Adopting an Ecological Approach to Misinformation: Understanding the Broader Scope and Impacts of Misinformation on Online Communities

Adopting an Ecological Approach to Misinformation: Understanding the Broader Scope and Impacts of Misinformation on Online Communities

by

Zhila Aghajari

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

Computer Science

Lehigh University

May 2025

Dissertation Signature Sheet

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Date

_____
Dissertation Advisor

Committee Members:

_____
Eric P. S. Baumer, Committee Chair

_____
Dominic DiFranzo

_____
Brian Davison

_____
Su Lin Blodgett

Acknowledgments

The journey of completing my PhD and this dissertation would not have been possible without the incredible support and mentorship of many individuals. First and foremost, I am grateful to my advisor, Eric Baumer. Eric, thank you for your belief in me and the trust you placed in my ability to pursue research directions that I am passionate about. Your support, mentorship, and guidance have been invaluable throughout my entire PhD. Your unique approach to advising has fostered significant personal and intellectual growth in me, and the lessons I learned from you will undoubtedly extend far beyond my doctoral program. I am also deeply thankful to Dominic DiFranzo for his continued mentorship and the insightful advice that helped my research to shape and to grow. Being part of your lab was an amazing experience, and I truly appreciate all the collaborations we explored together. Thank you to Brian Davison for the insightful comments, and thoughtful questions that significantly helped clarify key concepts in my dissertation. Brian's perspectives were particularly helpful in navigating some of the more complex aspects of my work, and I am grateful for that. I thank Su Lin Blodgett, for being such a positive mentor. Su Lin, your contributions to my work and your mentorship throughout my PhD were so important to me. I learned so much from our conversations. I am also grateful to the my peers in the HCSC lab at Lehigh University for their support and valuable feedback on my work. I am specially thankful of my lab mates, Amin, Chase, Lillian, and Asiyah, for their contributions to my research. To my parents, Shirin and Feridoun, and my siblings, your unwavering support and encouragement have been the very foundation of my journey, empowering me and instilling a deep sense of pride throughout my journey. Finally, and above all, to Amin, my love, my friend, and my mentor, I am profoundly grateful for you. You have been my rock

every single day of this PhD journey. Your encouragement at every step, your celebration of every achievement, and your comfort through every difficulty, all delivered with your beautiful smile and positive spirit, meant the world to me. Sharing this entire journey with you has been incredible, and I am so proud of everything we have accomplished together.

Table of Contents

List of Tables

List of Figures

Abstract

Misinformation plays a significant role in people's lives. While numerous interventions are designed to address misinformation and its impacts, these interventions primarily focus on addressing individual pieces of false and misleading content. This dissertation argues that such individualistic focus on misinformation de-emphasizes and draws attention away from the broader scope and impacts of misinformation. Instead, it advocates for conceiving of misinformation as a broad societal phenomenon that transcends any isolated, individual pieces of false or misleading content. This perspective to misinformation, in particular, emphasizes the crucial role of community-oriented mechanisms in the broader scope and impacts of misinformation, which are often under-looked in individualistic approaches to misinformation. Therefore, to study the broader scope of misinformation, and to account for the community-oriented mechanisms involved in this phenomenon, this dissertation takes an ecological approach to misinformation. Specifically, in an experimental setting, it demonstrates how false and misleading content, and community responses to such content, together contribute to the way misinformation influences perceptions about social norms within online communities, underscoring the way misinformation impacts online communities beyond misleading their members about any individual pieces of content. To examine the interplay between false and misleading content, and community responses in authentic online interactions in an observational setting, it then embraces the concept of framing from sociological research, and demonstrates how these elements and the interplay between them together contribute to the shifts in the way people come to understand the world's events, again impacting online communities beyond misleading their individual members about any individual pieces of content. The approach taken in this dissertation serves as an example that

can inform how future research might adopt an ecological approach to misinformation, and further expand knowledge about the broader scope and impacts of this phenomenon. This dissertation concludes by outlining the implications that it offers for researchers to study misinformation's broader scope, for community moderators to expand their moderation practices beyond individual content moderation, and for social media designers to leverage the role of community members in addressing the broader impacts of misinformation on how their communities run and evolve.

Chapter 1

Introduction

Misinformation has significant consequences, including political, economic, and social ones. Examples include the effects of misinformation on the US 2016 and 2020 presidential elections (Swire et al., 2017a; Allcott and Gentzkow, 2017; Bovet and Makse, 2019; Pennycook and Rand, 2021; Chen et al., 2021), its impacts on India's economy during the 2019 coronavirus disease outbreaks (fak, 2020), and its impacts on attitudes toward COVID-19 vaccines (Puri et al., 2020; Tasnim et al., 2020; Loomba et al., 2021; Enders et al., 2020). In several cases, online misinformation even inspired violent attacks in physical interactions (e.g., Rao, 2021; bbc, 2018).

Numerous interventions have been designed to address misinformation and its impacts (e.g., fact-checkers, signaling credibility of the content, signaling credibility of the source of content, or providing more perspectives on the content) (e.g., Facebook, 2020; AssociatedPress, 2021; **?**; Bhuiyan et al., 2021a). Despite the progress made in addressing misinformation, it is clearly still an ongoing and critical challenge (Lazer et al., 2018, 2017).

Part of the challenge involves conceptualizing what constitutes misinformation. Some researchers define misinformation as "false or misleading information" (Lazer et al., 2018; Wardle et al., 2018). Some others describe misinformation as "a claim that contradicts or distorts common understanding of verifiable facts" (Guess and Lyons, 2020). Both these definitions implicitly treat misinformation as individual pieces of content. While addressing pieces of false and misleading content is important, the full scope of the problem exceeds the bounds of such definitions. For example, the prevalence of false and misleading content

could create an information space in which individuals become skeptical of all claims, even the ones that are true, leading to a breakdown in trust in information space, in science, and in authorities (Lewandowsky et al., 2017, 2012).

Alternatively, we could conceive of misinformation as a broad phenomenon that transcends any individual piece of false or misleading content. In this view, misinformation is not an isolated phenomenon. Rather, it exists within a broader information ecosystem. That ecosystem involves a mixture of factually false statements, misleading content (regardless of their posters' intentions), true statements, as well as people's subjective discussions and opinions about various events. All of these elements and their interplay together contribute to people's views of the world's events, and impact their reactions to those events and the corresponding misinformation. For example, during the COVID-19 pandemic, people encountered a mixture of truthful, misleading, and uncertain content, in addition to personal opinions and experiences of other individuals. These elements are all part of misinformation about the COVID-19 pandemic and impacted how people viewed the pandemic and responded to its various aspects (e.g., masking, social distancing, vaccination, stress, and anxieties involved with the situation).

The impacts of misinformation could similarly extend beyond misleading individuals about a single piece of content. For example, misinformation contributes to broad societal issues such as the spread of distrust in science, and vaccine hesitancy (Bicchieri et al., 2021; bos, 2022; orević et al., 2021). These broad impacts of misinformation on society are arguably not the results of any single piece of untrue content. Instead, these effects also pertain to social factors surrounding such content, such as how people discuss these pieces of content and construct understanding related to these issues. Thus, to address misinformation and its broad impacts, it may be useful to incorporate approaches that extend beyond simply identifying and correcting individual pieces of untrue content. Put concisely, it may be beneficial to adopt **an ecological perspective on the phenomenon of misinformation** that considers the various elements within the information ecosystem and their influence on individuals' interpretation of and response to events. Such a comprehensive perspective may enable us to offer different strategies for approaching the problem of misinformation and its various aspects (Aghajari, 2023).

The concept of framing is a conceptual framework that allows us to take an ecological approach to the phenomenon of misinformation. While this concept is studied in different fields — from political communication (Froehlich and Rüdiger, 2006; Scheufele, 2000; Iyengar, 1996), to psychology (Quattrone and Tversky, 1988), to sociology (Goffman, 1974), to behavioral economics (Kahneman and Tversky, 1984) — this dissertation embraces a definition of framing from sociological research (Gamson, 1989; Scheufele, 1999; Benford and Snow, 2000; Druckman, 2001). Specifically, it treats framing as a dynamic set of processes by which people interpret and make sense of events around them and construct their understanding of those events. Within this sociological paradigm, processes of framing involve many aspects of interpretation and meaning making, including how people define problems, diagnose causes, make moral judgments related to those problems, and suggest remedies (Entman, 1993; Gamson and Modigliani, 1989). Understanding these functions that framing performs gives insights into how people interpret and view an event and respond to its surrounding aspects. In this way, examining framing processes may enable us to account for the broader scope of the information ecosystem wherein individuals encounter false and misleading content. Therefore, it may allow us to expand our vision of misinformation beyond individual pieces of content.

To account for the broader scope of misinformation, this dissertation, therefore, adopts an ecological approach to misinformation and expands understanding of this phenomenon and its impacts beyond pieces of false and misleading content. To do so, it first conducts a systematic literature review of prior work on countermeasures to address misinformation (Chapter 2). This review (published in the proceeding of CSCW 2023 (Aghajari et al., 2023b)), examines how these prior approaches account for factors that are involved in the phenomenon of misinformation and contribute to its impacts. It argues that prior approaches primarily focus on the pieces of content when addressing misinformation. As a result, they fall short in accounting for the other factors that are involved in the phenomenon of misinformation, such as the community-oriented factors that contribute to the spread and impacts of misinformation. Overlooking the interplay between false and misleading information and other elements in the information ecosystem that are involved in the phenomenon of misinformation can result in disregarding the impacts of misinformation

at the community level. This chapter concludes by highlighting the importance of considering community-oriented aspects of misinformation to effectively address the broader scope of misinformation and its systematic impacts.

To account for these community-oriented aspects, Chapter 3 examines the significant role of perceived norms as a factor that contributes to the spread and impact of misinformation at the community level (Colliander, 2019; Andı and Akesson, 2020; Gimpel et al., 2021; Koo et al., 2021). Given that there is less known about the mechanisms by which norms are perceived norms in this context, this chapter uses an experimental approach to investigate such potential mechanisms. Key findings show that the prevalence of false and misleading content, as well as the response of community members to such content, influence perceived norms and broader expectations about the community, including what the community as a whole might accept, and whether and how behaviors regarding misinformation might escalate within the community. In addition, the findings suggest that the response of community members to false and misleading content can mitigate the effects of such content on the perceptions about a community and its norms. Thus, Chapter 3 provides insights into the significant role of community response to misinformation on the way misinformation has its impacts. In addition, it also discusses implications for design around community response to address the broad impacts of misinformation (published in the proceeding of CHI 2023 (Aghajari et al., 2023a)).

However, such an approach may fall short in fully embracing the ecological perspective advocated here. While experimental methods are effective at isolating individual causal relationships, they may not fully account for interplay of false and misleading and the way online communities discuss and respond to an event. However, as mentioned earlier, adopting an ecological approach to addressing misinformation requires accounting for such complexities, including the nuances in community response and the interplay between false and misleading content with community response within the broader information ecosystem. To achieve this goal, this dissertation leverages the concept of framing, as explained above.

In order to examine framing with a dynamic, processual orientation, it may be advantageous to employ computational methods. Such techniques have the potential to account for some of the complexities across a wide, diverse information ecosystem. However, prior

computational work on framing focuses primarily on labeling specific frames *per se* (e.g., Baumer et al., 2015; Card et al., 2016), rather than analyzing framing as a set of dynamic processes involved in meaning constructions. Therefore, in Chapter 4, this dissertation takes a computational approach, and develops three models to test out and identify framing language. It examines the utility of these proposed models using a human-subject study, and evaluates how the patterns identified in the proposed models might be in line with what researchers look for when analyzing the framing processes. The results demonstrates the novel computational model designed and developed in this dissertation effectively provides evidence of framing language, and facilitates examining framing processes at wide, diverse information ecosystem.

