hate Lab, a DC-based research group, revealed that 80% occurred in BJP-led states, with 70% in areas facing 2023 and 2024 elections involving conspiracy theories, calls for violence, and boycotts against Muslims. Ahead of the 2024 Lok Sabha elections, a rise in misinformation campaigns merging with hate speech aimed to sway public opinion and heighten communal discord was noted. Specifically, conversations in Telegram channels analyzed encouraged attacks on Muslims in Haryana, evoking the 2002 Gujarat riots, among other communal clashes, by playing on fears of a Muslim majority.

Indian fact-checking organizations we spoke with on the ground noted that much of the misinformation they counter is aimed at inciting communal tensions, especially during elections, using advanced technologies like AI-generated deepfakes to seem credible. Recently, misleading claims have spread widely, such as false videos about the Supreme Court of India banning electronic voting machines and inaccurate posts about the Congress Party's election manifesto targeting Muslims and Christians in Telangana, leading to fears of violence. Additionally, outdated misinformation from the 2018 Telangana Congress was circulated as current in 2023.

DataLEADS, supported by Google, launched The PollCheck Election Academy to tackle widespread election misinformation. This initiative has trained over 300 media professionals in over 50 languages, enhancing fact-checking capabilities across various platforms. The Misinformation Combat Alliance has also united 16 Indian organizations specializing in fact-check reporting. Despite these efforts, the issue of misinformation and extremist content on Telegram, its spread to other platforms, and its potential to incite offline violence remains significantly under-researched.

Kendal Gee, MIA Candidate, Columbia SIPA

LABELING AI-GENERATED CONTENT IN INDIA

BLUF: Rapid advancements in generative artificial intelligence (AI) are producing realistic, yet fabricated, audiovisual content, posing significant global challenges that threaten democratic integrity and public trust. **This memo proposes a collaborative pilot study to refine the design of labels for AI-generated or altered content in India, leveraging strategic partnerships with Meta and the Misinformation Combat Alliance (MCA) to develop contextually sensitive labeling strategies.**

Opportunity: India's digital landscape is increasingly affected by AI-generated or altered misinformation, impacting politics and public discourse. Content labeling, the practice of placing visible alerts on media to inform users about its context and origins, is a proposed tool to combat misinformation risks associated AI-generated or altered content. This practice is supported by the Government of India,[18] and Meta has announced its intention to implement such labels on AI-generated images starting in May 2024[19]. While current initiatives aim to label content in multiple languages, including Hindi, Tamil, and Telugu, there remains a critical need for research on label design that accommodates the unique linguistic and cultural nuances of these languages and the regions where they are spoken.

Recommendations: The Global Engagement Center (GEC) is well-positioned to leverage its international partnerships and local insights to lead a pilot study researching how labeled content is perceived in India's most commonly spoken languages, providing technology platforms with context to make informed decisions regarding label design. This project should explore optimal designs for a dual-purpose labeling system that informs users of the content's technical origins and the potential to misinform and should be adaptable across multiple languages including Hindi, Tamil, and Telugu. Collaboration with the Misinformation Combat Alliance (MCA) —an Indian cross-industry coalition with a track record of recruiting participants for media literacy research within the region— and Meta, which can provide the study with access to AI-generated content samples shared on its platforms and has expertise in product design, will be crucial.

Recommendations	Intended Outcome	Metrics for Success
Strategic Partnership Formation	Establish a collaborative framework with Meta and the MCA to leverage their expertise and resources for the pilot study.	Successful acquisition of a wide range of AI-generated or altered content to be used in the study. Establishment of participant recruitment mechanisms that include a variety of demographics, encompassing different income and education levels within each language group.
Pilot Study Design and Execution	Design and execute a structured study to assess user perception of AI-generated or altered content labels using UX best practices such as prototyping and testing varied	Enhanced understanding of how users perceive and react to different label designs, using unlabeled content as the control.

