Understanding Universal Dynamics of Misinformation: Prevalence, Beliefs, and Solutions

Overview

This research aims to explore the universal dynamics of misinformation by recruiting a diverse group of participants: Indian, Colombian, and Mexican Americans, along with people from India, Colombia, and Mexico. We will use effective data donation methods aimed at sampling these groups from multiple social media platforms to understand the cultural and linguistic nuances of misinformation. By analyzing the collected social media data, our study will identify patterns of misinformation spread across different cultural and linguistic contexts, which is crucial for understanding both local and global characteristics of misinformation. Next, we aim to causally measure the beliefs and behaviors influenced by misinformation, providing insights into how narratives are consumed and absorbed across cultures. To address the challenges faced by fact-checkers, we will develop bottom-up solutions for misinformation, incorporating state-of-the-art large language models to offer on-device fact-checking resources. These solutions will be culturally sensitive, ensuring tailored approaches are developed for each setting, thereby reducing the burden on fact-checkers and enhancing the efficacy of misinformation management.

Intellectual Merit

The proposal outlines significant intellectual contributions to the field of misinformation research, by filling critical gaps related to the consumption, spread, belief, and mitigation of misinformation. First, it introduces advanced methods for data collection that preserve privacy, using opt-in data donation tools adaptable across major social platforms like WhatsApp, YouTube, and Facebook, which promise to set new standards for ethical, large-scale data gathering. Second, the project plans to refine misinformation detection across modalities – text, images, audio, and video – employing cutting-edge AI techniques such as mutlimodal embeddings. Third, it deepens the theoretical understanding of misinformation's spread, offering a perspective that considers both universal patterns and unique, culturally specific characteristics. Fourth, this research integrates social science methods with technical analyses to trace misinformation's impact, thereby enhancing our understanding of how misinformation narratives causally affect changes in beliefs and behaviors. Finally, it offers innovative, user-centric solutions for misinformation, focusing on on-device, privacy-preserving technologies that empower users to identify and counter misinformation effectively.

Broader Impacts

The proposed research stands to significantly impact society by advancing the understanding and the development of solutions to combat misinformation within underrepresented communities in the US. By constructing a unified theory of misinformation spread among diverse linguistic and cultural contexts, this project will guide global and local strategies, enhancing efforts by policymakers, educators, and technologists to craft interventions tailored to varied populations. Particularly, it emphasizes data collection from minority groups such as Indian American and Latinx communities in the U.S., aiming to illuminate misinformation dynamics within these communities. Additionally, the research will pioneer scalable, proactive solutions with on-device, privacy-preserving technologies that enable individuals to assess and counter misinformation effectively in real-time. This approach ensures practical, broad implementation, supporting individuals' capacity to navigate information critically. Moreover, the project will educate the next generation of scholars and practitioners, offering students hands-on experience in cutting-edge data science, AI, and social science methodologies. By engaging local communities and leveraging findings for public outreach, the research will bolster community resilience against misinformation, making significant educational and societal contributions.

1 Introduction

Social media platforms like WhatsApp, YouTube, and Facebook, each boasting over 2 billion active users globally, connect users across diverse linguistic and cultural landscapes, enabling (mis)information to propagate swiftly and widely. Misinformation, ranging from benign rumors to malicious fabrications, can significantly distort public perception and influence behavior on a massive scale [1, 2, 3, 4]. The challenge lies in understanding and addressing how misinformation narratives are crafted, shared, and perceived by global audiences, particularly as these narratives adapt to the cultural and contextual nuances of different user groups.

Existing studies on large scale social media data suggest that misinformation is rare and shared mainly by a minor subset of users [5, 6, 7, 8]. However, these studies have been criticized for focusing too narrowly on the definition of misinformation, primarily focusing on political misinformation in the US, and linked to particular URLs [2]. This narrow focus neglected the broader spectrum of misinformation, including benign rumors that can also have substantial impacts. Additionally, these efforts frequently overlooked the real-time evolution of misinformation campaigns and failed to adequately track the cross-platform spread of these narratives [9]. Understanding and addressing misinformation effectively presents significant challenges due to its complex and dynamic nature, which such naive approaches often fail to capture. Misinformation varies significantly across different cultures, languages, and media formats, making it difficult to create universal detection systems. Earlier efforts predominantly focused on particular regions, typically the U.S., and centered around political events like elections [10, 11]. These studies, while insightful, do not generalize well to the diverse and global nature of social media, where misinformation pervades everyday topics and various forms of media.

This proposal differs from previous methodologies by incorporating a perspective that accounts for the nuances of misinformation spread in different cultural settings. We utilize a robust sampling strategy that includes participants from multiple countries with shared languages and different cultural populations – Indians, Mexicans, and Colombians – living in the US and their home countries. This strategy enables us to observe and analyze the commonalities and variations in how misinformation is produced, spread, consumed, and believed across various global communities. Recognizing that misinformation can transition from local incidents to global crises, as evidenced by various incidents of localized rumors escalating into widespread misinformation campaigns [3, 12, 13], our sampling strategy is designed to effectively differentiate and tackle both scales of misinformation. Specifically, we aim to answer the following research questions:

RQ1. Consumption: What are the prevalent viral themes of misinformation across different linguistic and cultural groups? and how much overlap do they have? Are there any 'universal' themes of misinformation consumption? Specifically, can we develop a unifying theory of misinformation themes in diverse settings?

RQ2. Spread: How do misinformation narratives evolve and spread within and across different cultural and linguistic groups? Can we find evidence of temporal causality, establishing pathways for foreign influence, e.g., do narratives beginning in India spread to Indian Americans living in the US?

RQ3. Belief: Does the consumption of misinformation impact belief and behaviors of users? Can we provide causal evidence of the influence of misinformation?

RQ4. Solutions: How effective are current top-down strategies in combating misinformation, and is there a need for culture-specific solutions? Can we build privacy preserving, bottom up solutions that users trust?

2 Background and Intellectual Innovations

The proposal makes notable intellectual contributions across multiple areas, addressing critical challenges and advancing the state-of-the-art in misinformation research.

Methods for Data Collection. The study of misinformation has a long-standing history, tracing back to foundational works in the late 20th century [14]. However, it has gained significant momentum since 2015 [15], especially within the realm of computer science. Most of the research in this field has predominantly focused on analyzing misinformation through relatively small, single-platform datasets that are primarily text-based and in English [16]. Despite this growing interest, both in computer science and social sciences, there remains a noticeable lack of investment in developing robust infrastructures to effectively measure the prevalence of misinformation. The cross-platform spread of misinformation, particularly at a user level, remains a challenging frontier. Each platform's unique affordances significantly influence how misinformation is disseminated and consumed, making it difficult to track misinformation across different media environments [17].

This research will pioneer new methodologies for sampling and adaptively collecting data in a privacy-preserving manner. By developing innovative opt-in data donation and adaptive surveying tools, we will enable large-scale, ethical data collection from major social media platforms, including WhatsApp, YouTube, and Facebook. This methodology allows us to collect, for the first time, data from multiple platforms from the same user, over time, along with their demographics, and the ability to conduct surveys adaptively with users [18]. These methods are designed to be platform-agnostic, ensuring broad applicability and significant impact across diverse digital environments.

Misinformation Detection Across Multiple Modalities. The challenge of automatically detecting misinformation has engaged the computer science community for over a decade [19, 20]. A comprehensive survey by Shu et al. (2022) captures the scope of methodologies employed in this domain, noting that the majority of existing approaches treat misinformation detection as a classification problem [21]. Historically, research in this area has predominantly focused on textual content, with deep learning models being a common approach [22]. However, the emergence of advanced computational techniques has spurred interest in multi-modal methods that integrate various forms of media, such as text, images, and video [23, 24]. These multi-modal approaches are promising because they reflect the complex, varied nature of how misinformation is typically disseminated across different platforms. Despite these advancements, most detection architectures concentrate on developing new algorithms that either contextualize content or aim to provide explanations for their outputs [16, 25].

Yet, there remains a fundamental gap in understanding why misinformation detection is computationally intractable, largely due to its inherently subjective nature. Moreover, as Juneja et al. (2022) highlight, many methods that perform well in academic settings fail to translate effectively into practical applications [26]. This discrepancy underscores the challenges of applying theoretical models in real-world scenarios.