In developing and evaluating these models, this dissertation focuses on the COVID-19 pandemic as a testbed. Indeed, the COVID-19 pandemic is a current, major health crisis, wherein misinformation is a dominant concern (Loomba et al., 2021; Rocha et al., 2021; Rosenberg et al., 2020; Roozenbeek et al., 2020). Given the relationship between misinformation and framing articulated above, it is crucial to understand how people frame this pandemic. In addition, studying the COVID-19 pandemic and its surrounding events is valuable in terms of providing an understanding of how to communicate aspects of future such events that unfold in real-time, and how to address the phenomenon of misinformation related to similar emerging events.

Next, Chapter 5 utilizes the computational techniques developed in Chapter 4 to investigate the relationships between the processes of framing and the phenomenon of misinformation. Leveraging the developed computational techniques, this dissertation explores the interplay between the prevalence of false and misleading content and framing processes in online communities. Specifically, it examines whether and how the prevalence of false and misleading content might play a role in the wider changes in how a community perceives and interprets world events, including the way they frame different events, and the way they respond to framing presented by other sources around those events, either implicitly or explicitly. To do so, it focuses on the COVID-19 pandemic, and investigate the processes of framing in two communities where false and misleading content is less or more prominent, specifically, the r/science and r/conspiracy communities on the Reddit platform. It analyzes

the differences between how these communities frame this event in response to the framing of the same event in the news media articles. Through these analyses, this dissertation investigates how these communities collectively reinforce, revise, or in some cases completely reject the framing of official news governments while discussing an event and constructing understandings about it.

The findings of this chapter demonstrate that r/science more often reinforces, and in some case revises framings from news media. However, r/conspiracy, a polar opposite of r/science community in terms of prevalence of false and misleading content (Phadke et al., 2022), more often significantly revises or completely rejects news media framings of the pandemic. The conducted analysis in this study contributes to the broader understanding of misinformation, and its role in the way online communities interpret events, and how misinformation as a societal phenomenon contributes to the community-level effects in terms of shifts in the processes of meaning construction. In addition to providing empirical evidence to the broader scope of misinformation, the approach taken in this chapter provides a concrete example to study misinformation with an ecological approach, serving as a pathway for future research to attend to the broader scope and impacts of misinformation.

To reiterate, this dissertation argues for conceiving of misinformation as a broad phenomenon that transcends any individual piece of content. It suggests that adopting this stance and moving beyond an individualistic focus on misinformation allows us to understand the broader scope of the problem. In particular, it demonstrates how misinformation shapes the broader impressions about a community, thereby contributes to the significant community level impacts. In addition, it examines how false and misleading content could become entangled in the processes of meaning constructions (i.e., framing), and the ways individuals and communities respond to an event. In this way, this dissertation highlights how taking an ecological approach to misinformation can expand our understanding of this phenomenon, and its broad impacts, from shaping perceptions about online communities to influencing the processes by which communities interpret and understand the world's events and respond to them. Such a comprehensive understanding of misinformation may enable us to offer different strategies for approaching the problem of misinformation and its extensive ramifications.

This dissertation concludes by offering important implications for researchers, community moderators, and social media designers, and discuss. Specifically, it discuss how the approach taken in this dissertation towards studying misinformation as a societal approach and using an ecological perspective can enable these stakeholders to better understand the broader scope of misinformation and to incorporate this understanding into the design of interventions to address broader scope and impacts of misinformation (Chapter 6).

Chapter 2

Review of Interventions to Address Misinformation

2.1 Introduction and Motivations

Despite the development of various techniques to address misinformation (i.e., fact-checkers, signaling credibility of the content, signaling credibility of the source of content, and providing more perspectives on the content) (e.g., Facebook, 2020; AssociatedPress, 2021; FacebookNewsroom, 2017; Bhuiyan et al., 2021a), misinformation is still a persistent problem with significant impacts across various domains of society. To identify strategies for interventions that maybe under explored, this chapter examines how prior work that are designed to address misinformation define and scope this phenomenon. More precisely, it examines how prior work that aims to address misinformation and its impacts accounts for the underlying factors that are involved in the phenomenon of misinformation and contribute to its impacts.

To do so, this chapter conducts a two-stage review process, following standard practices in literature review methodology (Okoli, 2015). First, it reviews interventions designed to address misinformation. Second, after examining the citations within those papers, this review describes the underlying mechanisms that drive responses to misinformation. Next, it analyzes these two collection of papers to examine how the various underlying mechanisms

are operationalized and implemented in the reviewed interventions that are designed to address misinformation.

This chapter highlights how prior work to address misinformation embodies an individualistic focus on misinformation and highlights the consequences that arise from such an individualistic focus. In particular, it emphasizes the ways this individualistic perspective towards addressing misinformation draws attentions away from the systemic nature and impacts of this phenomenon. Therefore, it suggests that future work should expand its vision beyond this individualistic focus to misinformation and adopt an ecological standpoint in understanding the broader scope of this phenomenon and addressing it. Furthermore, it outlines how community-oriented factors can be leveraged to inform the design of interventions aimed at addressing the broad impacts of misinformation.

2.2 Methods

To identify under-explored strategies for misinformation interventions, this chapter reviews existing interventions, as well as the various drivers of misinformation those interventions leverage. To conduct this review, it employs the guidelines for a systematic literature review provided by Okoli (2015). This section provides a high level of the review process. For the detailed descriptions, please refer to the paper titled "Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus" Aghajari et al. (2023b).

Following Okoli (2015), this chapter first clarifies the purpose of the literature review. Specifically, it focuses on interventions for addressing misinformation, which it defines as *false or misleading information that has the potential to deceive people*, regardless of the intentions of the actor who spreads it. This definition draws on and synthesizes aspects from a variety of prior definitions. Lazer et al. (2018, p. 1094) define misinformation as "false or misleading information", and disinformation refers to "false information that is *purposefully* spread to deceive people" (emphasis added). Similarly, Wardle et al. (2018) refer to misinformation as false or misleading information that the person spreading it believes is true, and defines disinformation as false information that the person disseminating it knows

is false. Both Lazer et al. (2018) and Wardle et al. (2018) emphasize the deliberateness of disseminating false information to distinguish disinformation from misinformation.

In contrast, Starbird et al. (2019) argue that disinformation is a form of information operation, and they emphasize that disinformation aims to undermine the integrity of the information space, and overwhelm individuals to make sense of information. According to Starbird et al. (2019), therefore, disinformation is broader than a single piece of information, and is rather a collaborative work of

Rather than make a clear distinction between misinformation and disinformation, this dissertation's definition of misinformation combines elements from each of these prior definitions. Noting the difficulty in determining intent behind content, our definition makes no strong claim about the poster's intent to deceive (or lack thereof) (cf. Wardle et al., 2018). This approach enables us to include both interventions that do and that do not attempt to infer the poster's intent, making this review more inclusive. At the same time, the proposed definition in this dissertation draws on the insight from Starbird et al. (2019) that the phenomena surrounding misinformation and its effects often transcends any single piece of content. This stance enables this review to include both interventions that attempt to address larger, coordinated campaigns (e.g., Facebook, 2018; twi, 2018, 2019), and interventions that are geared toward individual actors or pieces of content(e.g., Jahanbakhsh et al., 2021; Bhuiyan et al., 2021a). While this approach reviews interventions designed to address the spread of misinformation with different intentions under the same category, it helps analyze the broader landscape of interventions around misinformation.

2.2.1 The Review Process

Following Okoli (2015), this study uses the above scope to formulate a set of steps to conduct this review. This subsection briefly overviews the review process.

Based on the goal of mapping existing interventions and identifying under-explored approaches, this review process started by collecting intervention papers (resulted in 67 papers in the intervention corpus). This corpus of interventions is used to identify various factors and mechanisms that act as drivers of misinformation (resulted in 84 papers in the drivers of misinformation corpus). These individual mechanisms and factors were induc-

tively analyzed and thematically categorized. These categories of drivers provided a means of analyzing the corpus of interventions to address misinformation.

After collecting relevant research for our review, following the guideline by Okoli (2015), the papers are read for the purpose of the planned analysis. Specifically, aiming to map the interventions to the drivers of misinformation, the analysis are began by investigating the papers on the drivers of misinformation. The goal was to collect all the drivers of misinformation that were identified in prior work and to recognize patterns that might exist within those factors. Therefore, a thematic, inductive analysis analysis is conducted (Braun and Clarke, 2006), looking for patterns, similarities, and trends in terms of the ways that these papers conceptualized the various mechanisms, entities, and factors that might drive misinformation.

Figure 2.1 summarizes this review process. It includes the number of papers identified or retained at each step, as well as the relationships among the different steps. The following sections describe each step in greater details.

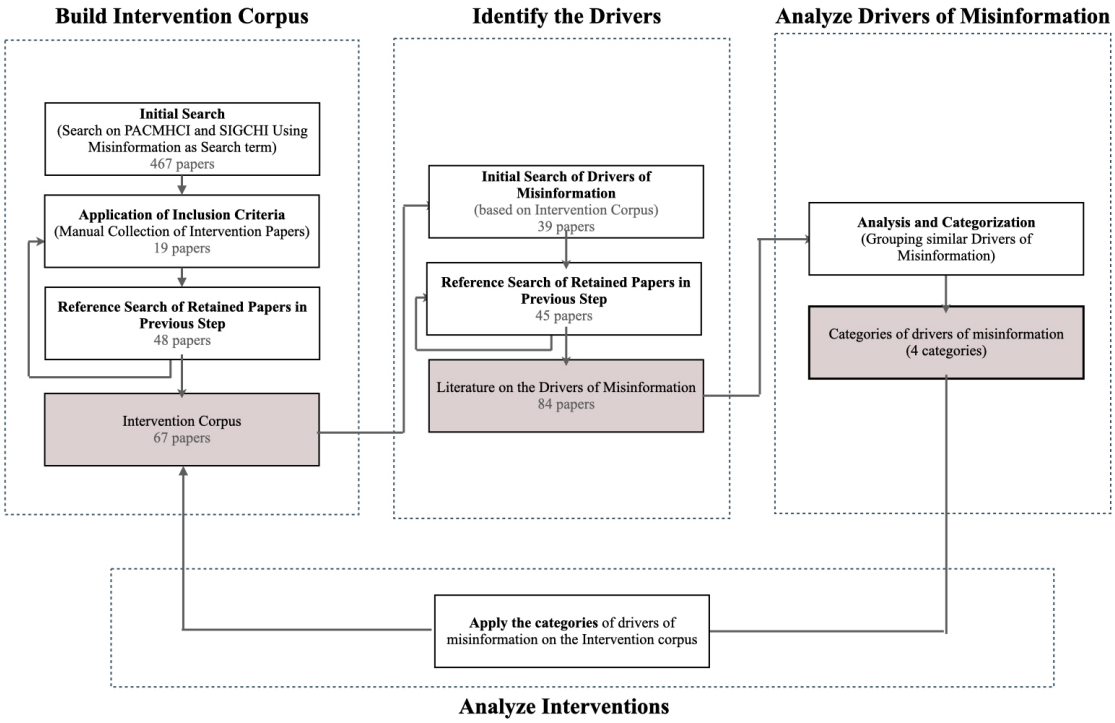


Figure 2.1: An overview of the review process. The number of papers examined and retained by search iteration is shown in the boxes.