[18] TOI Tech Desk, "Government Issues 'Warning' on AI-Generated Content," TIMESOFINDIA.COM, March 17, 2024, https://timesofindia.indiatimes.com/technology/tech-news/government-issues-warning-on-ai-generated-content-all-the-details/articleshow/108560121.cms.

[19] Nick Clegg, "Labeling AI-Generated Images on Facebook, Instagram and Threads," Meta, February 6, 2024, https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/.

	terminologies and formats across select demographics.	A comprehensive evaluation of label perception across varied media formats, languages, and demographics within language groups.
Iterative Label Optimization	Use the data and insights collected from the pilot study to continually improve and fine-tune AI-generated or altered content labels.	Development of culturally and linguistically tailored labels that are effective and provide actionable guidance for Meta and other platforms in the designing of labels that are noticeable yet minimize misconceptions about content accuracy.

Background: Extensive academic research, combined with insights from Trust and Safety (T&S) professionals and fact-checking organizations in India, have illustrated that warning labels can significantly reduce belief in and dissemination of debunked content.[20] Notably, there is empirical evidence that label design for misinformation impacts user reactions — such as reports from interface designers that corrections displayed alongside misinformation decreased users' acceptance of false claims.[21] This underpins the potential value of implementing a similar labeling process for AI-generated content.

Moreover, T&S professionals have flagged significant risks associated with inadequate label design or the absence of labels altogether. Poorly designed labels can create confusion about the authenticity or intent of content leading to incorrect assumptions that a label is a marker of trust or distrust rather than transparency. Some studies report that "disputed" labels could inadvertently increase the perceived accuracy of unlabeled false stories,[22] emphasizing the need for rigorous assessment of label design.

Building on this, a recent study from MIT Sloan Professor David Rand emphasizes the importance of cultural and semantic nuances in label comprehension. The study illustrates how different terms resonate across linguistic contexts; for instance, in most languages studied, "deepfake" is a widely recognized term, whereas connotations around the term "artificial" are varied. These findings suggest that label design and terminology should be specifically tailored to accommodate India's linguistic diversity to effectively counteract the negative impacts of misleading content.[23]

This pilot study is motivated by the recent initiatives from global platforms like Meta to begin labeling AI-generated content. By furthering research into how labels can be tailored to audiences in India, this initiative aims to improve these tools and enhance the credibility and effectiveness of digital communications.

[20] Cameron Martel and David G. Rand, "Misinformation Warning Labels Are Widely Effective: A Review of Warning Effects and Their Moderating Features," *Current Opinion in Psychology* 54 (2023): 101710, https://doi.org/10.1016/j.copsyc.2023.101710.

[21] Jeff Smith, "Designing Against Misinformation," Design at Meta, December 20, 2017, https://medium.com/designatmeta/designing-against-misinformation-e5846b3aa1e2.

[22] Gordon Pennycook, "Assessing the Effect of 'Disputed' Warnings and Source Salience on Perceptions of Fake News Accuracy," SSRN Electronic Journal (2017), https://doi.org/10.2139/ssrn.3035384.

[23] Epstein, Ziv, Mengying C. Fang, Antonio A. Arechar, and David G. Rand. 2023. "What Label Should Be Applied to Content Produced by Generative AI?." PsyArXiv. July 28. doi:10.31234/osf.io/v4mfz.

Yahui Sun, MPA Candidate, Columbia SIPA

COLLABORATIVE GAME-BASED PLATFORM FOR HEALTH DISINFORMATION ON X

BLUF: GEC should develop an innovative game inspired by Harmony Square, tailored to India's unique cultural and religious landscape, designed to tackle digital health disinformation on X (formerly Twitter). Integrating real-time X data, the game offers personalized, episodic adventures where players combat disinformation, reflecting the immediacy of India's health discussions. This direct X integration empowers users to actively debunk myths, fostering a digitally literate community. By blending engaging gameplay with educational objectives, we aim to enrich public health discourse in India.