Leveraging cutting-edge advancements in joint multi-modal embeddings along with experts in the loop, this project will enhance the detection of misinformation in various formats, including text, images, audio, and video. By employing state-of-the-art techniques such as ImageBind [27] and joint embeddings [28], we will improve the accuracy and efficiency of misinformation detection, addressing the complexity and diversity of modern misinformation narratives.

Theories of Misinformation Spread. Misinformation exhibits unique propagation characteristics, often spreading faster, wider, and deeper than factual information. This phenomenon is well-documented in studies by Vosoughi et al. (2018) and Shao et al. (2018), which analyze the virality and reach of misinformation [29, 30]. Understanding who consumes misinformation is equally crucial; research by Guess et al. (2019) and Grinberg et al. (2019) provides insights into the demo-

graphics and psychological profiles of individuals more likely to engage with misinformation [6, 7]. Furthermore, the motivations behind forwarding or spreading misinformation, as discussed by Chen et al. (2015), reveal social and psychological triggers that exacerbate this issue [31].

Global events, such as the COVID-19 pandemic, have underscored the universal nature of misinformation spread, affecting diverse populations across different geographies [32, 33, 34]. These events catalyze the rapid dissemination of globally relevant misinformation narratives concerning topics like vaccines, 5G technology, and immigration. However, while qualitative data highlights the universal resonance of these narratives, quantitative studies remain scarce. There is a critical need for more comprehensive quantitative research to understand how these universal consensuses on misinformation are formed and how they are tailored to fit local contexts [35].

Concurrently, misinformation can also manifest in highly localized contexts, sometimes escalating to significant societal issues. Instances such as the measles outbreaks in the United States [36] and widespread vaccine hesitancy [37] exemplify how localized misinformation can have substantial, tangible consequences. Moreover, localized rumors, such as those about child kidnappings, have occasionally gained momentum to become issues of global concern, as seen in cases documented across various countries, leading to severe offline consequences such as lynchings and violence [3, 12, 13].

This research will contribute to the theoretical understanding of misinformation as a global and local phenomenon. We will explore the universal aspects of misinformation spread, as well as the unique characteristics that are culturally and contextually specific. This dual focus will provide a comprehensive framework for understanding how misinformation propagates across different linguistic and cultural landscapes.

Understanding the Impact of Misinformation. Understanding whether individuals genuinely believe the misinformation they share, or whether they disseminate it for other reasons, such as social status or in group allegiance, remains a largely unresolved issue in the study of social media dynamics. Research by Pereira et al. (2018) and Reinero et al. (2023) explores these motivations, suggesting that the act of sharing misinformation can often be more about identity signaling than genuine belief [38, 39]. However, the effects of believing misinformation are undeniably serious, as studies like Loomba et al. (2021) demonstrate with regards to vaccine hesitancy [40].

Yet, establishing a direct link between exposure to misinformation and tangible offline consequences remains elusive. Persily (2020) notes the lack of solid evidence tying misinformation exposure to real-world actions, which aligns with broader findings by Kalla and Broockman (2018) that the persuasive effects of information campaigns are generally minimal [41, 42]. Despite this, qualitative studies, such as those analyzing the role of social media in events like the January 6 Capitol attack, suggest that misinformation can indeed have serious real world impacts [43].

By integrating adaptive surveying tools in our pipeline, we enable social science experimentation with advanced technical methods such as Hawkes processes, thus offering causal evidence of misinformation consumption on beliefs. We will examine both the temporal and spatial patterns of misinformation dissemination, providing insights into the impact of these narratives on different communities and their behavioral responses.

Developing Bottom up Solutions for Misinformation. Current interventions to combat misinformation can be broadly categorized into five main approaches, each with its specific strategies and associated challenges: (i) Boosting/Prebunking [44, 45], (ii) Providing media literacy [46, 47], (iii) Nudging/Accuracy prompts [48], (iv) Fact checking [49, 50, 51], and (v) Labelling [52, 53]. While these interventions have proven effective in experimental settings or over short durations, they face substantial challenges in real-world application [54, 55, 56]. Current solutions often adopt a top down 'push' approach where the burden of implementing the corrective tools falls solely on the platforms, which may not be inclined to incorporate such solutions due to a lack of incentives. In the case of bottom up solutions like tiplines [51], they depend on users without providing them

with a value proposition for their time. This leads to a situation where the tools available are underutilized, and misinformation continues to propagate.

The proposed research will innovate new strategies for combating misinformation that prioritize bottom-up, user-centric approaches. We will explore on-device, privacy-preserving solutions that empower users to critically assess and counter misinformation in real-time. By harnessing the future potential of AI, these solutions will be designed to be adaptive, scalable, and respectful of user privacy, offering a paradigm shift from traditional top-down interventions.

3 Methods

In this section, we will first describe our data collection methodology (Section 3.1), followed by our four methodological contributions: (i) detecting misinformation in multi-modal, real-world setting (Section 3.2), (ii) developing insights on the universal properties and spread of misinformation (Section 3.3), (iii) understanding the causal impact of misinformation on beliefs (Section 3.4), and, (iv) developing solutions to alleviate misinformation (Section 3.5).

3.1 Sampling And Data Collection

One of our key contributions will be the implementation of a novel and ambitious data collection strategy. Unlike many studies that rely on convenience samples, which severely limit the generalizability of their findings, our approach integrates advanced methods from social science and survey methodology to actively sample and survey participants. We plan to recruit users from multiple cultural and linguistic backgrounds and obtain data from multiple platforms from these users. This strategy is designed to provide a comprehensive and representative understanding of misinformation dynamics. Our strategy requires significant infrastructure, much of which has already been developed and deployed by the PI in various contexts.

3.1.1 Data Donation Tools

We will develop and deploy data donation tools that facilitate data collection from multiple platforms, including WhatsApp, YouTube, and Facebook [57, 58, 59]. These social media platforms each have over 2 billion active users worldwide, making them ideal for capturing a broad spectrum of digital consumption behaviors. Users typically engage with friends on WhatsApp, follow pages of interest on Facebook, and consume news on YouTube, providing a rich and diverse dataset. This list is not exhaustive: there might be other social networks of interest that appear or existing ones might wane out.

Our data donation tools allow users to easily export their data, which platforms are legally required to permit under regulations such as the GDPR's "Right of Access" [60]. These tools are user-friendly, accessible to individuals with low literacy, and can be used to donate their data even on their phones within five minutes. The data donation process captures both user consumption patterns and choices, along with detailed demographic data.

Additionally, we will deploy a WhatsApp bot to adaptively survey and request information from the user in an easy and non-intrusive manner. Such bots have been used successfully at scale in previous studies to mitigate misinformation and deliver treatments [61]. This bot will collect weekly information on users' beliefs and actions related to the (mis)information narratives they consume, ensuring minimal intrusion while providing valuable data. The PI has extensive experience with data donation, both in-field and online [58, 59] (partially funded by previous NSF projects, see Section 5), and has collaborated with multiple partners to conduct large-scale surveys [62].

3.1.2 Sampling Users And Recruitment

Our sampling strategy is designed to explore the local and global patterns of misinformation. We will primarily sample two different populations which allow us to measure differences in the local and global nature of information spread. (i) **Cultural Coherence**: We will start with Indians living in various parts of the world: in India, the U.S., and Europe. The Indian diaspora is the largest globally, with over 20 million people, constituting significant populations in the U.S., the UK, and parts of Europe; and, (ii) **Linguistic Coherence**: Our second sample will focus on involving participants from two countries in Latin America – Mexico and Colombia, representing two of the biggest Latinx immigrant populations in the U.S. [63]. We will recruit participants both in the U.S., and in their respective home countries. Note that the methodology developed in this proposal is not specific to these countries and can be easily extended to other countries/contexts.

Figure 1 shows a sample schematic of the populations we recruit and an outcome we plan to achieve. By comparing the overlap in consumption of misinformation narratives across these different cultural and linguistic communities, we can identify different properties of how misinformation spreads across different cultural and linguistic communities. For instance, how much overlap exists between narratives for Indians living in India versus those residing in the United States? Do they consume the same narratives? Does misinformation originating in India influence beliefs within the diaspora in the U.S. and Europe, or vice versa? Moreover, it would be interesting to understand the universal nature of misinformation narratives. The overlap of content and narratives might be expected across a community sharing a nationality/culture, but how common is information overlap across very different countries which share a language?, say between Mexico and Colombia? For example, consider the rumors of Joe Biden being communist, which are prominent in certain regions in the Latinx diaspora in the US [64]. Do users from both countries consume such similar rumors? Do such rumors originate in the home countries and spread to the US? Our priors on these questions indicate that there might not be much overlap, but small scale pilots on WhatsApp by the PI in Colombia and Brazil indicate a non-trivial overlap in narratives across the two countries.