2.2.1.1 Build the Intervention Corpus

This section describes the steps we took based on Okoli (2015)'s guideline to build the intervention corpus, and overviews the intervention corpus obtained from this step.

2.2.1.1.1 Initial Search. For our initial literature search, we used a search of the ACM Digital Library to collect papers. We searched papers published in PACMHCI and in conferences and journals where SIGCHI is a sponsor or co-sponsor. This library was chosen as our initial database because one of the key elements of our literature scope is investigating how the CSCW community and the broader SIGCHI community approach misinformation and design interventions to address this phenomenon. This initial literature search helps to ensure a comprehensive coverage of the relevant literature in SIGCHI community.

We used the search term "misinformation" to search for any papers that included the word "misinformation" anywhere in the paper. We used a single search term (i.e., misinformation) since we were specifically interested in the phenomena of misinformation (as apposed to "disinformation" and "fake news," both of which prior work distinguishes from "misinformation" (Wardle et al., 2018; Lazer et al., 2018; Starbird et al., 2019; mis, 2022)). Other prior reviews have similarly conducted their search processes around very constrained search terms (e.g., Kim et al., 2021; Brynjarsdottir et al., 2012; DiSalvo et al., 2010; Baumer et al., 2014; Boehner et al., 2007) as this approach helps to focus exactly on the phenomenon of interest and to exclude other related but dissimilar phenomena.

2.2.1.1.2 Application of Inclusion Criteria. In the resulting papers, we looked at the title and the abstract of each paper to identify the papers that met the following inclusion criteria:

- The paper was a peer-reviewed published work.
- The paper presented an intervention to address the spread or impacts of misinformation, or the paper investigated the efficacy of already existing interventions to combat misinformation.

- The intervention in the paper focused primarily on influencing the *response* to misinformation (e.g., when reading the news) rather than influencing the *creation* of misinformation (e.g., by a news reporter).

If the title and abstract were not clear enough about whether or not the paper focused on an intervention, we read the paper in more detail. To do so, we looked at the introduction and methods sections of the paper.

Papers that did not match these criteria were excluded. Specifically, while we included approaches that aimed at helping online users identify misinformation, we excluded the various methods that were aiming at improving classifiers to detect misinformation. In addition, extended abstract and working papers were excluded from our corpus, partly because it would have been resource-intensive to analyze that many papers, and partly because they did not always contain enough information for our analysis. Given the goal of this work (i.e., identifying under explored interventions), our review was not bound within any particular years to ensure that all the explored interventions in this community are covered in our review.

2.2.1.1.3 Reference search of papers identified in previous step. Following the guidelines from Okoli (2015), for each paper we went through the reference list to conduct a second pass of search. That is, we included in the review the intervention approaches that were cited by these resulting papers, but were not included in PACMHCI or SIGCHI proceedings. References that seemed relevant (i.e., the papers which include an intervention to address misinformation or investigated the efficacy of such interventions), were selected based on the title and the abstract of the papers.

Next, we repeated the process of going through the references and seeking the relevant papers until we reached a stage of saturation (i.e., where there were no more new publications being added to the corpus). Each interaction resulted in fewer relevant manuscripts and more cross-reference within the already found publications. Prior research noted such a systematic approach of paper seeking to be very effective at constructing a corpus of publications that are related to the same theme (Kitchenham and Charters, 2007).

2.2.1.1.4 Overview of Intervention Corpus. The first step of literature search on ACM digital library (i.e., Section 2.2.1.1) returned a total of 467 manuscripts that include the word "misinformation" anywhere in the paper. After excluding papers that did not meet our aforementioned criteria, this step identified 19 papers around interventions to address misinformation.

According to steps two and third in our approach, we repeated the process of going through the references and seeking the relevant papers until we reached a stage of saturation. This process resulted in 48 papers. Therefore, the whole process yielded 67 intervention papers being included in our review of intervention papers.

2.2.1.2 Identify the Drivers of Misinformation.

2.2.1.2.1 Initial search of drivers of misinformation. Next, we sought to identify the factors that drive misinformation. Most of the papers in our intervention corpus (generated in Section 2.2.1.1) include a subsection, usually within related work, where factors that drive misinformation are discussed. We used the citations that were introduced in those reviews to begin building our review of drivers of misinformation. We collected all the factors and mechanisms that were described as influencing the way people identify misinformation and respond to it. These drivers of misinformation were described based on prior work in journal such as PNAS (Proceedings of the National Academy of Sciences of the United States of America), Science, and Nature, as well as psychology journals, such as Journal of Experimental Psychology: General, Applied Cognitive Psychology, Psychological Bulletin. This step identified 39 papers around drivers of misinformation.

2.2.1.2.2 Reference search of papers identified in previous step. Similar to our method in building the intervention corpus, we again applied the literature review methodology offered by Okoli (2015). That is, for each paper identified at each step, we went through its reference list to conduct a second round of searching. This process resulted in 45 papers. Therefore, the whole process resulted in 84 papers on drivers of misinformation being included in our review.

2.2.1.3 Analyze Drivers of Misinformation.

After collecting relevant research for our review, following the guideline by Okoli (2015), we started reading the papers and extracting the information required for our analysis. Aiming to map the interventions to the drivers of misinformation, we began our analysis by investigating the papers on the drivers of misinformation (from Section 2.2.1.2). The goal was to collect all the drivers of misinformation that were identified in prior work and to recognize patterns that might exist within those factors. Therefore, we conducted a thematic, inductive analysis (Braun and Clarke, 2006), looking for patterns, similarities, and trends in terms of the ways that these papers conceptualized the various mechanisms, entities, and factors that might drive misinformation.

To do so, the authors read the papers and met periodically throughout the reading and analysis process to discuss observations, and emergent themes to reach consensus (Braun and Clarke, 2006). This step resulted in the development of a list of drivers of misinformation that prior work introduced as influential on people's response to misinformation. The authors then discussed this list of drivers of misinformation and identified a pattern based on the entities that each of these drivers and influential factors can be related to. The results of this step are discussed in Section 2.3.

2.2.1.4 Analyze Interventions.

We applied the categories of drivers of misinformation on the intervention corpus to analyze the papers in the intervention papers. Specifically, we aimed to map each intervention paper to one or more of the categories of drivers. To provide this mapping, the authors examined each intervention paper to determine what drivers of misinformation are used as the motivation for, or as the focus of, the design of the intervention. If a paper used more than one driver of misinformation to design an intervention, we allow the intervention to sit in multiple categories. To conduct this analysis, the authors read the papers, and discussed the observations to reach a consensus on the category (or the categories) that each paper falls under. The results of this step are discussed in Section 2.3.

2.3 Results

Using the methods described above, this study reviewed prior work that investigated factors that influence the way people identify misinformation and respond to it, and identified four categories: content factors (e.g. Igartua and Cheng, 2009; Sundar et al., 2007; Jahanbakhsh et al., 2021), source factors (e.g. Dias et al., 2021; met, 2019a; Bhuiyan et al., 2021b), individual users factors (e.g. Nickerson, 1998; Bradshaw et al., 2018; Matz et al., 2017; Chen, 2016), and community factors (e.g. Rimal and Real, 2003; Colliander, 2019; Koo et al., 2021; Anspach, 2017).

Based on the analysis described in Section 2.2.1.4, we categorized the interventions designed to address misinformation into four groups: *Content-Based strategies* used to address misinformation; *Source-Based strategies* used to address misinformation; *Individual User-Based strategies* used to address misinformation; *Community-Based strategies* used to address misinformation. Table 2.1 shows a list of papers in each of these categories.

This section reports the different categories of drivers of misinformation identified by prior work, as well as the interventions that are designed to address the spread of misinformation. With the goal to identify strategies that maybe under explored, this section provides a mapping of the interventions to address misinformation to the drivers of misinformation.

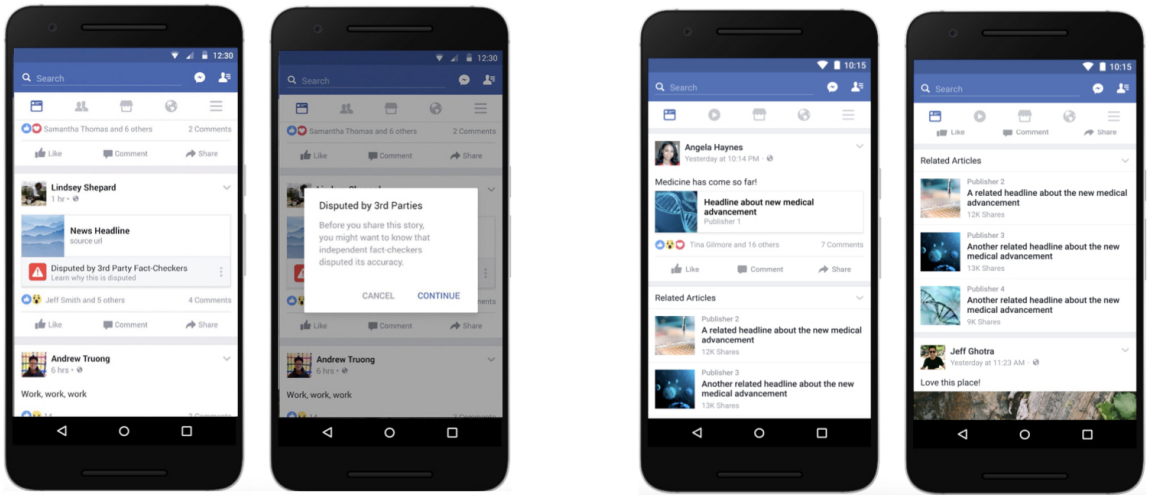
2.3.1 Factors about the Content

2.3.1.1 Content-Related Mechanisms

The characteristics of an individual piece of content (e.g., a news article) can influence how people interact with and respond to it (Igartua and Cheng, 2009; Sundar et al., 2007; Jahanbakhsh et al., 2021; Vosoughi et al., 2018; Spezzano et al., 2021). Such characteristics include the number of quoted sources (Sundar, 1998), prior exposure to a news article (Pennycook et al., 2018), supporting evidence in the article (Jahanbakhsh et al., 2021), whether or not a piece of content appears biased or not (Jahanbakhsh et al., 2021), and the number of related articles written about the same news event by other news organizations

Table 2.1: List of Papers in Intervention Corpora Categorized Based on The Driver of Misinformation that the Interventions Focus on

The orientation of Intervention	**Number of Articles**	**References**
Content	40 papers	(Facebook, 2020; AssociatedPress, 2021; Facebook, 2021; Vlachos and Riedel, 2014; Hunt, 2017; Castillo et al., 2011; Nagura et al., 2006; Potthast et al., 2016; Zhang et al., 2018; Facebook, 2013; Garber, 2012; Cappella and Jamieson, 1994; Clayton et al., 2020; Yaqub et al., 2020; Zubiaga et al., 2016; Shin et al., 2017; Shin and Thorson, 2017; Garrett and Weeks, 2013; Ortutay, 2017; Arif et al., 2017; Stray, 2017; Nguyen et al., 2018a; **?**; Kirchner and Reuter, 2020; Chan et al., 2017; Garrett and Poulsen, 2019; Nyhan and Reifler, 2010; The Economic Time, 2020; Bode and Vraga, 2015; Hamborg et al., 2017; Team, 2019; Buntain et al., 2021; met, 2019b; Meta, 2020; Epstein et al., 2020; Spezzano et al., 2021; Horne et al., 2020; Sultana and Fussell, 2021; Gao et al., 2018; Mosleh et al., 2021)
Source	10 papers	(Dias et al., 2021; Pennycook and Rand, 2019; Schwarz and Morris, 2011; Im et al., 2020; Epstein et al., 2020; Bhuiyan et al., 2021a; Dias et al., 2020; Horne et al., 2020; Spezzano et al., 2021; Kane et al., 2018)
Individual Users	16 papers	(Roozenbeek and van der Linden, 2019; Basol et al., 2020; van Der Linden et al., 2020; Basol et al., 2021; Tsipursky and Morford, 2018; Karduni et al., 2019; Pennycook et al., 2020, 2021; Jahanbakhsh et al., 2021; Bhuiyan et al., 2018; Facebook, 2018; twi, 2018, 2019; Cox and Koebler, 2019; Bilton, 2019; Jeon et al., 2021)
Community	4 papers	(Andı and Akesson, 2020; Nguyen et al., 2018b; Kim et al., 2018; Bhuiyan et al., 2021a)



((a)) ((b))

Figure 2.2: a) Facebook uses third-party fact-checkers to mitigate the spread of misinformation (Isaac, 2016). b) Facebook suggests related-articles to provide additional information on articles with low credibility (Constine, 2017).