Opportunity: Addressing health disinformation on X represents a pivotal opportunity for GEC to make a significant impact within the complex socio-cultural and linguistic fabric of India. The platform's algorithmic emphasis on engagement over accuracy and its global reach makes it a critical arena for intervention against health disinformation, which can have profound implications on public health and safety. By targeting disinformation in this context, GEC can leverage X's extensive user base and real-time information flow to promote digital literacy and critical thinking, addressing disinformation at its source. This approach not only navigates the nuanced landscape of Traditional versus Western medicine and the influence of prominent figures but also respects cultural sensitivities and linguistic diversity. Through innovative solutions like gamified learning, GEC aims to foster a discerning public discourse, making it a strategic priority to ensure a well-informed public and mitigate the adverse effects of health disinformation.

Recommendation: Creating an episodic adventure game necessitates collaboration with X, health experts, the local game company, and fact-checking organizations to ensure accuracy, engagement, and effectiveness. Health experts will provide insights into crafting content aligning with accurate health information and fostering critical thinking. The local game company's expertise in development and cultural integration ensures resonance with the audience, combining education with an engaging gameplay experience. Integrating real-time fact-checking features empowers players to combat disinformation encountered on X, promoting digital literacy. The implementation strategy merges digital outreach and educational collaboration for strong initial engagement and sustained influence. The launch involves vibrant social media campaigns and partnerships with influencers for broad visibility. In tandem, the platform will be woven into educational programs to bolster students' digital literacy and critical thinking. Additionally, a game-based badge system will reward and reinforce fact-checking activities and credible behaviors.

Partner	**Role and Key Initiatives**	**Metrics of Success**
X (formerly Twitter)	Facilitate real-time monitoring and dissemination of alerts on health disinformation. Utilize advanced AI technologies and X's API for real-time health topic incorporation.	○ Increase in alerts generated and disseminated. ○ Reduction in the spread of health disinformation on the platform ○ Number of users earning and displaying badges.
Health Organizations with Experts	Collaborate to ensure game content accuracy and facilitate expert participation in Q&A sessions. Provide reliable advice and clarifications on complex health issues.	○ Increased accuracy of health-related content. ○ Enhanced user trust and credibility of the platform. ○ Positive feedback from user interactions with experts.

		○ Behavioral change in disinformation sharing practices among participants.
Local Game Company	Develop the game, manage day-to-day operations, and integrate educational content based on partner inputs. Create an episodic adventure game combining India in the context of culture and religion.	○ User base growth and active participation rates. ○ Engagement with educational content and behavioral changes among users. ○ Knowledge acquisition and sustained user engagement with the platform.
Fact-Checking and Debunk Organizations	Partner to provide real-time fact-checking features within the game. Enable users to report and verify the accuracy of health-related information encountered on X.	○ Volume and variety of health misinformation instances reported by users. ○ Reduction in the spread of health misinformation narratives before and after the platform's introduction. ○ Increased public awareness and critical evaluation of health information.

Background: In recent years, the proliferation of health disinformation on social media platforms has emerged as a significant public health concern, particularly in India where the digital ecosystem is vast and varied. X, with its expansive reach and influence, has become a critical battleground for the dissemination and contestation of health information. The challenge is compounded by the platform's algorithmic preferences, which often prioritize engagement over accuracy, inadvertently amplifying false or misleading content.

This scenario is further complicated by the cultural and linguistic diversity of India, where health disinformation can exploit traditional beliefs and regional differences, making universal strategies for disinformation mitigation less effective. The debate between traditional and Western medicine, often fueled by influencers and unverified claims, adds layers of complexity to public health communication and literacy.

Recognizing the urgent need for innovative interventions, the proposed project seeks to leverage the power of gamification and digital literacy to combat health disinformation. By engaging users in an interactive and educational experience, the initiative aims to enhance critical thinking skills, promote media literacy, and empower individuals to distinguish between credible and false health information. This approach not only addresses the immediate challenge of disinformation but also contributes to the long-term resilience of public discourse on health matters.