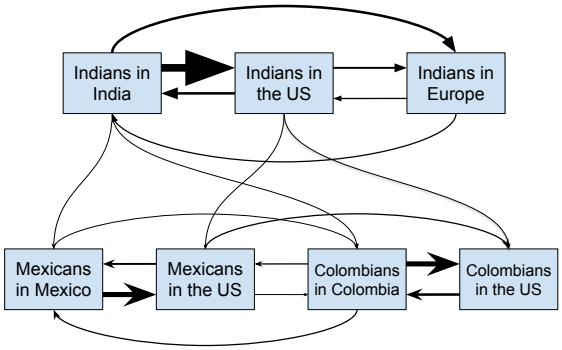

Figure 1: Schematic showing the flow of (mis)information between different populations in our study. The thickness of the arrows would indicate the amount of (mis)information flow from one country to the other. The study will use actual data to quantify this overlap and flow of (mis)information, even though the diagram currently shows only hypothetical values.

We will be recruiting 300 participants at each location trying to cover various demographics like age, and gender. The 300 participants in each location are sampled depending on the local survey data, so we can use survey weighting techniques [65] to obtain population level estimates. The recruitment will be done primarily through online sample providers like Lucid, CloudResearch, targeted Facebook ads, alongside on-ground NGO partners such as The London Story in the UK and Europe (see letter of support). The PI brings extensive experience in working with survey firms and NGOs to recruit survey participants [58, 59, 62, 66].

3.1.3 Recruitment Flow

The recruitment process will vary by context, with users recruited either in person or online. Once recruited, participants will complete a pre-survey capturing demographics, social media usage, consumption habits, political support, and attitudes. Following this, participants will use a custom web tool to consent to data donation from the platforms they use. After the initial data collection, participants will complete a post-survey and sign up for regular updates from our WhatsApp bot. The bot will secure consent to recontact participants for future surveys, allowing regularly polling about their exposure to misinformation narratives. Combining surveys with digital data collection methods is ambitious but feasible given the PI's previous success in similar projects [58].

3.2 Identifying Misinformation

Given the vast collection of multi-modal, multi-lingual, and multi-platform data, the first challenging task would be to detect misinformation. Though there is ample work in the space of automated detection of misinformation (as detailed in Section 2), most methods are still not applicable in real world, particularly in a complex situation like ours. In this section, we propose a human in the loop methodology to identify and tag misinformation from our collected data. Our approach involves embedding multi-modal content into the same space, cluster the embeddings to identify coherent narratives, and having human experts annotate these narratives for misinformation. These embeddings and narratives can be further used to train machine learning models to detect misinformation automatically.

3.2.1 Narrative Detection

We first define the task of narrative detection, which refers to the identification and analysis of thematic continuity across various forms of content, encompassing text, images, audio, and videos. Narratives serve as a fundamental structure through which individuals make sense of the world, encapsulating a sequence of events or concepts that evolve over time. For example, a narrative might involve various media representations of a presidential debate, where the debate is discussed in videos on YouTube, depicted through memes on Facebook, and debated in text messages on WhatsApp. Each platform offers a different modality – video, image, and text, respectively – but they all contribute to a cohesive narrative about the debate.

Narrative detection (also referred to as story detection [67]) is a well-established line of work, particularly for news articles – [68] provides a detailed survey of over 100 papers in the field. However, one of the key missing aspects in existing literature and a primary contribution of this work is the expansion of narrative detection to multi-modal settings. We define the problem of narrative detection as follows:

Problem Formulation: Let us define a set of contents $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ shared by a user u_i across a set of platforms $\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$ within a specific time duration T (e.g., a week). Assume these contents are aggregated into a stream A. The problem of narrative detection is to identify and group subsets of $\mathcal{C}$ that share common thematic elements. This process is formalized as follows:

1. Continuously receive content stream A, which consists of elements from $\mathcal{C}$ occurring within sliding time windows $\{W_1, W_2, \ldots\}$.
2. For each window W_k, update a set C_{W_k}, which is a subset of contents from A that falls within W_k. Each C_{W_k} should reflect coherent narratives emerging or evolving within the time span of W_k.

3. Apply unsupervised learning techniques to group contents in C_{W_k} based on thematic similarity. This grouping forms the basis of narrative clusters that span multiple platforms in $\mathcal{P}$.
4. As new contents are added to the stream A, incrementally adjust the narrative groups in C_{W_k} to include new insights or to refine existing narratives.

One potential way to detect narratives will be to use methods that can process and analyze multi-modal content. Using state-of-the-art techniques like ImageBind [27] or Video LLaMA [28], or LlaVa [69] we will embed various modalities into a unified space. ImageBind (or other multi modal embedding techniques) utilizes large-scale paired datasets, including image-text pairs, and combines them with other naturally occurring data pairings such as video-audio. By learning a single joint embedding space, ImageBind effectively aligns text embeddings with various other modalities like images, video and audio. This alignment facilitates zero-shot recognition capabilities in these modalities, enabling the model to recognize elements it has not explicitly been trained on, without the need for direct semantic or textual associations. These open-sourced models built on large web-scale data can then be fine tuned for our own datasets. Embedding multi modal content into a unified space is currently fastly developing and we expect it to be practical and applicable on in-the-wild datasets over the next few years. Such embeddings then allow us to build down stream analysis tasks such as classifiers capable of detecting misinformation and other harmful content across different platforms and content types.

The use of multi-modal embeddings is one way to identify narratives. This is not guaranteed to yield the results we anticipate. In case those methods do not work, we will try out classic alternatives involving other signals. This could involve methods such as the construction of user-content bipartite graph and using community detection methods to identify consistent content narratives. Another fool proof approach would involve transcribing everything to text (e.g., speech-to-text for audio, OCR/image captioning for images, video descriptions) and using text analysis methods like topic modeling to identify coherent narratives.

3.2.2 Manual Narrative Annotation For Misinformation Detection

To ensure the accuracy and relevance of our narratives, as well as to identify misinformation narratives, we will employ local experts, such as fact-checkers and journalists to go over the identified narratives. These annotators, who are familiar with the cultural and contextual nuances of the content, will meticulously label the narratives for various issues, including misinformation, political polarization, and hate speech. By involving local experts, we can ensure that the annotations are contextually accurate and reflective of the community's specific challenges. Depending on the data volume, we will sample and annotate a subset of narratives. The sample will be carefully chosen to get prevalence estimates at population level [65].

The manually annotated dataset, encompassing multiple platforms and modalities, is a significant contribution to the research community. Currently, there is no existing dataset that offers a clean, multi-platform, multi-modal collection of annotated content. Our dataset will fill this gap, providing a valuable resource for future research in misinformation detection and analysis.

Next, we use the annotated data to build classifiers to detect misinformation automatically. The classifiers will be applied to the larger dataset to detect misinformation and political propaganda at scale. This large-scale application will enable us to quantify the prevalence of these issues across different platforms and modalities. In case the classifiers do not perform well for this task, given the scale of the manual annotation, and the precise sampling methodology, we can still provide estimates on the prevalence of misinformation at a population scale.

3.3 Prevalence And Spread Of Misinformation

Understanding the prevalence and spread of misinformation is pivotal in framing effective interventions and policies. This section addresses key aspects of misinformation dynamics, providing a structured overview based on a multi-dimensional approach.

Theorization and Categorization of Misinformation Types. Initially, it is essential to develop a theoretical framework that delineates the various types of misinformation. This involves a mixed-methods approach that categorizes misinformation into distinct categories, such as orchestrated, explicitly coordinated, information operations by bad actors [70] and the emergent, organic behaviors of an online crowd [71]. By defining these categories specifically in our data collection context clearly, we can better understand the mechanisms through which misinformation arises and propagates. Following this categorization, we can examine how narratives like political rumors, health misinformation, or cultural myths transfer or transform across different cultures, countries and demographics.