(Sundar et al., 2007; Jahanbakhsh et al., 2021). Put differently, each of these content-related factors can work as a heuristic that guides individuals' response to (mis)information.

Such factors can also be manipulated to influence perceptions about the reliability of content, and information processing. For example, examining the role of framing in individuals' assessment of articles that evaluate truthfulness of news, Kreiner and Gamliel (2021) show that framing effect influence the perceived reliability of the evaluation articles when the outcomes were favorable but not the outcomes were unfavorable.

2.3.1.2 Content-Based Strategies Used to Address Misinformation

We find three general approaches that focus on specific pieces of content: disputing False information Using Fact-checkers, signaling credibility of content, and reducing the visibility of misleading content. For each, we describe some of the dominant approaches, and summarize findings from studies of how these countermeasures are used in practice.

2.3.1.2.1 Disputing False information Using Fact-checkers One of the most common approaches to mitigate the spread of misinformation is disputing false information using fact-checkers (Facebook, 2020; AssociatedPress, 2021). Fact-checkers assess the veracity of different claims made by public figures (e.g., politicians, pundits, corporations, etc.) that

are likely to be misleading based on several signals. These signals include observing expression of dis-beliefs to these content, the speed by which these content spread, and the output of machine learning models that predict false information (Facebook, 2021; Vlachos and Riedel, 2014; Cohen et al., 2011).

A variety of entities work to provide accurate fact-checkers, from social media companies such as Facebook and Twitter (Facebook, 2020; AssociatedPress, 2021; Hunt, 2017), to journalists (The Washington Post, 2013; Journal, [n. d.]), to academic institutions (Castillo et al., 2011; Nagura et al., 2006; Potthast et al., 2016; Zhang et al., 2018). There are different forms of fact-checkers, from websites such as PoliticFact and Snopes, which evaluate factual claims of news, to credible news media such as Washington Post and Wall Street Journal, to content moderation techniques (either paid agents or volunteers) used by platforms such as Facebook and Reddit (Facebook, 2013; Garber, 2012), to automatic fact-checkers (Hassan et al., 2015). Figure 2.2(a) shows an example of disputing a piece of news on Facebook using fact-checkers. Prior studies demonstrate the success of fact-checkers in identifying low credible information and mitigating its spread (Cappella and Jamieson, 1994; Clayton et al., 2020; Yaqub et al., 2020).

A large body of research has investigated the ways online users engage with fact-checkers and show individual responses to fact-checkers vary based on different factors (Zubiaga et al., 2016; Shin et al., 2017; Shin and Thorson, 2017; Garrett and Weeks, 2013; Ortutay, 2017; Sultana and Fussell, 2021; Mosleh et al., 2021; Gao et al., 2018). For example, examining people's response to rumors on Twitter before and after their veracity is determined by fact-checkers, Zubiaga et al. (2016) show many people share tweets that support a rumor which is still not verified. However, once the rumor has been debunked, they are less likely to make the same effort to communicate to their followers that the content they previously shared was untrue. In another case, Shin et al. (2017) investigate the spread of rumors on Twitter during the 2012 U.S. presidential election and show that people choose the outcome of fact-checkers subjectively. That is, partisans selectively shared fact-checkers' messages that were advantageous to their group and denigrated the opposing group. As a result, rumors had been propagated even after fact-checker organization debunked them. This subjective use of fact-checkers occurs because many individuals tend not to question the credibility of

information unless it contradicts their view and prompts them to do so (Lazer et al., 2018). These findings align with the phenomenon of selective exposure, where individuals prefer to read information that is in line with their prior beliefs (Garrett, 2009a), and avoid the content that contradicts their beliefs (mut, 2006).

Mosleh et al. (2021) investigate downstream consequences of social corrections on the users' future content sharing behavior. Specifically, (Mosleh et al., 2021) investigate the behavior of large scale users on Twitter who shared false information and received replies to their false tweet with links to fact-checkers. Examining of the users' subsequent activities shows that this method decreased the quality of content they shared, and increased language toxicity.

2.3.1.2.2 Signaling Credibility of Content Relatively less work has explored different ways to give insights about the credibility of content (**?**Kirchner and Reuter, 2020; Hamborg et al., 2017). Facebook, for instance, displayed red flags on articles that were disputed by fact-checkers to signal their lack of credibility. Warning against misinformation has been shown to reduce its perceived accuracy (Chan et al., 2017; Lewandowsky et al., 2012; Kirchner and Reuter, 2020). However, prior research suggests that strong language or visualization in warnings can backfire and strengthen prior beliefs (Seifert, 2002; Garrett and Poulsen, 2019; Lewandowsky et al., 2012; Nyhan and Reifler, 2010).

In addition, warnings such as red flags are likely to prompt people to click on the false content (Lewandowsky et al., 2012; Ortutay, 2017). Clicking on false information in turn increases the visibility of content, the risk of repeating false information, and may result in the increase in accepting misinformation as true (**?**Lazer et al., 2018). Lastly, a red flag can only signal the credibility of content that is false, and cannot communicate any information on the veracity of partly false and unproven content.

Therefore, Facebook removed the red flag feature in favor of a new feature, named "Related Articles" (Ortutay, 2017). Unlike the fact-checkers that inform whether a content is "true" or "false", this approach aims to provide people with additional information (e.g., reporting from a certified third-party fact-checker or the stories published by another publisher) and help them decide by themselves whether the news is misleading or not. Figure

2.2(b) shows an example of how related articles are suggested under low credible or unproven articles. Similarly, Twitter uses labels or warning messages to provide additional information or clarifications about tweets that contain harmful and unconfirmed claims (The Economic Time, 2020).

Kirchner and Reuter (2020) examine the efficacy of different warning-based approaches (i.e., a simple warning that shows the article is disputed, related articles underneath the post with headlines contradictory to the false claim, and a warning extended by a short explanation). The results of their experiment show that all these warning messages are effective, but adding an explanation to warning messages is the most effective warning-based approach. Specifically, adding an explanation to a warning is shown to be more effective than the related articles approach. In another case, Bode and Vraga (2015) show when people have strong prior beliefs about an issue (e.g., anti-vaccine beliefs), providing related stories are less likely to correct initial misconceptions that misinformation creates. Instead, in this case corrective information might even backfire and result in accepting false information as correct more strongly.

2.3.1.2.3 Reducing visibility of misleading content Reducing the visibility of harmful content, including misleading information, has been used by several social media companies to mitigate the spread of such content (Team, 2019; Buntain et al., 2021)

One approach to lower the visibility of misleading content is to use the crowds to rate trustworthiness of content, and use this information in the ranking algorithms of news feeds (met, 2019b; Meta, 2020). Epstein et al. (2020) investigate the efficacy of this approach, and examine whether or not laypeople game this crowdsourcing mechanism to promote content of their own interests. Their findings suggest people are less likely to game the system. Indeed, the participants trusted mainstream sources much more than hyper-partisan or fake news sources, regardless of their partisanship. However, many people distrusted unfamiliar outlets. Epstein et al. (2020) argue that while using the crowds trust ratings are effective in discerning between high and low quality content, this approach may still result in a rise in polarization.

In another case, YouTube claims it uses a combination of machine learning methods and

human evaluation to identify misleading, harmful videos and videos that include borderline content (Team, 2019). While YouTube does not remove these videos, it removes them from recommendation systems as a way to lower their exposure. Reddit uses this YouTube approach to lower the visibility of misleading content. Buntain et al. (2021) investigate the efficacy of this approach on Twitter and Reddit, and demonstrate that de-recommendation results in a significant decrease in sharing conspiracy labeled content on both platforms. However, this approach resulted in an increase of sharing conspiracy labeled content on conspiracy oriented communities on Reddit. These findings suggest that reducing exposure to harmful, misleading content might be beneficial for some communities, but harmful for some others (e.g., the communities on the borders).

Another approach to reduce visibility of misleading content, employed by Facebook, is reducing the size of low credible headlines and articles. Kirchner and Reuter (2020) conducted semi-structured interviews to investigate the efficacy of this approach (in addition to several other approaches of addressing misinformation). While some of the participants found this method favorable as it draws attention away from misleading content, some others believe that it is rather an extreme approach and almost like censoring content. Overall, participants preferred warning based approaches that provide additional information over reducing the size of misleading content.

2.3.2 Factors about the Source

2.3.2.1 Source-Related Mechanisms

In addition to factors related to individual pieces of content, factors related to the source of information can influence individuals' assessment of that information (Dias et al., 2021; Lewandowsky et al., 2012; met, 2019a; Bhuiyan et al., 2021b; Hovland and Weiss, 1951; Dias et al., 2020). Lewandowsky et al. (2012) explain that when people lack the motivation and the knowledge to investigate a message in detail, they tend to defer to their assessment of the source's credibility. If the source is perceived as credible, people are more likely to consider the content as credible (Eagly and Chaiken, 1993; Petty and Cacioppo, 1986). In this case, a lack of detailed information about a source may decrease users' perceptions of

the credibility of content from that source (Freeman and Spyridakis, 2004).

Multiple factors can influence the perceived credibility of a source. These factors include the "look and feel" of a website (Fogg et al., 2001, 2003), how information on the site is structured (Fogg et al., 2001), the professionalism of the site design (McKnight and Kacmar, 2007), and official-looking logos and the domain names (Wineburg and McGrew, 2017).

Kahan (2017) argues that many people's assessment of news sources can also be influenced by their partisan bias. In this case, as a result of motivated assessment, unreliable content may be perceived as reliable. Pennycook et al. (2020), however, demonstrate that failing to assess misinformation often results from a lack of reasoning about, rather than a from motivated assessment of, news sources. In another study, Epstein et al. (2020) show that if people are asked to reflect on trustworthiness of news sources, their judgment may not be disproportionately swayed by their partisanship. Pennycook et al. (2020) also argue that if prior experience with an outlet and the content the outlet shares is necessary to shape an accurate assessment of its reliability, many people cannot judge most outlets as there are so many outlets with which people may have not experienced. Future research is therefore required to investigate how people assess the legitimacy of news outlets, and to examine the link between their partisanship and their assessment of the news sources.