The backdrop of this initiative is a digital landscape where disinformation not only spreads rapidly but also entrenches deeply held beliefs and misconceptions, undermining public health efforts and eroding trust in medical authorities. Against this backdrop, the proposed game-based learning platform represents a timely and necessary intervention, designed to harness the very technologies that facilitate disinformation for educational and empowering purposes.

Gunnar Wainwright, MPA Candidate, Columbia SIPA

AUGMENTING AND SUPPORTING DIGITAL LITERACY MEDIA INTERVENTIONS IN INDIA

BLUF: To combat mis/disinformation in India, the Global Engagement Center (GEC) should **1) host "Misinformation Mirror" series** through the American Center in New Delhi for **U.S. and Indian student cohorts to share experiences about misinformation in their lives, 2)** initiate a content **contest** for American Center student cohorts for 90 second – 3 minute digital literacy video **3) make grants available to expand the reach of and evaluate the efficacy of promising and existing digital media literacy efforts** by Mythos Labs and DataLeads.

Opportunity: In a wide range of regulatory contexts, and especially in India, digital media literacy interventions are increasingly becoming one of the few, viable methods for partially inoculating users to mis/disinformation. There is an opportunity to scale existing programming to new and older users, an equally important cohort with significant social influence at the familiar and community levels. Also, in past interventions participants valued having accreditation for digital media literacy training with DOS branding, a key motivator of participation for future efforts. Finally, few cross-cultural exchange programs focus on the effects of misinformation on daily life, providing an opportunity to foster a "mirror" for cohorts in the U.S. and India. These spaces could help cohorts identify patterns in each other's countries, better equipping them to scrutinize misinformation in their own contexts, potentially inspiring behavior change.

Recommendations:
*Activity 1***:** Leverage existing *Citizen Engagement Exchange Program* at American Center to host "Misinformation Mirror Series", a virtual digital media literacy training and experience sharing space for university age students in the New Delhi area and in the US.

- **Intended Outcomes:** Facilitate storytelling and experience sharing space for U.S. and Indian students. Students can recognize common impacts of mis/disinformation on their lives and communities in different countries. Students also learn critical digital literacy skills and would receive "certification" from American Center.
- **Metrics For Success:** Students can identify common effects of misinformation in familiar/community life, number of students who join the session, an after-action survey to gather student feedback and interest for future training and experience sharing spaces.

Activity 2: Call for proposals targeting participants of "Mirror" series in New Delhi to create 90 second to 3-minute video about lateral media consumption and the emotional appeal of misinformation after the exchange series.

- **Intended Outcomes**: Facilitate further exchange, collaboration, and creative thinking among participating students for addressing issues. Two new pieces of comedic digital media literacy content for university student cohorts to share with themselves, friends and families.
- **Metrics for Success**: Two videos -90 second to 3-minute in length - formatted for Instagram reels and YouTube, Positive anecdotal feedback from participating students on collaboration experience and developing media/communication skills, and content views and engagement on selected videos.

Activity 3: American Center or appropriate funding mechanism makes funding available for existing, credible digital media literacy programs to expand their reach and for randomized control trials (rct) to evaluate these interventions for efficacy.

- **Intended Outcome:** Expand the reach of current digital media literacy efforts by organizations such as Mythos Labs (digital format) and DataLEADS (in-person) to cohorts and communities not yet reached by FactShala program.

- **Metrics for Success:** Fund one digital media literacy initiative formatted for digital consumption and one in-person digital media literacy campaign, number of views and engagement on media literacy campaign for digital consumption, number of people reached through in-person intervention, conduct post-intervention on each literacy campaign format to evaluate efficacy.

Background

The current regulatory climate in India is less conducive to a broad range of options available for pre- and post-facto efforts to combat mis/disinformation. The fact checking industry in India is facing more scrutiny against its operations leading up to the 2024 elections. Tools and techniques at the technical level for social media and messaging platforms, such as WhatsApp group size limiting, fact-checking tiplines and limits on content sharing, are important choke points against misinformation spread. However, social media platforms themselves are facing an increasingly challenging regulatory environment. For the widespread implementation of these technical changes to continue, one source at a major platform said they need more diplomatic support from the U.S. government to deconflict this regulatory environment.