Universal Characteristics of Misinformation. A critical aspect of misinformation is its universal features and the extent to which these are influenced by either top-down (institutionally driven) or bottom-up (community-driven) processes. Given our extensive and targeted data collection (Section 3.1), we are rightly placed to answer and analyze universal misinformation patterns to reveal commonalities and differences across cultural and linguistic boundaries. This includes investigating whether certain misinformation themes are universally prevalent and how cultural contexts shape the reception and spread of misinformation. Specifically, we plan to answer questions such as those posed in Section 3.1.2, providing vital insights for developing universally applicable yet locally sensitive misinformation mitigation strategies (see Section 3.5).

Cross-Platform Dynamics. Misinformation does not confine itself to a single platform but proliferates across multiple digital ecosystems. Given our effective data collection strategy collecting per user consumption behavior on multiple platforms, we can answer questions on the cross-platform dynamics of misinformation. Previous efforts to match users across platforms post-hoc have been limited by the reliance on usernames or other identifiers [72]. Utilizing our joint embeddings created in Section 3.2.1, we can trace how a misleading video on one platform correlates with an image or text on another, providing a comprehensive view of cross-platform misinformation dynamics. This analysis not only helps in understanding the spread of misinformation across platforms but also in the role of platform affordances, thus allowing us to define platform specific as well as general tools for moderating misinformation effectively.

3.4 Belief And Influence Of Misinformation

3.4.1 Identifying The Causal Impact Of Misinformation On Beliefs

One of the biggest challenges and open questions in the misinformation literature is a clear picture of misinformation on beliefs [41]. Understanding the causal impact of misinformation on user beliefs and behaviors, such as voting or other offline actions, presents significant challenges due to the multitude of influencing factors and concurrent events. In dynamic (social) media environments, users are exposed to a plethora of narratives, interactions, and external stimuli that can simultaneously affect their beliefs and actions. Isolating the specific impact of misinformation from other influences requires sophisticated analytical techniques and a comprehensive approach to data collection and analysis.

Our methodology leverages the ability to identify multi-platform and multi-modal narratives in real-time using the automated techniques (Section 3.2.1) and experts in the loop (Section 3.2.2), combined with the deployment of a WhatsApp bot (Section 3.1.1) designed to reach out to par-

ticipants. This infrastructure enables us to measure the beliefs and impacts of misinformation in real-time and enables clean causal estimates.

The process is as follows: Each week, we gather data from users across various social media platforms to identify prevalent misinformation narratives. The WhatsApp bot actively engages users, regularly collecting information about their beliefs and behaviors concerning the narratives they encounter. This includes all types of narratives, not solely those classified as misinformation. Users are prompted with a standardized set of questions to ascertain whether they have encountered a specific narrative, whether they believe it, and whether they plan to act on it.

Subsequently, we organize users into groups using standard matching techniques such as propensity score matching [65]. This allows us to form well-defined treatment and control groups by matching users based on demographics, activity levels, and exposure to misinformation. Essentially, the two groups are identical in terms of their location, demographic characteristics, and social media activity, differing only in their exposure to misinformation narratives. By comparing beliefs and actions across matched sets of users, we can identify the influence of different misinformation narratives.

Our approach enhances our understanding of misinformation by collecting and analyzing data in real-time. This ensures that we capture the immediate impact of misinformation, while the longitudinal data collection allows us to observe changes over time. This real-time monitoring contrasts sharply with most existing interventions that rely on synthetic or hypothetical scenarios due to limitations in data collection infrastructure [73, 74]. Our system's ability to test beliefs and behaviors based on the actual content consumed by users ensures that our causal estimates are both precise and representative of true dynamics in the field.

Challenges. The success of our project hinges on several interrelated components that each present their own set of technical and operational challenges. Firstly, the effective detection of narratives through automated systems must be robust. Following this, incorporating humans in the loop to identify misinformation effectively and then communicating findings to users via an automated bot adds layers of complexity. The PI has previously successfully demonstrated his ability to work on such large data collection and annotation challenges [75]. Integrating human judgment with automated systems poses significant challenges, though the concept of real-time crowdsourcing, even with live crowds, has been proven viable on a large scale in past research [76]. Given that we plan to employ only a small number of experts specifically recruited for this task, we believe the integration will be feasible.

3.4.2 Modeling The Temporal Influence Of Misinformation

Another important component of our approach is the ability to collect temporal data, particularly from multiple countries to identify whether misinformation which started in one country/culture impacts other countries/cultures. To achieve this, will model the spread of narratives as temporal networks, which provide a representation of the pathways through which misinformation reaches different user groups.

To identify causality within these networks, we will employ Hawkes processes [77], a statistical method well-suited for modeling the self-exciting nature of information spread. Hawkes processes account for the temporal dependencies between events, enabling us to infer the causal relationships between exposure to misinformation and subsequent actions. This method allows us to answer questions such as: (i) the extent to which exposure to specific narratives increases the likelihood of certain beliefs or behaviors, (ii) the impact of a narrative posted first on certain platforms (say an encrypted platform like WhatsApp) on the viral spread of these narratives other open platforms such as Facebook or YouTube, or, (iii) whether misinformation which started in a certain

country/culture impacted beliefs in other countries/cultures.

We model the process of posting narratives on different platform with each platform corresponding to a process, and an event occurs each time a narrative is posted on a platform. Events on one process can cause impulses that can increase the likelihood of subsequent events, including other processes, e.g., a user might see a narrative on one platform (say YouTube) and post about it on their family WhatsApp group, or post it on Facebook. This approach allow us to assess the causality of events, hence it is a far better approach when compared to simple approaches like looking at the timeline of specific pieces of content [78].

Hawkes processes will also be used to model the influence of misinformation over time, identifying how initial exposure can lead to a cascade of effects within the network. By analyzing these cascades, we can determine the most influential sources of misinformation (e.g. the country/culture of origin) and the pathways through which it spreads. This analysis provides a deeper understanding of the mechanisms and sources driving the spread of misinformation and its impact on user behavior.

3.5 Solutions

3.5.1 Motivation: Need For Bottom-Up, Local Solutions

We hypothesize from findings in Section 3.3 that while some misinformation remains localized, affecting small community-level environments without gaining national traction, other misinformation becomes viral, spreading across communities. This distinction highlights a significant challenge: the current centralized model of fact-checking, which primarily addresses widespread, viral misinformation, often overlooks the subtleties and specificities of local misinformation.

Localized misinformation, which can profoundly influence smaller communities (particularly minorities), might be debunked using straightforward tools like reverse image searches or targeted keyword searches on Google [79, 80]. However, the prevailing centralized approach to combating misinformation – relying heavily on professional fact-checkers who process large volumes of data – struggles with efficiency, particularly when addressing content that involves local nuances or requires a culturally contextual understanding [81, 82].

To address these inefficiencies, we propose a shift towards innovative, bottom-up solutions that empower local communities to tackle misinformation directly. This approach would foster a more localized and responsive strategy, integrating the understanding of local cultural contexts that central fact-checkers may lack.

To operationalize this strategy, we envision developing a conversational AI agent capable of performing localized fact-checking tasks. This AI would be fully deployed on users' devices, handling local content independently while established fact-checkers address global misinformation trends. By operating within the cultural and contextual framework of their designated communities, these agents would prioritize user privacy and leverage local networks, ensuring that interventions are both culturally appropriate and contextually relevant. This platform-agnostic solution would be adaptable to various digital environments, significantly enhancing the capability to combat misinformation at a granular level.

This localized, on-device approach represents a critical shift, treating local and global misinformation with the distinct strategies they require, and promises to bridge the gap between global capabilities and local needs effectively.

3.5.2 Functionality

Our proposed solution capitalizes on the latest advancements in large language models (LLMs) that are adept at handling multi-modal content [28]. The AI assistant (deployed as a standalone mobile app) has access to all data on the user's device, including screenshots, messages, images, videos, and text. It can index and understand the content the user interacts with, providing a comprehensive analysis of their information consumption. Users can activate the AI assistant for specific platforms they wish to (e.g. only on WhatsApp but not YouTube). The assistant works in the background, completely on the device, monitoring content consumption and alerting the user if they encounter misinformation. No user data is exported to a central server.