2.3.2.2 Source-Based Strategies Used to Address Misinformation

Recent work investigates leveraging the impacts of the news source and its perceived credibility on individuals' assessment of a message (Dias et al., 2021; Bhuiyan et al., 2021a; Pennycook and Rand, 2019; Spezzano et al., 2021). For example, Bhuiyan et al. (2021a) employed a nudge-based intervention based authority of a source, referred to as the *reliable* nudge. They designed a browser extension for Twitter, named NudgeCred to investigate the efficacy of this nudge-based intervention. Through a five-day field experiment, this study demonstrates that NudgeCred improves individuals' recognition of misinformation. That is, signaling authority of source influenced the perceived credibility of the content. In particular, the participants rated posts with *reliable nudge* as more credible. (This study also investigated a community-based nudge, as discussed in Section 2.3.4.2).

In another case, Schwarz and Morris (2011) present a visualization tool which augments

search results with information on the credibility of web pages to help online users with credibility assessment. Conducting a user study, they demonstrate that signaling credibility of web pages can help people to assess information and identify credible content from non-credible content. However, this approach can result in perceiving the content published in new outlets whose credibility is unknown as less credible (Schwarz and Morris, 2011). In a similar approach, Im et al. (2020) design an approach to provide social signals about online accounts and inform others about whether a certain account engages in the spread of misinformation. The accounts that spread misinformation will receive a tag of misinformation where all other users can see. Through a field study, they investigate the utility of the approach and show the participants find these social signals helpful. Future research is required to investigate whether and how such signals might impact the behaviors of those users who receive these misinformation tags on their profile. Specifically, future research should investigate whether this approach prompt the receiver of this tag to engagement more with misinformation spread.

Epstein et al. (2020) investigate using crowdsourcing to identify outlets that produce misinformation. Next, they use the crowds' ratings as an input to social media ranking algorithms. The results of their study demonstrate layperson trust ratings are an effective way to distinguish between high and low quality news outlets. The results of their experiment show the participants tend to trust mainstream sources much more than fake news sources. To test whether the participants would use this approach as a way to game the system, the participants were told their trust rating of news outlets will influence ranking algorithms of social media feeds. This information did not influence participants to game the system and promote their content of interest. In addition, Epstein et al. (2020) argue that focusing on the source evaluation is more practical than focusing on individual articles for two reasons. First, rating credibility of news sources requires a much lower volume of rating. Thus, source-level rating is more scalable due to fewer number of news sources than news articles. Second, source-level ratings seem less susceptible to variation based on the idiosyncrasies of specific headlines. In another case, Pennycook and Rand (2019) also investigate the crowds' rating of news sources trustworthiness, and suggests using these ratings in social media ranking algorithms. They also acknowledge crowd-sourced trust ratings can

effectively differentiate more versus less reliable sources (i.e., the participants of their study rated mainstream sources as more trustworthy than hyperpartisan or fake news sources.)

In another example, using survey experiments, Dias et al. (2021) investigate whether or not making the publisher information of a post more visible via adding a logo banner can improve assessment of the post. They demonstrate that publisher information had no significant effects on whether participants perceived the headline as accurate. When the headline was accurate but the publisher was a distrusted source, people tend to rate the content as less credible. Put differently, providing publisher information could increase the likelihood of mistakenly perceiving true headlines as false. Future research is required to investigate how providing different types of information about the source (e.g., adding a logo banner vs. detailed information) might have different efficacy in news assessment.

2.3.3 Factors about the Individual User

2.3.3.1 Individual User-Related Mechanisms

The characteristics of individual users can impact how they interact and respond to different pieces of misinformation (Nickerson, 1998; Garrett, 2009a; Moravec et al., 2018; Aufderheide, 2018; Association et al., 2009; Doughty et al., 2017; Chen, 2016; Buchanan and Benson, 2019). These characteristics include people's prior beliefs (Nickerson, 1998; Garrett, 2009a; Moravec et al., 2018), their media literacy skills (Aufderheide, 2018; Association et al., 2009; Jones-Jang et al., 2021; Buchanan, 2020), as well as their personalities (Buchanan and Benson, 2019; Doughty et al., 2017; Chen, 2016; Buchanan and Benson, 2019). The remainder of this subsection describes how each of these attributes play a role in people's response to misinformation.

2.3.3.1.1 Prior Beliefs. Individuals' prior beliefs about different topics influence news consumption (Nickerson, 1998; Garrett, 2009a; Moravec et al., 2018; Minas et al., 2014), and news assessment (Nickerson, 1998), but do not necessarily predict the way people share information (Pennycook et al., 2021).

Individuals' *news consumption* are influenced by their prior beliefs. That is, many individuals prefer to read and seek out information that is in line with their prior beliefs (Met-

zger and Flanagin, 2013; Jahanbakhsh et al., 2021). This phenomenon is known as selective exposure effect (Garrett, 2009a,b), and contributes to the spread of misinformation and its impacts. For example, Del Vicario et al. (2016) show that selective exposure plays a role in the spread of conspiracy theories on Facebook. Selective exposure can also result in homogeneous and polarized communities (Del Vicario et al., 2016; Lehmann and Ahn, 2018; Moore and Tambini, 2018), wherein people are more exposed to consonant content and less exposed to incompatible arguments. Such a closed system contributes to the creation of echo chamber (Garrett, 2009a; Pariser, 2011; Prasetya and Murata, 2020), wherein people mostly see content that agrees with their preexisting.

Prior beliefs also influence the process of *news assessment* (Nickerson, 1998). That is, many people tend to believe information that confirms their preexisting beliefs, a concept known as confirmation bias (Nickerson, 1998). Due to confirmation bias, many people do not question the veracity of information with which they agree (Nickerson, 1998; Moravec et al., 2018; Lazer et al., 2017). Confirmation bias can even influence people not to respond to corrective information objectively (Lazer et al., 2018; Hameleers and van der Meer, 2020). In one case, Sleegers et al. (2019) demonstrate many people tend to interpret ambiguous feedback on their incorrect beliefs in favor of their beliefs. In another case, Hameleers and van der Meer (2020) illustrate that many people tend to avoid the articles that evaluate truthfulness of news (i.e., fact-checkers) if those outcomes are incompatible with their prior beliefs (Hameleers and van der Meer, 2020). Similarly, prior beliefs can prevent some people from accepting corrective information (Taber and Lodge, 2006; Lazer et al., 2018). Taber and Lodge (2006) argue that when people's prior beliefs are questioned, it might have a boomerang or backfire effect (Byrne and Hart, 2009). This backfire effect occurs because many people tend to actively counterargue incongruent evidence (i.e., disconfirmation bias) (Taber and Lodge, 2006). They engage in "motivated reasoning" (Flynn et al., 2017; Kunda, 1990), bringing arguments to defend their prior beliefs (Kahan, 2012). As a result, their prior beliefs will become stronger and they are more likely to believe in the original false content.

While prior beliefs influence *news consumption*(Garrett, 2009a; Metzger and Flanagin, 2013), and *news assessment* (Nickerson, 1998; Moravec et al., 2018; Jahanbakhsh et al.,

2021; Hameleers and van der Meer, 2020), recent research suggests *sharing information* is not necessarily a reflection of prior beliefs (Pennycook et al., 2021). Individuals' decisions about information sharing are more likely influenced by peripheral cues, such as the source of the information, and are moderated by individuals' self-presentation goals (Ceylan and Schwarz, 2020). For example, individuals who want to fit in with a group prefer to share popular information from sources generally well-known and perceived as credible within their group. However, individuals who aim to stand out tend to share a piece of content regardless of the popularity of the content or its source (Ceylan and Schwarz, 2020).

2.3.3.1.2 Media-Related Literacy. Researchers have explored the link between different types of media-related literacy (i.e., media literacy, information literacy, news literacy, and digital literacy) and the way people identify and respond to misinformation (Aufderheide, 2018; Association et al., 2009; Jones-Jang et al., 2021; Buchanan, 2020). For example, investigating the role of digital media literacy on the way people respond to misinformation, Aufderheide (2018) shows that people with greater digital media literacy are less susceptible to the misinformation and its impacts. In another case, Jones-Jang et al. (2021) show that while both digital media literacy and information literacy are important, the latter is comparatively more influential on the way people respond to misinformation.

Sirlin et al. (2021) investigate the role of digital media literacy, and acknowledged that of digital literacy is associated with less ability to tell truth from falsehood. This result does not vary based on the participants' partisanship and the type of the news. However, the skills of digital media literacy does not make people less likely to *share* false information. This finding could be because people are less motivated with the accuracy of content than with factors such as their emotion, and social feedback (Pennycook and Rand, 2020; Huang et al., 2015; Martel et al., 2020). Buchanan (2020) also argues that while media literacy is an important variable mediating the spread of misinformation and its impacts, people may know that a piece of content is untrue and still spread it anyway. They might be sympathetic to the content's intentions, or aim to signal their social identity or adherence to some political group or movement.

2.3.3.1.3 Personality. A line of research investigates whether personality traits have any effects on people's engagement with and their response to misinformation (Doughty et al., 2017; Chen, 2016; Buchanan and Benson, 2019). To measure personality traits, the Five Factor Model of personality (FFM), also known as the "Big Five" Costa Jr and McCrae (2008), has been widely used (Buchanan and Benson, 2019; Chen, 2016). FFM includes extraversion, agreeableness, conscientiousness, neuroticism, and openness (See (Costa Jr and McCrae, 2008) for more details on these factors.)

Chen (2016) shows that *openness* has a positive influence on sharing misinformation, while *neuroticism* has a negative influence on the sharing of misinformation. However, Buchanan and Benson (2019) demonstrates that of the five dimension of personality, only *Agreeableness* personality had an impact on people's response to misinformation.

Zhu et al. (2010) argue to investigate why people are influenced by misinformation, the role of personalities should be investigated along with other individuals' characterises that can induce false memories. They investigate the interaction effects between personality characteristics and cognitive abilities on individuals' vulnerability to misinformation. The result of their study demonstrate that low fear of negative evaluation, low harm avoidance, high cooperativeness, high reward dependence, and high self-directedness in combination with relatively low cognitive abilities make people more vulnerable to misinformation. In another study, Doughty et al. (2017) show that personality characteristics are associated with memory conformity, which can make some people more susceptible to accepting misinformation as true. Specifically, they suggest low openness, extraversion, low neuroticism, and high agreeableness are related to memory conformity, and can make people more likely to accept misinformation as true.

As is apparent in this discussion, the results of prior studies that investigate the link between personality characteristics and people's susceptibility to the spread of misinformation draw different conclusions. Therefore, future research is required to clarify the sources of these mixed findings.

2.3.3.2 Individual User-Based Strategies Used to Address Misinformation

Among the papers that focus on addressing misinformation at the individual user level, we found four general approaches: interventions that work towards improving individuals' media literacy skills (e.g., Roozenbeek and van der Linden, 2019; Basol et al., 2021), interventions aiming at shifting an individual's attentions towards credibility of content (Pennycook et al., 2021, 2020; Jahanbakhsh et al., 2021; Kane et al., 2018), interventions that aims at identifying inauthentic accounts (e.g., Facebook, 2018; twi, 2018, 2019), and interventions that work towards identifying and deplatforming dissimulators of misinformation (e.g., Twitter., 2021).