Digital media literacy interventions maintain an advantage in that they can build individual and community resiliency to misinformation without targeting specific content that may drive political disagreements. The Democratic Erosion Consortium (DEC) survey of Global South Misinformation experts found digital media literacy interventions as the preferred intervention for this sample. DEC notes that there have only been 16 randomized control trials (RCTs) on digital media literacy interventions, less than half of which were in countries experiencing digitization. In DEC's database, only one example of a literacy intervention in India was accompanied by a RCT. In this survey, successful interventions focused on three outcomes: (1) **discernment**, which is the ability to distinguish true from false information; and/or (2) the **intent** to share false (relative to true) information and (3) **emotion** or enhanced understanding of the emotional appeal of mis and disinformation.

Our research identified some promising approaches to media literacy interventions that could be replicated, funded and/or leveraged to expand reach. DataLEADS, a digital media and technology company based in New Delhi, implemented a pan-Indian, in-person digital media literacy program, FactShala, which is supported by Google News Initiative. FactShala uses a "train the trainer" approach, conducting over 2000 trainings across India, in multiple languages to reach over 70,000 participants. A team of about 250 trainers, including journalists, fact checkers and media educators, trains communities, especially in more rural, less digitally connected areas. There is an opportunity to leverage DataLEADS efforts and network by expanding their current programming to reach more and new cohorts or evaluating their current interventions through RCTs, adding valuable and needed research to this field.

Mythos Labs, a company focusing on innovative approaches to combating misinformation in India and a grantee of previous DOS funding, launched the SMILE digital literacy campaign, a video series featuring a prominent comedian using comedy to communicate key literacy techniques and concepts to viewers. A representative from Mythos Labs highlighted that leveraging comedy to communicate key concepts resulted in this campaign being one of the most successful to date in both reach and impact (based on post-completion surveys). Additionally, participants valued having accreditation for digital media literacy with DOS branding, a key motivator of participation especially amongst university age cohorts. Mythos Lab's #Checkkiya campaign specifically targets older users, using comedy again to create content about situations many younger users may find themselves in when interacting with older family members. The need to invest in digital literacy campaigns targeting older users was a consistent finding in our interviews with different stakeholders. DOS can further leverage Mythos Lab's expertise in producing digital campaigns on media literacy to reach more and new cohorts, especially old users, an often overlooked but socially important cohort.

Ashwini Sivaganesh, MIA Candidate, Columbia SIPA

ENGAGING THE INDIAN DIASPORA IN THE UNITED STATES

BLUF: Establish a triannual forum with Indian diaspora communities that will allow for open dialogue and explore programming opportunities with the U.S. Department of State.

Opportunity: The United States hosts the largest Indian diaspora globally, with close to 5 million people of Indian origin or descent heavily concentrated in California, New Jersey, and Texas. The Indian diaspora primarily stays connected to India through contact with family and friends on social media and messaging apps, forming a tight-knit network that's challenging for outsiders to access. Due to the high degree of trust amongst members, these private group chats can sometimes become conduits for the spread of disinformation. Engaging actively and effectively with the Indian diaspora provides the GEC a way to understand and integrate into this network, while also building meaningful relationships. This could significantly benefit the GEC's objectives to improve the global information-sharing ecosystem and enhance its broader engagement strategies, especially during times of crisis.

Recommendations	**Intended Outcomes**	**Metrics for Success**	**Goal**
Cultivate diverse network of contacts from existing Indian diaspora organizations to create a triannual forum	A formalized line of communication between the U.S. Department of State and the Indian diaspora	Engagement with all if not most of the diaspora community, quantitatively and qualitatively tracked through an organization's membership and impact	Efficient and reliable channel to the Indian diaspora which can be used to mitigate problems that arise within or related to the community
Curate resources, tools, and information relevant to the Indian diaspora that can be shared during forums, whether that be from the State Department, within the diaspora community, or other relevant experts/stakeholders	Further U.S. interests and values at home and aboard in India	Depending on the program/initiative, this would entail tracking engagement and/or observing systematic changes	