The misinformation detection process within the app incorporates multiple stages. It begins with an initial screening, where the app cross-references incoming user queries against a dynamic database of hashes from previously fact-checked content [83]. If there is no exact match, a local classifier, powered by an LLM [84], assesses the content for potential misinformation. If the classifier identifies misinformation with high confidence, the user is immediately alerted. If the issue remains unresolved, the app further engages the user by asking for their perspective and the possibility of involving their network for peer review, which educates users and taps into collective intelligence for verification. Additionally, the system employs federated learning, where local experts contribute to, and enhance the LLM's accuracy over time. In cases where misinformation is complex or unresolved, these are escalated to a dedicated fact-checker tip line for professional review.

3.5.3 Technical Implementation

The principal component of this system is a mobile application designed to intelligently monitor and analyze user interactions privately on the user's device. We will develop an app that capitalizes on proven methodologies for capturing user interactions through screenshots of selected apps, inspired by the success of similar technologies in capturing mobile behavior as demonstrated by Reeves et al. [85]. This method echoes the functionality seen in features like Pixel Screenshots on the Pixel 9 phone [86] and Microsoft's Recall on Copilot [87]. Additionally, leveraging Android's accessibility features will allow us to capture app-specific content with greater precision [88]. The PI, Garimella, brings significant expertise in developing such sophisticated mobile applications [89].

For content analysis, our solution incorporates state-of-the-art NLP and computer vision models, fine-tuned on large-scale, annotated narrative datasets as referenced in Section 3.2.1. This will allow the app to discern and analyze misinformation across different media types, including text, images, and videos.

Central to our approach is the use of federated learning, which enables the AI models to learn from data across multiple devices without necessitating centralized data storage. This methodology not only upholds privacy but also benefits from real-world data diversity. We plan to leverage Federated Transfer Learning, where a generalized pre-trained model on misinformation detection is adapted locally on each device using specific user data, which enhances its relevance and accuracy in diverse linguistic and cultural settings.

Given the non-IID nature of data in such applications—where data points on individual devices are not independent and identically distributed—we face technical challenges in model training and performance. To address these, we will employ techniques like FedProx and FedDyn, which are designed to improve convergence in federated settings by managing the heterogeneity of client updates [90, 91]. FedProx introduces a proximal term to the local objective function, which mitigates the issue of client drift by keeping local updates closer to the global model. FedDyn, on the other hand, dynamically adjusts local objective functions based on the divergence of local models from

the global model, effectively countering the negative effects of data heterogeneity. These methods ensure more stable and effective learning across the diverse devices involved in our project.

To augment the system's efficacy, we will implement a framework for AI assistants to communicate and share information within a user's personal network, employing influence maximization strategies to prioritize interactions with influential nodes [92, 93]. This not only enhances accuracy but also extends the reach of misinformation detection. Moreover, a gamified verification system will incentivize user participation through points and rewards, fostering a community-based approach to misinformation identification and correction.

The outputs from both automated and collaborative verification processes will be aggregated and relayed back to the user, providing a transparent and comprehensive view of content veracity. This ensures that users are not only recipients of information but active participants in the verification process.

4 Broader Impacts

The proposed research has significant potential to benefit society and contribute to the achievement of specific, desired societal outcomes. The broader impacts of this work include:

Focus on Minority Communities in the U.S.. This research will be one of the first to understand the consumption of misinformation by two minority communities – Indian and Latinx Americans – who are often underrepresented in misinformation studies. By understanding the consumption patterns and misinformation dynamics within these groups, this project will fill a critical gap in existing research and provide insights that can inform more inclusive and effective public health and communication strategies. Moreover, the research will also help in understanding the impact of domestic actors from their respective countries (India, Mexico, Colombia) on Americans and their beliefs.

Practical solutions for misinformation. The solutions (in Section 3.5) proposed are directly applicable and designed to be scalable with an easy to install app. This enables the easy dissemination of the app through journalists, fact checkers and other community partners enabling a possible mass adoption.

Interdisciplinary contributions. This project represents the possibilities of computational social science implemented on a large scale, integrating cutting-edge tools from computer science with dynamic applications to real-world content to tackle pressing social science questions. By employing advanced machine learning models, we analyze multi-modal and multi-platform user-consumed content in real-time. Experts participate actively in the loop, annotating content to identify misinformation. We then complete this cycle by surveying users about their beliefs, directly addressing crucial social science questions about the impact of misinformation. This innovative interdisciplinary methodology provides insights, tools and datasets for both computer scientists and social scientists and the paradigms developed will help lay the foundation for the much needed interdisciplinary research on misinformation in the future.

Educational contributions. This project will also have a significant educational impact, training the next generation of scholars and community members in cutting-edge techniques for studying and combating misinformation. See Section 7 for more details.

5 Results from Prior NSF Support

Garimella is a PI on a 'SaTC: CORE: Small: Towards a Privacy-Preserving Framework for Research on Private, Encrypted Social Networks' Award number 2318843. Period: 09/2023 to 09/2026.

Intellectual Merit: The project develops privacy preserving techniques to collect data and study the prevalence of misinformation on encrypted social networks like WhatsApp. It also develops solutions that make use of the peer-to-peer nature of WhatsApp, built upon the wisdom of crowds to tackle misinformation. **Broader Impacts:** The project provides a first look at misinformation on a platform like WhatsApp and how it differs from other platforms like Facebook or Twitter. It provides tools to enable journalists and civil society organizations to open the black box of WhatsApp. **Publications and products:** The project enabled research and tools to collect data from WhatsApp. WhatsApp Explorer, a tool to collect data at scale from WhatsApp, was partly funded by this project [58].

Garimella is a recipient of an NSF Convergence accelerator grant (Phases 1 and 2). The project was titled "FACT CHAMP - Fact-checker, Activist, and Academic Collaboration Tools: Combating Hate & Abuse towards Minorities as well as Misinformation & Propaganda In Online Networks" in Phase 1 and "Co-Insights: Fostering Community Collaboration to Combat Misinformation" (in Phase 2). Award number: 49100421C0035, Period: 09/2021 to 09/2024. **Intellectual Merit:** The Co-Insights project provides a unique platform that enables community, fact-checking, and academic organizations to work together to respond effectively to misinformation targeting Asian-American and Pacific Islander (AAPI) communities. **Broader Impacts:** The grant showed that misinformation claims often form around common themes, persistent stereotypes, and patterns of deception. By building taxonomies and using machine learning to map claims to them, we can move to a proactive model where interventions are available before claims spread widely. **Publications and products:** Publications: [59]. Tools for data collection were developed as part of this project: WhatsViral [89] and Diaspora Watch [57].

6 Relevance to SaTC

This proposal is being submitted to SaTC. The project is directly aligned with SaTC's focus on supporting a safe, secure, resilient, and trustworthy cyberspace. In particular, our focus on identifying and developing solutions for misinformation on social networks directly addresses several topics of interest in the Information Integrity topic area of SaTC. The data collection includes various minority communities in the US and identifies the influence they might have from their own countries, providing novel approaches to understanding the scale and impact of misinformation on Americans. Finally, we develop realistic and bottom up interventions based on the use of state of the art AI tools, which can be scaled independently without depending on the platforms. While we aim at developing a generic framework for identifying and countering the misinformation on three of the most popular social networks, the outcomes of this research can be applied to misinformation in other domains and platforms as well. This project is also aligned with the SaTC program's vision of taking an interdisciplinary, comprehensive, and holistic approach to cybersecurity issues.

7 Education plan

The primary objective of the education component is to raise awareness about misinformation among minority communities in the US, particularly those who use non-traditional platforms like WhatsApp, mostly in non English, low resource languages. These communities are often understudied and excluded from mainstream fact-checking ecosystems, making them vulnerable to misinformation. By leveraging our research findings, we aim to educate these communities about the nature of misinformation they encounter and empower them with tools to identify and mitigate its effects.

Our plan has two distinct, yet interacting components: (i) Community outreach with Indian and Latinx population of adults in New Jersey, (ii) Opportunities for Rutgers undergraduates and masters students. In order to streamline the design of relevant educational materials, our timeline integrates both adult community member and Rutgers student populations in several waves of curriculum development, guided in collaboration with research mentor, Prof. Rebecca Reynolds at the School of Communication and Information.

7.1 Community Outreach

We will conduct community workshops and seminars in collaboration with community centers, libraries, and local organizations, while developing a toolkit of resources for self-education including pamphlets, videos, and online resources. Particularly, we will partner with community organizations based in New Jersey like Indo-American Cultural Foundation of Central New Jersey [94], Latino Leadership Alliance of New Jersey (LLANJ) [95], and libraries based in Edison, Fort Lee, and Palisades Park in Central and North New Jersey with large populations of Indian and Latinx American populations [63].