2.3.3.2.1 Improving Individuals' Media literacy Skills Various approaches have been designed to improving individuals' media literacy skills (Roozenbeek and van der Linden, 2019; Basol et al., 2020; van Der Linden et al., 2020; Tsipursky and Morford, 2018). In recent years, several game-based psychological interventions have been investigated to improve individuals' skills to identify misinformation (Micallef et al., 2021; Roozenbeek and van der Linden, 2019; Jeon et al., 2021; Muscat and Duckworth, 2018). For example, Roozenbeek and van der Linden (2019) designed a game, named "the bad news game", wherein players learn various skills to mitigate misinformation spread, such as detecting discredit, polarized arguments, impersonation, etc. Basol et al. (2021) also designed another game, named "Go Viral!" to improve abilities of individuals to detect manipulation techniques that are used in COVID-19 misinformation. The game increased the abilities of the participants to identify misinformation, and reduced their willingness to share misinformation with others. In another case, Jeon et al. (2021) designed a game, named "ChamberBreaker", which is designed to increase a player's awareness of echo chamber effect and the importance of maintaining diverse perspectives when consuming information. After playing the game, the players showed greater intention to see information from more diverse perspectives and more awareness of the possible echo chambers.

Karduni et al. (2019) designed a visual analytic system, named "Verifi2", that help social media users distinguish misinformation. Verifi2 highlights different aspects of a piece

of news, such as its linguistics, its networks of spread, and highlights image features related to the news to help online users to learn dimensions that characterize misinformation, and learn how suspicious news are different with true content.

In another case, Tsipursky and Morford (2018) designed an intervention named the Pro-Truth Pledge (PTP) where the signees agree to abide by twelve behaviors to mitigate the spread of misinformation, including verify, balance, cite, clarify, acknowledge, reevaluate, defend, align, fix, educate, defer, and celebrate. (See (Tsipursky and Morford, 2018) for more details). The participants, including both private citizens and public figures, self-reported the impact of this pledge on their behaviors regarding their response to misinformation.

While valuable, improving people's skills to identify false and misleading content does not always lead to changing people's intention to share misinformation. For example, many people may know that a piece of content is untrue and still spread it anyway for reasons such as signaling their social identity or adherence to some groups (Buchanan, 2020; Jones-Jang et al., 2021). To improve the efficacy of the interventions around media literacy, other factors such as the role of prior beliefs should also be taken into account. In addition, the studies that are reviewed here analyse the efficacy of the intervention based on self-reported data. Future research should investigate how these approaches may help in real-word settings, outside the experiment. Additionally, future investigation is required to assess the long-term efficacy of these prebunking-based and pledge-based interventions.

2.3.3.2.2 Shifting an Individual's Attention to the Credibility of Content Only recently researchers focused on reducing the spread of misinformation via shifting attention of online users toward the credibility of what they share (Pennycook et al., 2021, 2020; Jahanbakhsh et al., 2021; Bhuiyan et al., 2018). Recent work shows that many people are motivated by their emotions and social feedback when sharing a piece of content online (Huang et al., 2015; Martel et al., 2020). Motivated by these insights, Pennycook et al. (2021) examine whether encouraging online users to reflect on the accuracy of content could make them less likely to spread misinformation (Pennycook et al., 2021). Using a field experiment on Twitter, they selected Twitter users who regularly shared misinformation. Next, the researchers send those users private messages, asking them to rate the accuracy

of a headline. By inviting these users to reflect on the accuracy aspect of the news, this intervention increased the average accuracy of the news that those users shared in the next 24 hours.

In another example, Jahanbakhsh et al. (2021) investigated the effects of two behavioral nudges which request accuracy assessments and rationales, on sharing false news. Specifically, the first nudge asked people to assess the accuracy of the content they were about to share, and the second nudge asked people why they think the content is or is not accurate at the time of sharing it. The results of this study revealed that both accuracy assessment nudge and rational nudge can reduce the sharing both false content and true content. However, these nudges also reduced sharing of true content to a lesser degree compared to the sharing of false content, resulting in an overall decrease in the fraction of shared content that is false (Jahanbakhsh et al., 2021).

2.3.3.2.3 Identifying and Removing Disseminators of Misinformation At the first glance, this category of interventions focus on the creators of misinformation (e.g., identifying and removing inauthentic accounts, deplatforming disseminators of misinformation), which does not fall in the scope of this review. However, these interventions, by removing the disseminators of misinformation and signaling the audience about it, may still impact people's perceptions of and their *response* to the content they have previously read from these accounts. Therefore, this review includes and examines these approaches as well.

One of the approach in this category is *identifying and removing inauthentic accounts* (Facebook, 2018; twi, 2018, 2019). This approach focuses on the activities of individual users and addresses the users who engage in suspicious behaviors and spread misinformation(e.g., artificially boost the popularity of content, or impersonating another person such as politicians or celebrities). Platforms such as Facebook and Twitter, identify and remove accounts that exhibit inauthentic behavior (Facebook, 2018; twi, 2018, 2019). For example, Facebook introduced a concept named "coordinated inauthentic behavior" (CIB), which is defined as the use of Facebook or Instagram assests (i.e, accounts, pages, groups, or events) to mislead people (Facebook, 2018). The networks of people who engage in coordinated inauthentic behavior focus on two activities. In one case, they create fictitious, independent media

entities and personas to engage unwitting individuals to amplify their content and expand their reach. In another case, they drive people to other websites that their networks control (Facebook, 2018). Facebook identities and disables these accounts. Upon disabling an account, all the content within the account becomes inaccessible to other users.

While identifying and removing these networks of accounts might mitigate the spread of misinformation (met, 2021), recent research reveals that the success of these accounts heavily lies on the activities of online crowds (as apposed to coordinated networks of accounts) and the ways they engage in the spread of misinformation and its impacts (Vosoughi et al., 2018; Grinberg et al., 2019; Starbird et al., 2019). The engagement of online communities with misleading content, amplifying it, and expanding its reach makes it possible for coordinated networks to expand their networks and push their goals. Therefore, policies that focus on addressing misinformation by identifying "coordinated" inauthentic behaviors fall short in addressing the circulation of misleading content that is shared by online crowds.

Figure 2.3: The figure shows deplatforming of @realDonaldTrump account, which occurred on January 8th, 2021. Twitter announced the account due to the risk of further incitement of violence (Twitter., 2021).

Another approach in this category is *deplatforming*, a moderation strategy that refers to the permanent ban of conversational influencers for spreading misinformation, conducting harassment, or violating other platform policies (Grimmelmann, 2015). Facebook, Twitter, Instagram, YouTube and other platforms have employed deplatforming to mitigate the spread of misinformation, hate speech, and conspiracy theories in various cases (Cox and Koebler, 2019; Koebler, 2018; Bilton, 2019). For example, upon investigating tweets from the @realDonaldTrump account and the way they were being interpreted on and off Twitter, Twitter suspended the account permanently to prevent the risk of further inducement of

violence by this account (Twitter., 2021). (See Figure 2.3).

While deplatforming has been shown to decline toxicity level of supporters (Jhaver et al., 2021), the efficacy of this approach to combat misinformation is nuanced by a lot of factors. For example, in the case of combating anti-vaccine misinformation, Armitage (2021) argues that deplatfoming of anti-vaccine campaigners is likely to reinforce individuals' strongly held beliefs about vaccination and vaccine conspiracies. In addition, Innes and Innes (2021) argue that de-platforming might result in "re-platforming", which refers to behaviors such as developing a network of alternative accounts and signaling their presence, and migrating to one or more other platforms.

2.3.4 Factors about the Community

2.3.4.1 Community-Related Mechanisms

Prior work has acknowledged the different mechanisms by which online communities contribute to the spread of misinformation and its impacts (Schwarz et al., 2007; Begg et al., 1992; Swire et al., 2017b; Thorson, 2015; Hasher et al., 1977; Hermida et al., 2012; Geeng et al., 2020). These mechanisms can be organized into roughly three groups: individual level mechanisms, such as familiarity bias and trust in community-shared content (Schwarz et al., 2007; Begg et al., 1992; Swire et al., 2017b; Thorson, 2015; Hasher et al., 1977; Huckfeldt et al., 1995; Hermida et al., 2012; Geeng et al., 2020); network mechanisms, such as social network structure and homophily (Hermida et al., 2012; Geeng et al., 2020; DiFonzo et al., 2013); and social norms, especially perceived norms (Rimal and Real, 2003; Colliander, 2019; Koo et al., 2021; Gimpel et al., 2021). The remainder of this subsection describes how each of these mechanisms works in the spread and the impacts of misinformation.

First, online communities can influence the way individual members respond to misinformation via different mechanisms. For example, the activities of online communities (e.g., sharing, liking, commenting) can increase the likelihood of individual members' exposure to misinformation. For instance, when a contact of ours "likes" a post, we are more likely to see the content (Anspach, 2017). This increased visibility can also increase individuals' exposure to misinformation, which contributes to familiarity bias (Schwarz et al., 2007; Begg

et al., 1992; Swire et al., 2017b; Thorson, 2015; Hasher et al., 1977), wherein an individual remembers the content itself, but forgets the contexts and details around it. Familiarity bias increases the likelihood of accepting familiar but false information as true (Schwarz et al., 2007; Begg et al., 1992; Swire et al., 2017b; Thorson, 2015; Hasher et al., 1977), leading to the further spread of misinformation.

Another mechanism by which online communities influence individual members' response to misinformation is via increasing trust in the content that is shared in the community (Hermida et al., 2012). That is, many people are more likely to trust what their networks share (Hermida et al., 2012) and less likely to question content that is shared by their networks (Hermida et al., 2012). As a result, the content that is shared by an individual's communities are more likely to be perceived as true (Geeng et al., 2020).

Second, the structure of social networks within and across online communities play a role in how misinformation spread and has its impacts (Friggeri et al., 2014; Lazer et al., 2017; Lerman et al., 2016). For example, online communities might form homophilous network structures, where there are more connections among people who share similar views (Osmundsen et al., 2020; Quattrociocchi et al., 2016). Such networks make people more likely to be exposed to consonant information, and less likely to observe arguments that might challenge their view (mut, 2006; Smith et al., 2013; Conover et al., 2011; Adamic and Glance, 2005; Bishop and Myers, 1974; Burnstein and Vinokur, 1973; Huckfeldt et al., 2004; Scheufele and Krause, 2019). For example, examining two polarized communities on Facebook, science and conspiracy communities, Quattrociocchi et al. (2016) show that people of each community tend to be connected only with like-minded people and not to interact with people of the other community. In such networks, people's views are less likely to be challenged, leading many individuals to become more confident in their views (Sunstein, 1999). Homophily can also explain why people in the same network are more likely to believe in the same rumor (DiFonzo et al., 2013). This way, networks that connect similar individuals make it possible for misinformation to spread more quickly within a community.

Third, social norms are another key mechanism by which online communities play a role in the spread of misinformation and its impacts. Social norms refer to people's perceptions

around what others do (i.e., descriptive norms), what others approve of and what they condemn (i.e., injunctive norms), and how an individual thinks they are expected to behave (i.e., subjective norms) (Rimal and Real, 2003; Cialdini and Trost, 1998). These perceptions, combined with the desire to be liked or to obtain approval from others, can influence people's behaviors (Rimal and Real, 2003; Cialdini and Trost, 1998). These behaviors include the types of content that people share within their community (Olson and Zanna, 1979; Atkin, 1985; Andı and Akesson, 2020; Park et al., 2012; Wojcieszak et al., 2020), and the way they identify misinformation and respond to it (Colliander, 2019; Koo et al., 2021; Gimpel et al., 2021). For example, descriptive norms around content sharing (i.e., perceptions about what other members of the community share) can influence the perceived popularity of certain information. The perceived popularity of a claim in turn increases its perceived reliability (Bacon, 1979; Begg et al., 1992), making it more likely to be accepted as true. Injunctive norms can influence people's perceptions about what a community approves or disapproves of sharing. People also form perceptions about how they are expected to behave (i.e., subjective norms), which influence their decision about whether or not to share a piece of content (Colliander, 2019). Perceptions of social norms can also influence the way people assess and respond to misinformation (Colliander, 2019; Koo et al., 2021; Gimpel et al., 2021). For example, people who perceive of correcting misinformation as a common practice within a community may be more likely to correct misinformation themselves (Koo et al., 2021).