Recommendations:

Activity 1: The GEC should first identify and organize Indian diaspora community contacts from various existing organizations. These contacts should include local leaders and organizers who will act as representatives for their respective communities at a triannual forum. The focus should be on organizations formed around shared cultural backgrounds (e.g. Tamil Sangam, Gujarati Samraj, United Sikhs, etc.), as well as those consisting of members from specific professional fields or industries (e.g. American Association of Physicians of Indian Origin, US-India Business Council, The IndUS Entrepreneurs, etc.). Once the GEC acquires these contacts, it can then broaden its outreach efforts to target individuals within the Indian diaspora who are not affiliated with these organizations. This approach will help ensure comprehensive coverage and engagement within the community.

Activity 2: The U.S. Department of State should host the triannual forum, with participants from the GEC, one or two representatives from each Indian diaspora organization, and other key stakeholders. The GEC would be responsible for setting the agenda as it relates to disinformation and related issues, which could include items like:

- Updates from the State Department on relevant topics for awareness, transparency, and information on programs, initiatives, or grant opportunities

- A presentation by a diaspora community organization, sharing about their group's focus or discussing concerns and challenges they've observed
- Presentations from relevant stakeholders, which could involve experts from within the community or external figures. For instance, journalists covering India could discuss media literacy, healthcare professionals might tackle vaccine hesitancy, among other topics
- Q & A session relevant to the meeting's presentations

Intended Outcomes: Establishing a structured forum between the GEC and the Indian diaspora will create a vital relationship that enhances communication and collaboration, which is pivotal during crises or when addressing specific challenges, such as misinformation and disinformation. This triannual forum facilitates the exchange of information, enabling the GEC to share details about available programs and resources for U.S. residents, and to discuss current issues related to the Indian diaspora and India.

This strategy emphasizes positive engagement and reinforcement, as well as an active interest from the GEC to strengthen relationships with the Indian diaspora. This ensures the community does not feel negatively singled out, especially during crises or problems. It also familiarizes the diaspora with diverse tools and programs from the State Department and other agencies, designed to effectively manage and mitigate the effects of crises. An illustrative example of this is addressing vaccine hesitancy during the COVID-19 pandemic. Through the forum, the GEC could have provided accurate information for diaspora leaders to disseminate within their communities, thereby amplifying the reach and impact of critical health messaging. Similarly, addressing misinformation and disinformation can be achieved by involving relevant experts and stakeholders in the discussion.

Metrics For Success:

- Success will be gauged by the number and diversity of diaspora organizations actively participating, alongside growth in their participation in triannual forums
    - Metrics will include event attendance, program participation rates, and observed systematic community changes due to these initiatives
- Effectiveness will be assessed by the creation of diverse and inclusive communication channels between the State Department and a broad cross-section of Indian diaspora organizations
- Success will also be measured by the promotion of U.S. interests and values, evidenced by policy influences, cooperative projects, and changes in perceptions within the United States and India
- Continuous feedback collection and the subsequent adaptation of strategies and initiatives based on this feedback will indicate success in meeting the community's needs and expectations

Background: The U.S. government and Indian diaspora previously collaborated successfully on the U.S.-India Civil Nuclear Agreement, however, now more than ever, continuous engagement with the diaspora is essential, particularly as disinformation spreads to and from the Indian diaspora. With more than 32 million people, it is the world's largest diaspora and boasts prominent leaders who are gaining influence in global politics and helming some of the world's leading companies. This community's significant economic contributions, including nearly $125 billion in remittances in 2023, about 3.4% of India's GDP, along with high educational and professional attainment—80% of Indian Americans have college degrees and median household incomes double the national average—highlights their role in strengthening U.S.-India economic and innovative partnerships. Engagement with Indian diaspora communities will, inter alia, enhance information accuracy, media literacy, empower diaspora voices, and support disinformation research to uphold democratic values, and foster global stability and prosperity.