7.2 Education Outreach For Rutgers Students

Students from the Rutgers School of Communication and Information's Information Technology and Informatics (IT&I) undergraduate major and the Master of Information (MI) programs will actively participate in the design and implementation of the community education materials, providing them with practical experience in educational design and community service. This cohort of students is particularly well placed for a bootcamp we propose because of the diverse background of the student body in IT&I and MI programs.

7.3 Curriculum Development

We will engage in design of (a) informal community learning experiences where learning objectives will be focused on critical literacies for lay publics of the target participants (workshops, presentations, self-learning modules to teach about the problems of misinformation using culturally responsive methods in situ), as well as, (b) formal learning experiences for Rutgers students wherein learning objectives will be focused on a case-based approach to teaching both (i) critical literacies and (ii) critical programming/computational thinking/data literacies (undergraduate IT&I curriculum modules built into existing classes such as Information Security, or Social Informatics; master's MI curriculum modules built in existing classes such as Understanding, Designing, and Building Social Media; and a mis- and dis-information data and systems engineering boot camp). All teaching materials, worked examples of misinformation, and design and programming modules will stem from the primary grant research and technology development processes, results and outputs; they will also be pedagogically evidence-driven in learning sciences design based research methods.

7.3.1 Critical Literacies Curriculum For Lay Publics

The workshops and materials for community outreach will build upon the main project research findings to illustrate the types of misinformation targeting these communities, framing the socio-technical contexts in which they appear, providing case-based worked examples of existing misinformation operations occurring in the socio-technical systems contexts most frequently engaged by the target lay publics, and critical thinking activities to spark understanding, critique, and change

leading to new behaviours, practices and routines with social media. Interactive sessions will enable participants to learn and practice misinformation identification and evaluation skills using pedagogical methods such as the Apt-AIR framework for epistemic cognition which posits that epistemic thinking includes epistemic Aims and value, epistemic Ideals and Reliable epistemic processes, foregrounding learners' epistemic competence to successfully achieve epistemic aims across diverse situations and contexts [96]. The Apt-AIR framework further identifies five interweaving aspects of competent engagement with epistemic aims, ideals, and processes: (1) cognitive engagement in epistemic performance; (2) adapting epistemic performance to diverse situations and contexts; (3) metacognitively regulating and understanding epistemic performance; (4) caring about and enjoying epistemic performance; and (5) participating in epistemic performance together with others [96, 97, 98]. This framework will be integrated with the misinformation "ecological literacy" approach outlined in [99] which posits a multi-pronged educational curriculum to combating mis- and dis-information by building learners' complex socio-technical imaginations and critical literacies.

7.3.2 Critical Programming/Data Literacies Curriculum For Rutgers Students

In the evolving landscape of misinformation, the ability to navigate and analyze digital content critically is indispensable; so is preparation of an informed, critically minded STEM workforce. Our initiative, aims to equip Rutgers undergraduate and master's degree students in the STEM discipline of information science with the necessary tools to both understand and combat misinformation effectively through a synthesized educational approach that integrates critical mis/disinformation perspectives with foundational programming skills that are already built into their curricular and disciplinary course of study in IT&I and the MI. Students will learn to apply computational thinking processes such as decomposition, pattern recognition, abstraction, and algorithm design to analyze and understand the structure and propagation mechanisms of misinformation on various social media platforms. Emphasizing the critical role of data in understanding and addressing misinformation, the curriculum will cover data collection from various social media, analysis and visualization techniques. Students will learn how to critically assess data sources, use statistical tools to interpret data, and apply data visualization techniques to visualize the spread of misinformation.

Finally, students will acquire practical programming skills, particularly in Python, that help them detect misinformation. This includes learning how to use libraries and APIs for natural language processing, machine learning, and data donation tools on various social media platforms and use of data analytics and visualization systems such as Pandas, Plotly and Matplotlib, equipping them with the capabilities to build applications that can filter, analyze, interpret and fact-check large volumes of information quickly and efficiently.

Timeline. In the first year, the education plan aims to engage with partner organizations for recruitment and conduct interviews with 15-20 individuals per target population to understand their misinformation practices and challenges, followed by data analysis and reporting based on the Apt-AIR framework. Over the subsequent years, the program will develop and refine a culturally tailored critical literacies curriculum through iterative feedback, pilot testing, and community workshops, culminating in the final implementation and assessment of workshop materials in the fourth year.

Budget. A budget of $30,000 has been allocated for recruiting specialists, community engagement and a summer bootcamp, covering staff, and resources, with plans to expand and secure additional funding based on the initiatives' success.

References Cited

[1] S. Torkington, "These are the 3 biggest emerging risks the world is facing," https://www.weforum.org/agenda/2024/01/ai-disinformation-global-risks/, 2024, [Accessed 13-06-2024].

[2] U. Ecker, J. Roozenbeek, S. van der Linden, L. Q. Tay, J. Cook, N. Oreskes, and S. Lewandowsky, "Misinformation poses a bigger threat to democracy than you might think," *Nature*, vol. 630, no. 8015, pp. 29–32, 2024.

[3] C. Arun, "On whatsapp, rumours, lynchings, and the indian government," *Economic & Political Weekly*, vol. 54, no. 6, 2019.

[4] E. F. Thomas, L. Bird, A. O'Donnell, D. Osborne, E. Buonaiuto, L. Yip, M. Lizzio-Wilson, M. Wenzel, and L. Skitka, "Do conspiracy beliefs fuel support for reactionary social movements? effects of misbeliefs on actions to oppose lockdown and to "stop the steal"," *British Journal of Social Psychology*, 2024.

[5] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[6] A. Guess, J. Nagler, and J. Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on facebook," *Science advances*, vol. 5, no. 1, p. eaau4586, 2019.

[7] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake news on twitter during the 2016 us presidential election," *Science*, vol. 363, no. 6425, pp. 374–378, 2019.

[8] S. Baribi-Bartov, B. Swire-Thompson, and N. Grinberg, "Supersharers of fake news on twitter," *Science*, vol. 384, no. 6699, pp. 979–982, 2024.

[9] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[10] A. Chadwick, C. Vaccari, and N.-A. Hall, "Covid vaccines and online personal messaging: the challenge of challenging everyday misinformation," 2022.

[11] S. Badrinathan and S. Chauchard, "Researching and countering misinformation in the global south," *Current Opinion in Psychology*, p. 101733, 2023.

[12] M. Stojanoski, "Fake news, child-kidnapping gangs and violence against the roma community in france: making social media accountable," *International Journal on Rule of Law, Transitional Justice and Human Rights*, vol. 11, no. 11, pp. 53–65, 2020.

[13] T. Rahman and I. Jahan, "The role of social media rumors in social unrest of bangladesh," *International Journal for Studies on Children, Women, Elderly and Disabled*, vol. 11, 2020.

[14] P. Godfrey-Smith, "Misinformation," *Canadian Journal of Philosophy*, vol. 19, no. 4, pp. 533–550, 1989.

[15] K. Starbird, J. Maddock, M. Orand, P. Achterman, and R. M. Mason, "Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing," *IConference 2014 proceedings*, 2014.

[16] K. Shu, S. Dumais, A. H. Awadallah, and H. Liu, "Detecting fake news with weak social supervision," *IEEE Intelligent Systems*, vol. 36, no. 4, pp. 96–103, 2020.

[17] T. Bucher and A. Helmond, "The affordances of social media platforms," *The SAGE handbook of social media*, vol. 1, pp. 233–254, 2018.

[18] Y. Velez, "Crowdsourced adaptive surveys," *arXiv preprint arXiv:2401.12986*, 2024.

[19] Á. Figueira and L. Oliveira, "The current state of fake news: challenges and opportunities," *Procedia computer science*, vol. 121, pp. 817–825, 2017.

[20] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, and D. Rothschild, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[21] K. Shu and H. Liu, *Detecting fake news on social media*. Springer Nature, 2022.

[22] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: a survey and new perspectives," *Social Network Analysis and Mining*, vol. 10, no. 1, p. 82, 2020.

[23] N. Micallef, M. Sandoval-Castañeda, A. Cohen, M. Ahamad, S. Kumar, and N. Memon, "Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 651–662.