At the same time, individuals' may have inaccurate perceptions of social norms (Rimal and Real, 2003). Such misperceptions can contribute to the "majority illusion effect" (Kempe et al., 2003), wherein a few highly connected users with misleading views can skew the perceptions of many others and even trigger a rapid change in the community's view (Kempe et al., 2003; Lerman et al., 2016). For example, misperceptions due to majority illusion effect can make anti-vaccine views seem like the majority's opinion, even when they are not (Song and Gruzd, 2017; Johnson et al., 2020). The majority illusion effect induces some people to overestimate the prevalence of their view in the population and assume the majority share their view (Luzsa and Mayr, 2021), a concept known as "false consensus" (Marks and Miller, 1987; Luzsa and Mayr, 2021). Additionally, majority

illusion can make people incorrectly think the opposite of their view is held by the majority of the population when in fact the majority share their view, a concept known as "pluralistic ignorance" (Prentice and Miller, 1996; Miller and McFarland, 1991). As a result of the pluralistic ignorance effect, group members may behave contrary to their own preferences in favor of what they think is popular. The pluralistic ignorance effect may be a key factor that contributes to people's hesitation to correct perceptions of others about misinformation (Lewandowsky et al., 2012). Thus, the perception of social norms can play a key role in how communities respond to and spread misinformation.

2.3.4.2 Community-Based Strategies Used to Address Misinformation

Despite the significant role of online communities in the spread of misinformation and its impacts (Discussed in Section 2.3.4), much less attentions has been paid to designing around community factors that contribute to the spread of misinformation. We found only two interventions that are designed based on community factors for this purpose (Andı and Akesson, 2020; Bhuiyan et al., 2021a). We found several interventions that leverage the crowds to improve the efficacy of approaches that aim at identifying misinformation (Kim et al., 2018; Nguyen et al., 2018b,b). However, these interventions are not designed around the community-oriented factors of misinformation spread.

In one case, Andı and Akesson (2020) investigate whether a social norm-based nudge can result in sharing less misinformation. To do so, they use a message that inform participants about the abundance of false information online and warns them that most responsible people think twice before sharing a piece of news with their networks (i.e., descriptive norms intervention). The message is displayed above the articles that the participants see and reads as "NOTICE: There is a lot of misleading and false information online. Most responsible people think twice before sharing content with their friends and followers". The participants who were nudged using this message expressed less willingness to share false information. However, it is not immediately clear whether the result is solely due to priming effect of informative message about the abundance of false information, or due to conformity to social norms.

In another case, Bhuiyan et al. (2021a) employed a nudge-based intervention based

on users' collective opinion on a report. This nudge is designed to highlight the number of question marks in the comment section, and is named as the *questionable* nudge. To investigate the efficacy of this approach, Bhuiyan et al. (2021a) designed a browser extension for Twitter, named NudgeCred. Through a five-day field experiment, they demonstrate the NudgeCred influenced the perceived credibility of the content. That is, the participants rated posts with *questionable nudge* as less credible.

A few other studies leverage online crowds to improve the efficacy of automatic fact-checkers (Kim et al., 2018; Nguyen et al., 2018b). For example, Nguyen et al. (2018b) combine machine learning techniques with the crowd annotations to improve the efficacy of fact-checking approach in terms of predictive performance, its speed, as well as interpretability of the the predictive model. In particular, the crowd helps to provide explanations about the reputation of the news source to improve acceptance of the outcome of the model. The provided explanations about the reputation of the news source improved users' satisfaction and trust in model predictions. While the presented model improves transparency of the outcome, and helps speeding the fact-checking process, the focus of this study is not on leveraging the influence of community-oriented factors to impact response to misinformation. Instead, the model still focuses on identifying individual pieces of misinformation, and helping online users to be informed of the credibility or lack of credibility of different news sources.

2.3.5 Discussion: Consequences that arise from an individualistic focus on addressing misinformation

This section argues that neither a focus on individual pieces of misinformation (either based on attributes of content itself or based on attributes of the source of content) nor a focus on individual users will be sufficient in addressing the issues associated with misinformation. To do so, it first highlights a series of two implicit *assumptions* that our analysis reveals in approaches that focus on individual content or users. In practice, we argue, these assumptions rarely hold. Second, it points out *blind spots* that arise from a focus on individual content (either based on the content itself or its source) or on individual users. These blind spots occur in part because this individualistic focus makes crucial aspects of the issues

around misinformation either less readily apparent or entirely out of scope.

2.3.6 Assumptions

2.3.6.1 Humans are rational actors persuaded by additional or corrective information.

Most individual-focused approaches are predicated on the assumption that individuals are rational actors who engage in logical reasoning about the content view. Many such approaches provide additional, corrective information or alternative stories. For example, TwitterTrails (Metaxas et al., 2015, p. 71) "does not answer directly the question of a story's validity, [but] it provides information that a critically thinking person can use." Implicitly, a person with more information will make the correct decisions. Many other interventions similarly provide additional information to help the user be more informed (**?**Kirchner and Reuter, 2020; Chan et al., 2017; Lewandowsky et al., 2012).

However, much of an individual's decision-making stems from sources other than pure rationality. Self-perceived rationality of the content and its source (Kow et al., 2019), community narratives (Sloman et al., 2018), motivated reasoning (Kunda, 1990; Kahan, 2012), familiarity bias (Swire et al., 2017b) and other factors can have a greater influence than purely providing more information. For example, people may even know a piece of content is untrue and still spread of anyway for reasons such as signaling their social identity (Buchanan, 2020). In such cases, providing additional or corrective information is less likely to influence an individual's decision about sharing misinformation.

Furthermore, Lazer et al. (2018) discuss how tools that attempt to provide corrective information, such as fact-checkers, can actually reinforce the false information they seek to correct. This downside occurs via a two-step process. First, fact-checkers simply increase the familiarity with claims that are false, contributing to familiarity bias. Second, familiarity bias makes people more likely to accept information that is familiar to them (Schwarz et al., 2007; Begg et al., 1992; Swire et al., 2017b; Thorson, 2015; Hasher et al., 1977). This way, familiarity bias further contributes to the spread of misinformation and its impacts (Lazer et al., 2018). However, approaches such as fact-checkers that hinge on identifying individual

pieces of untrue content, not only fall short in addressing the broader effects of familiarity bias, but also contribute to the strength of this phenomenon. As a result, these approaches are less likely to influence people's response to the spread of misinformation and not sufficient in mitigating its impacts (Lazer et al., 2017; Garrett and Weeks, 2013).

2.3.6.2 Each individual encounter with misinformation occurs in isolation.

The individualistic approaches we reviewed make it difficult to account for the role of prior beliefs and the social contexts in which individuals encounter misinformation. However, as discussed throughout Section **??**, an individual's prior beliefs (Nickerson, 1998; Moravec et al., 2018), as well as the social context in which misinformation is encountered, can drive an individual's response to misinformation.

For example, if corrective or additional information contradict an individuals' prior beliefs, it is less likely to change the person's opinion about the original misinformation (Lazer et al., 2017). Indeed, corrective or additional information might even fuel "backfire effect" (Anderson et al., 1980; Garrett and Weeks, 2013; Nyhan and Reifler, 2010). This effect occurs because many people do not assess information objectively (Nickerson, 1998; Moravec et al., 2018). Instead, they process information to confirm to their preexisting beliefs, a phenomena known as'motivated reasoning" (Flynn et al., 2017; Kunda, 1990; Kahne and Bowyer, 2017). For example, if the corrective information is perceived as an identity threat, people can become more defensive and bring different reasons to counter the corrective information (Kunda, 1990; Kahan, 2012; Lodge and Taber, 2013). This phenomenon can in turn strengthen their beliefs in the original false information.

Additionally, many people trust information that is shared by their network (Hermida et al., 2012) and often accept such information at face value (Geeng et al., 2020). In such cases, corrective or additional information is less likely to change people's opinion about the original misinformation (Lazer et al., 2017).

Furthermore, social contexts in which people encounter misinformation influence people's perceptions of what that community accepts and what that community condemns (Piccolo et al., 2020). Such perceptions play a role in the assessment of information and the way people respond to misinformation. However, the individualistic approaches that we

reviewed implicitly assume individuals identify and respond to misinformation only based on misinformation itself.

2.3.7 Blind spots

2.3.7.1 The social context(s) in which people encounter misinformation.

People's response to different content is not only a function of the content itself, or even individuals' knowledge about the content (Lazer et al., 2017; Nickerson, 1998; Moravec et al., 2018). Rather, people's response to misinformation is influenced by different factors, including the social context in which they encounter misinformation and the ways others within a community engage with and respond to misinformation (Colliander, 2019; Koo et al., 2021; Sunstein and Vermeule, 2009; Phadke et al., 2021). For example, due to a desire to gain social approval, people may consider how others respond to an issue, such as in comments (Geeng et al., 2020), when deciding on how to respond to it themselves (Piccolo et al., 2020).

As a result, even if identifying all the false or misleading arguments were possible, and even if it was possible to combat misinformation by influencing individual responses to misinformation, it is not sufficient to address misinformation only based on misinformation itself.

2.3.7.2 Statements that are factually true but misleading.

Approaches at content level only focus on and evaluate the truth value of individual claims. However, not all misleading claims are factually incorrect. In many cases, the actors state their argument based on factually correct pieces of content, but reshape the true statements and fit the pieces together to push a misleading argument (Starbird et al., 2019).

For example, actors with anti-vaccine beliefs might argue vaccines include certain chemical ingredients, e.g.: "Thimerosal is a mercury-based preservative used in vaccines. mercury is a known neurotoxin." Both of these statements are factually correct. However, the actors who spread the content, purposefully omit the fact that medical research suggests that the portion of these chemical ingredients in vaccines is safe for the human body (reu, 2020). For

another example, anti-immigrant groups often argue that immigrants take away jobs from American workers (Borjas, 2016), thus claiming that immigrants are a threat to Americans. However, they do not mention that immigrants create new jobs by forming new businesses, paying taxes, and contributing to the productivity of businesses in the U.S. (Sherman et al., 2019).

Indeed, for these misleading arguments to be influential, they need to be stated based on a core of verifiable information (Bittman, 1985). Thus, approaches that solely focus on the truth or falsity of individual pieces of content simply are unable to recognize such instances of misleading arguments that are based on factually true statements.

2.3.7.3 The systemic nature of the impacts of misinformation.

The individualistic approaches draw attention away from the way an online community as a whole responds to misinformation. The response of a community to misinformation is influential on the way people perceive a community, and make inferences about what a community as a whole accepts and what the community disapproves of (Piccolo et al., 2020; Rimal and Real, 2003). These perceptions about a community influence how one thinks they are expected to behave within a certain community (Chung and Rimal, 2016). These behaviors include responses to misinformation.

However, a community is more than the aggregate of its members, and the response of a community to misinformation goes beyond an aggregation of the responses of its individual members to misinformation. Thus, the overall community response cannot be improved simply via changes in individual responses to specific pieces of misinformation. Rather, the response of a community to misinformation is also influenced by other properties of the community, including the pattern of interactions, norms of content selections and content sharing, and the way the community as a whole views various issues. Together, these properties and the perceptions they create form how a community as a whole responds to misinformation, and impact the response of its members to various issues.