[24] A. Wilson, S. Wilkes, Y. Teramoto, and S. Hale, "Multimodal analysis of disinformation and misinformation," *Royal Society Open Science*, vol. 10, no. 12, p. 230964, 2023.

[25] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 395–405.

[26] P. Juneja and T. Mitra, "Human and technological infrastructures of fact-checking," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–36, 2022.

[27] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.

[28] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[29] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[30] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, and G. L. Ciampaglia, "Anatomy of an online misinformation network," *PLOS ONE*, vol. 13, no. 4, pp. 1–23, 04 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0196087

[31] X. Chen, S.-C. J. Sin, Y.-L. Theng, and C. S. Lee, "Why do social media users share misinformation?" in *Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries*, 2015, pp. 111–114.

[32] J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. Freeman, G. Recchia, A. M. Van Der Bles, and S. Van Der Linden, "Susceptibility to misinformation about covid-19 around the world," *Royal Society open science*, vol. 7, no. 10, p. 201199, 2020.

[33] M. M. F. Caceres, J. P. Sosa, J. A. Lawrence, C. Sestacovschi, A. Tidd-Johnson, M. H. U. Rasool, V. K. Gadamidi, S. Ozair, K. Pandav, and C. Cuevas-Lou, "The impact of misinformation on the covid-19 pandemic," *AIMS Public Health*, vol. 9, no. 2, p. 262, 2022.

[34] E. Ferrara, S. Cresci, and L. Luceri, "Misinformation, manipulation, and abuse on social media in the era of covid-19," *Journal of Computational Social Science*, vol. 3, pp. 271–277, 2020.

[35] G. A. Fine and B. Ellis, *The global grapevine: Why rumors of terrorism, immigration, and trade matter*. Oxford University Press, 2013.

[36] P. Gahr, A. S. DeVries, G. Wallace, C. Miller, C. Kenyon, K. Sweet, K. Martin, K. White, E. Bagstad, and C. Hooker, "An outbreak of measles in an undervaccinated community," *Pediatrics*, vol. 134, no. 1, pp. e220–e228, 2014.

[37] S. L. Benoit and R. F. Mauldin, "The "anti-vax" movement: a quantitative report on vaccine beliefs and knowledge across social media," *BMC public health*, vol. 21, pp. 1–11, 2021.

[38] A. Pereira and J. Van Bavel, "Identity concerns drive belief in fake news," *https://psyarxiv.com/7vc5d*, 2018.

[39] D. A. Reinero, E. A. Harris, S. Rathje, A. Duke, and J. J. Van Bavel, "Partisans are more likely to entrench their beliefs in misinformation when political outgroup members fact-check claims," 2023.

[40] S. Loomba, A. De Figueiredo, S. J. Piatek, K. De Graaf, and H. J. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature human behaviour*, vol. 5, no. 3, pp. 337–348, 2021.

[41] N. Persily, J. A. Tucker, and J. A. Tucker, "Social media and democracy: The state of the field, prospects for reform," 2020.

[42] J. L. Kalla and D. E. Broockman, "The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments," *American Political Science Review*, vol. 112, no. 1, pp. 148–166, 2018.

[43] R. Bleiman, "Understanding the united states republicans' susceptibility to political misinformation," in *The International Conference on Cybersecurity, Situational Awareness and Social Media*. Springer, 2023, pp. 169–192.

[44] J. Roozenbeek, S. Van Der Linden, and T. Nygren, "Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures," 2020.

[45] S. Lewandowsky and S. Van Der Linden, "Countering misinformation and fake news through inoculation and prebunking," *European Review of Social Psychology*, vol. 32, no. 2, pp. 348–384, 2021.

[46] M. Hameleers, "Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the us and netherlands," *Information, communication & society*, vol. 25, no. 1, pp. 110–126, 2022.

[47] T. Dame Adjin-Tettey, "Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education," *Cogent arts & humanities*, vol. 9, no. 1, p. 2037229, 2022.

[48] G. Pennycook and D. G. Rand, "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation," *Nature communications*, vol. 13, no. 1, p. 2333, 2022.

[49] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer, "Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks," in *Proceedings of the 24th international conference on World Wide Web*, 2015, pp. 977–982.

[50] E. Porter and T. J. Wood, "The global effectiveness of fact-checking: Evidence from simultaneous experiments in argentina, nigeria, south africa, and the united kingdom," *Proceedings of the National Academy of Sciences*, vol. 118, no. 37, p. e2104235118, 2021.

[51] A. Kazemi, K. Garimella, G. K. Shahi, D. Gaffney, and S. A. Hale, "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*, vol. 3, no. 1, 2022.

[52] S. Zannettou, ""i won the election!": an empirical analysis of soft moderation interventions on twitter," in *Proceedings of the international AAAI conference on web and social media*, vol. 15, 2021, pp. 865–876.

[53] O. Papakyriakopoulos and E. Goodman, "The impact of twitter labels on misinformation spread and user engagement: Lessons from trump's election tweets," in *Proceedings of the ACM web conference 2022*, 2022, pp. 2541–2551.

[54] J. Roozenbeek, A. L. Freeman, and S. Van Der Linden, "How accurate are accuracy-nudge interventions? a preregistered direct replication of pennycook.(2020)," *Psychological science*, vol. 32, no. 7, pp. 1169–1178, 2021.

[55] C. M. L. Wong and Y. Wu, "Limits to inoculating against the risk of fake news: a replication study in singapore during covid-19," *Journal of Risk Research*, vol. 26, no. 10, pp. 1037–1052, 2023.

[56] A. Seelam, A. Paul Choudhury, C. Liu, M. Goay, K. Bali, and A. Vashistha, ""fact-checks are for the top 0.1%": Examining reach, awareness, and relevance of fact-checking in rural india," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW1, pp. 1–34, 2024.

[57] Garimella, Kiran, "Diaspora watch: Facebook data donation," https://www.diaspora-watch.us/, 2022.

[58] K. Garimella and S. Chauchard, "Whatsapp explorer: A data donation tool to facilitate research on whatsapp," *arXiv preprint arXiv:2404.01328*, 2024.

[59] J. Couto and K. Garimella, "Examining (political) content consumption on facebook through data donation," *Under Review*, 2024.

[60] "Right of access," https://gdpr-info.eu/issues/right-of-access/, accessed: 2024-06-10.

[61] M. Offer-Westort, L. R. Rosenzweig, and S. Athey, "Battling the coronavirus 'infodemic'among social media users in kenya and nigeria," *Nature Human Behaviour*, vol. 8, no. 5, pp. 823–834, 2024.

[62] A. Collis, K. Garimella, A. Moehring, M. A. Rahimian, S. Babalola, N. H. Gobat, D. Shattuck, J. Stolow, S. Aral, and D. Eckles, "Global survey on covid-19 beliefs, behaviours and norms," *Nature Human Behaviour*, vol. 6, no. 9, pp. 1310–1317, 2022.

[63] U. C. Bureau, "American community survey," 2022. [Online]. Available: https://www.census.gov/programs-surveys/acs

[64] A. Seitz, "Inside the 'big wave' of misinformation targeted at Latinos — apnews.com," https://apnews.com/article/latinos-misinformation-election-334d779a4ec41aa0eef9ea80636f9595, 2021, [Accessed 16-07-2024].

[65] S. Stantcheva, "How to run surveys: A guide to creating your own identifying variation and revealing the invisible," *Annual Review of Economics*, vol. 15, no. 1, pp. 205–234, 2023.

[66] K. Garimella, B. Nayak, S. Chauchard, and A. Vashistha, "Deciphering viral trends in whatsapp: A case study from a village in rural india," 2024. [Online]. Available: https://arxiv.org/abs/2407.08172

[67] S. Yoon, D. Lee, Y. Zhang, and J. Han, "Unsupervised story discovery from continuous news streams via scalable thematic embedding," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 802–811.

[68] B. F. Keith Norambuena, T. Mitra, and C. North, "A survey on event-based news narrative extraction," *ACM Computing Surveys*, vol. 55, no. 14s, pp. 1–39, 2023.

[69] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[70] C. A. Bail, B. Guay, E. Maloney, A. Combs, D. S. Hillygus, F. Merhout, D. Freelon, and A. Volfovsky, "Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017," *Proceedings of the national academy of sciences*, vol. 117, no. 1, pp. 243–250, 2020.

[71] K. Starbird, A. Arif, and T. D. Wilson, "Disinformation as collaborative work," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, pp. 1 – 26, 2019.

[72] C. Buntain, M. Innes, T. Mitts, and J. Shapiro, "Cross-platform reactions to the post-january 6 deplatforming," *Journal of Quantitative Description: Digital Media*, vol. 3, 2023.

[73] S. Badrinathan, "Educative interventions to combat misinformation: Evidence from a field experiment in india," *American Political Science Review*, vol. 115, no. 4, pp. 1325–1341, 2021.

[74] A. A. Arechar, J. Allen, A. J. Berinsky, R. Cole, Z. Epstein, K. Garimella, A. Gully, J. G. Lu, R. M. Ross, and M. N. Stagnaro, "Understanding and combatting misinformation across 16 countries on six continents," *Nature Human Behaviour*, vol. 7, no. 9, pp. 1502–1513, 2023.

[75] S. Chauchard and K. Garimella, "What circulates on partisan whatsapp in india? insights from an unusual dataset," *Journal of Quantitative Description: Digital Media*, vol. 2, 2022.

[76] W. S. Lasecki, C. D. Miller, and J. P. Bigham, "Warping time for more effective real-time crowdsourcing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2033–2036.

[77] S. Linderman and R. Adams, "Discovering latent network structure in point process data," in *International conference on machine learning*. PMLR, 2014, pp. 1413–1421.

[78] Y. Golovchenko, C. Buntain, G. Eady, M. A. Brown, and J. A. Tucker, "Cross-platform state propaganda: Russian trolls on twitter and youtube during the 2016 us presidential election," *The International Journal of Press/Politics*, vol. 25, no. 3, pp. 357–389, 2020.

[79] B. Paris and J. Donovan, "Deepfakes and cheap fakes," 2019.

[80] M. Pérez-Escolar, E. Ordóñez-Olmedo, and P. Alcaide-Pulido, "Fact-checking skills and project-based learning about infodemic and disinformation," *Thinking Skills and Creativity*, vol. 41, p. 100887, 2021.

[81] S. Shaar, N. Babulkov, G. Da San Martino, and P. Nakov, "That is a known lie: Detecting previously fact-checked claims," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3607–3618.

[82] A. Kazemi, K. Garimella, D. Gaffney, and S. Hale, "Claim matching beyond english to scale global fact-checking," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4504–4517.

[83] J. C. Reis, P. Melo, K. Garimella, and F. Benevenuto, "Can whatsapp benefit from debunked fact-checked stories to reduce misinformation?" *Harvard Kennedy School Misinformation Review*, vol. 1, no. 5, 2020.

[84] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros, "Dataset distillation," *arXiv preprint arXiv:1811.10959*, 2018.

[85] B. Reeves, N. Ram, T. N. Robinson, J. J. Cummings, C. L. Giles, J. Pan, A. Chiatti, M. Cho, K. Roehrick, and X. Yang, "Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them," *Human–Computer Interaction*, vol. 36, no. 2, pp. 150–201, 2021.

[86] P. Beaton, "Google's pixel 9 might use ai to help you search for screenshots," https://archive.is/p4cLx, 2024, [Accessed 09-07-2024].

[87] M. Wojo, "Co pilot Recall Overview — learn.microsoft.com," https://archive.is/GU2dm, 2024, [Accessed 09-07-2024].

[88] A. Alshayban, I. Ahmed, and S. Malek, "Accessibility issues in android apps: state of affairs, sentiments, and ways forward," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, 2020, pp. 1323–1334.

[89] Garimella, Kiran, "Whatsviral: Find out what is viral on whatsapp," https://www.whats-viral.me/, 2022.

[90] A. Khaled, K. Mishchenko, and P. Richtárik, "First analysis of local gd on heterogeneous data," *arXiv preprint arXiv:1909.04715*, 2019.

[91] D. A. E. Acar, Y. Zhao, R. M. Navarro, M. Mattina, P. N. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," *arXiv preprint arXiv:2111.04263*, 2021.

[92] L. Qiu, X. Tian, S. Sai, and C. Gu, "Lgim: A global selection algorithm based on local influence for influence maximization in social networks," *IEEE Access*, vol. 8, pp. 4318–4328, 2019.

[93] K. Garimella, A. Gionis, N. Parotsidis, and N. Tatti, "Balancing information exposure in social networks," *Advances in neural information processing systems*, vol. 30, 2017.

[94] IACFNJ, "IACFNJ — iacfnj.org," https://iacfnj.org, 2023, [Accessed 09-07-2024].

[95] L. L. Alliance, "Latino Leadership Alliance — latinoleadershipalliance.org," https://latinoleadershipalliance.org/, 2024, [Accessed 09-07-2024].

[96] S. Barzilai and C. A. Chinn, "On the goals of epistemic education: Promoting apt epistemic performance," *Journal of the Learning Sciences*, 2017.

[97] S. Barzlai and C. A. Chinn, "A review of educational responses to the post-truth condition: Four lenses on post-truth problems," *Educational psychologist*, vol. 55, no. 3, pp. 107–119, 2020.

[98] C. A. Chinn, S. Barzilai, and R. G. Duncan, "Education for a "post-truth" world: New directions for research and practice," *Educational Researcher*, vol. 50, no. 1, pp. 51–60, 2021.

[99] B. Paris, G. Marcello, and R. Reynolds, "Cultivating ecological literacy: A critical framework for understanding and addressing mis-and disinformation," *Proceedings of the Association for Information Science and Technology*, vol. 59, no. 1, pp. 479–485, 2022.

[100] P. Melo, J. Messias, G. Resende, K. Garimella, J. Almeida, and F. Benevenuto, "Whatsapp monitor: A fact-checking system for whatsapp," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, 2019, pp. 676–677.

[101] J. Angwin, "Few are addressing one of social media's greatest perils," https://www.nytimes.com/2023/05/06/opinion/fear-speech-social-media.html, 2023, [Accessed 17-07-2024].

[102] G. Shih, "Inside the vast digital campaign by hindu nationalists to inflame india," https://www.washingtonpost.com/world/2023/09/26/hindu-nationalist-social-media-hate-campaign, 2023, [Accessed 17-07-2024].

Synergistic Activities

Throughout my career, I have consistently prioritized the integration and dissemination of knowledge alongside its creation, as demonstrated by my commitment to making research outputs accessible and useful to both academic and public sectors.

1. **Public Release of Research Tools and Code**: My projects often incorporate a strong public component, exemplified by the release of research tools and codes. For instance, under the NSF project 49100421C0035, my team developed a data donation tool for Facebook, which was documented in our publication [59]. We also launched a public dashboard available at kiran-research2.comminfo.rutgers.edu/plotly to facilitate broader access to our data. This platform enables users to replicate our research findings and explore additional insights, significantly enhancing the transparency and reproducibility of our work.

2. **Broader use of Data Donation Tools**: Another example of my research being of broader use is the open sourcing of our data donation methodologies for WhatsApp (funded by NSF projects 2318843 and 49100421C0035) [89, 58]. These tools have been adopted by researchers at institutions such as the University of Chicago and the University of Pennsylvania, aiding them in establishing their own data collection infrastructures and fostering a collaborative environment for scientific advancement.

3. **Collaborations Beyond Academia**: My work extends beyond traditional academic boundaries through active collaborations with fact-checkers, journalists, and NGOs. For example, a dashboard we created for monitoring WhatsApp data has been utilized by over 100 journalists and fact-checking organizations across India and Brazil [100]. This tool has not only facilitated more informed journalism but also contributed to more robust public discourse.

4. **Engagement with Media and Dissemination of Knowledge**: My research has regularly been featured in both national and international media outlets, such as The New York Times and The Washington Post [101, 102]. Through these engagements, I strive to promote the practical implications of our research findings to a global audience, thereby broadening the public understanding of technological impacts on society.

5. **Technology Deployment with Non-Profits and Companies**: In partnership with the non-profit tech company Meedan, we deployed a technology that matches claims sent via a tipline to previously verified fact-checks [82, 51]. Funded by NSF project 49100421C0035, this technology is now operational across 12 tiplines in six languages in India, supporting thousands of queries daily and significantly enhancing the efficiency and reach of factual verification efforts.

These activities not only illustrate my commitment to the broader impacts of my work but also highlight my efforts in fostering an ecosystem where academic research and practical application converge to serve a global community.