The individualistic approaches draw attention away from the broader, systemic issues to which the spread of misinformation contributes. For example, the circulation of misinformation can cause long-term attitudinal and behavioral shifts (Zhu et al., 2012; Pluviano

et al., 2017). As another example, the spread of misinformation allows actors with political goals to influence and reshape social structures and the types of conversations that take place around certain issues, as well as the arguments that are heard more within a community (Starbird et al., 2019). Manipulating social structure for political goals can contribute to broader impacts of misinformation, including the rise in polarization (Bakshy et al., 2015; Starbird et al., 2019; Lazer et al., 2018; Stroud, 2010), amplifying political divisions among a society, and undermining faith in authorities and science (Hamilton et al., 2015; Lewandowsky and Oberauer, 2016). These broad, systemic impacts of misinformation can be more harmful than the spread of factually incorrect content (Anderson and Rainie, 2020). However, focusing primarily on individual pieces of content or users constrains our vision, both in terms of what misinformation can do to online communities and in terms of how we might most effectively respond.

2.3.8 Summary

To reiterate, as discussed throughout this section, current approaches of combating misinformation usually treat misinformation as individual pieces of content that need to be addressed. Given the definition of misinformation as "false or misleading information" (Lazer et al., 2018; Wardle et al., 2018; Starbird et al., 2019), which refers to information as pieces of content, this finding is perhaps unsurprising. However, in addition to the content of misinformation, there are other factors that drive an individual's response to misinformation, including their prior beliefs, as well as the social contexts in which they encounter misinformation.

In addition, while misinformation itself refers to false or misleading information, its impacts go beyond misleading people about individual factually untrue statements. Therefore, rather than completely abandon the existing approaches of combating misinformation that focus on individual pieces of misinformation, we should instead complement those approaches with interventions that address the broader, systemic nature of misinformation and its impacts.

2.4 Contributions and Future Work Directions

This chapter contributes to advancing our understanding of the scope of misinformation. It provides a comprehensive review of the factors that are involved in the phenomenon of misinformation, and acknowledges that misinformation is indeed broader than individual pieces of false and misleading content. In addition, it shows that the impacts of misinformation are similarly beyond misleading individuals about individual pieces of content. Specifically, it highlights the involvement of community oriented factors in the spread and impacts of misinformation on individuals, as well as the way they contribute to the broad community level effects of misinformation. This chapter also conducts a review of approaches that are designed to address misinformation and its impacts. Despite the various factors involved in the phenomenon of misinformation, however, it demonstrates that most prior approaches that are designed to study and address misinformation bound the scope of this phenomenon to individual pieces of false and misleading content. That is, they aim to address the impacts of false and misleading content at the individual level, and overlook the involvement of other factors, especially community oriented factors in this phenomenon. This individualistic focus on misinformation and overlooking the role of other factors involved results in broad impacts at the community level.

This chapter advocates for adopting an ecological approach to account for and address the broader scope of misinformation. An ecological approach enables us to go beyond current individualistic approaches, and account for the various factors involved in the information ecosystem that contribute to misinformation and its impacts. It enables us to consider the important role of community-oriented factors in the spread and impacts of misinformation at the community level, which have been largely neglected by existing individualistic approaches.'

Indeed, there is a wealth of different community-level processes that could be relevant. These processes include familiarity bias, social network structures, majority illusion, among many others (Kempe et al., 2003; Schwarz et al., 2007; Begg et al., 1992; Swire et al., 2017b; Thorson, 2015; Hasher et al., 1977; Huckfeldt et al., 1995; Hermida et al., 2012; Geeng et al., 2020; DiFonzo et al., 2013). The activities of community members either

directly or indirectly play a role in most of these mechanisms, thereby contribute to the impacts of misinformation at the community level. In particular, the response of community members to the world's events influences perceptions about what people in the community do, what they approve of, and how they expect other members to behave with respect to a certain event. Such perceptions are referred to as *perceived norms* (Rimal and Real, 2003; Rimal et al., 2005). Perceived norms in turn influence people's behaviors (Rimal and Real, 2003; Teunissen et al., 2012; Masur et al., 2021). As perceived norms significantly influence people's behavior in various contexts, in particular in the context of misinformation (Kim et al., 2020; Colliander, 2019), it is important to both protect them from the impacts of misinformation. At the same time, it is also important to investigate approaches to design around the influence of perceived norms in mitigating the impacts of misinformation at the community level. Thus, the next chapter examines how misinformation contributes to perceptions about social norms and explores the potential of online communities in mitigating the community-level impacts of misinformation.

Chapter 3

The Mechanisms by which Misinformation Impacts Perceptions of an Online Community's Norms

This chapter investigates the mechanisms by which individuals perceive the norms of an online community around conspiratorial content and related misinformation. In particular, motivated by the insights from prior work on norm perceptions (Matias, 2019; Colliander, 2019; Masur et al., 2021), it examines the role of the prevalence of false and misleading content, the community response to such content, and the community's established rules on norm perceptions. To do so, it employs an experimental approach, and using simulation-based studies, it manipulates and examines the effects of the aforementioned elements on norms perceptions. Moreover, this chapter investigates whether and how perceptions about a community's norms around misinformation might lead to broader impressions taken about a community. The remainder of this section describes the significant role of perceived norms within the contexts of misinformation and delves into the potential effects misinformation may have on this community-oriented aspect. Next, it describes the methods taken to examine the mechanisms of perceived norms, followed by a discussion of the results, and

broader implications thereof.

3.1 Introduction and Motivations

As discussed in Chapter 2, online communities play a significant role on how people perceive and interpret the events around them, thereby influencing their responses to surrounding misinformation about those events. For example, the way a community responds to conspiratorial content (e.g., sharing, liking, commenting) can significantly contribute to the influence that such content has (Begg et al., 1992; Swire et al., 2017b; Thorson, 2015; Hasher et al., 1977; Hermida et al., 2012; Geeng et al., 2020; Colliander, 2019), both at the individuals level and at the broader community level. At the individual level, prior work has shown how the activities of community members contribute to the exposure of other members to conspiratorial content (Anspach, 2017), thereby influencing their opinions about and their responses to such content (Colliander, 2019; Gimpel et al., 2021). However, less work has investigated the community level effects of members' responses to such content, such as perceived norms around conspiratorial content, increasing political polarization (Mackie, 1986), or mistrust in science and authorities (Bicchieri et al., 2021). In particular, the response of community members to an event influences perceptions about what people in the community do, what they approve of, and how they expect other members to behave with respect to a certain event. Such perceptions are referred to as *perceived norms* (Rimal and Real, 2003; Rimal et al., 2005). Perceived norms in turn influence people's behaviors (Rimal and Real, 2003; Teunissen et al., 2012; Masur et al., 2021). Indeed, prior work demonstrates the influence of perceived norms on people's behaviors in a variety of domains, from drinking habits (Teunissen et al., 2012), to decision making about the disclosure of personal information online (Masur et al., 2021), to individual language use (Allison et al., 2019). Recent work has specifically acknowledged the role of perceived norms on the way people respond to misleading and conspiratorial content (Koo et al., 2021; Colliander, 2019).

As acknowledged by prior work (e.g., Matias, 2019; Findor et al., 2021; Roozenbeek et al., 2020), perceived norms are important not only by influencing individual behaviors, but also through their broad impacts. Examples include the ways perceived norms contribute to

social tolerance of certain actions (e.g., Findor et al., 2021; Oyamot Jr et al., 2017), escalated behaviors (e.g., Matias, 2019), and impressions about popularity of (dis)trust in science (Constantinou et al., 2021; Imhoff and Bruder, 2014; Roozenbeek et al., 2020), among others. Each of these impacts can have important consequences. For instance, perceptions of social tolerance impact not only the way people respond to individual actions, but also the impressions that people take about a community's opinions as a whole around different topics (e.g., homosexual marriage Oyamot Jr et al., 2017; Wei, 2018), (e.g., mandatory vaccination Findor et al., 2021). In addition, perceived norms can impact expectations around escalated behaviors, where people form exaggerated perceptions about a community's norms and consider escalated behaviors as normative as well (Sutton and Douglas, 2022; Matias, 2019). These perceptions in turn can influence expectations about how a community responds to such attitudes. Thus, it is important to investigate the ways by which perceived norms are formed and to explore strategies to design around them.

However, relatively less is known about the mechanisms by which norms around conspiracy theories and related misinformation are perceived in online communities. In other contexts, prior work has explored various mechanisms to manipulate the perceived norms of a community. Examples include the prevalence of content exhibiting certain behaviors (Bicchieri, 2005; Cialdini et al., 1991; Cialdini and Trost, 1998; Masur et al., 2021), the response of community members to those behaviors (e.g., via supportive or opposing comments) (Colliander, 2019; Koo et al., 2021; Gimpel et al., 2021), and the expectations established by the community (e.g., explicit rules on social media platforms) (Cialdini and Goldstein, 2004; Matias, 2019). With one exception (Masur et al., 2021), these prior studies employed these mechanisms to manipulate social norms (e.g., using confederates) and explored the effects of those norms on individuals' behaviors, but they did not directly measure whether these mechanisms actually influenced participants' *perceptions* of those norms. Thus, it is not clear whether the effects of these mechanisms on an individual's behaviors occur through influencing perceived norms. In addition, it is also not clear by which of these mechanisms a community's norms are most readily perceived in the context of conspiratorial content and related misinformation. Furthermore, little is known about how perceived norms around conspiratorial content might then lead to other broad impressions

about online communities.

Therefore, this chapter investigates the effects of the aforementioned mechanisms on perceived norms around conspiratorial content and related misinformation in online settings. To do so, it presents experimental studies that focus on perceived norms around anti-vaccine conspiracies and related misinformation. These experiments investigate how the previously mentioned mechanisms (i.e., the prevalence of certain behaviors, the community's response to such behaviors, and the presence or absence of explicit community rules) can influence norm perception, as well as the broader impacts of such perceived norms.

The results from the conducted experiments provide insights about how the prevalence of misleading content, even when constituting a minority of posts in the feed, impacts different types of perceived norms and broader impressions about the community. However, the way other members of the community respond to such content can also strongly impact norm perceptions, thus in some cases mitigating the effects of misleading content on norm perceptions. In addition, while prior work suggests that making a community's rules explicit might impact individuals' behaviors (e.g., Matias, 2019), none of the conducted experiments support the effects of this intervention in addressing the perceptions about a community's norms. Moreover, perceived norms contribute to broader expectations about the community as a whole. Specifically, people's perceptions of norms regarding a certain conspiracy theory guide their expectations around how the community would respond to other conspiracy theories not directly observed in the community, as well as expectations around escalated behaviors both within and outside the community. The chapter offers a nuanced discussions of the implications that the findings suggest for the design, policies, and governing online communities.

3.2 Related Work on Social Norms and Their Roles on Response to Conspiratorial Content and Related Misinformation

This section reviews prior work on the definition and impacts of social norms. It illustrates two different types of social norms, known as *collective norms* and *perceived norms* (Rimal and Real, 2003; Rimal et al., 2005; Rimal and Lapinski, 2015), and highlights the importance of perceived norms compared to collective norms on people's behavior. It illustrates how perceived norms influence people's behaviors in a variety of subjects (e.g., Teunissen et al., 2012; Lindström et al., 2018; Masur et al., 2021; Bursztyn et al., 2020; Cialdini and Trost, 1998; Rimal and Real, 2003; Rimal and Lapinski, 2015). Next, it outlines the role that perceived norms play on individuals' response to conspiratorial and related misinformation (Koo et al., 2021; Colliander, 2019; Gimpel et al., 2021).

3.2.1 Social Norms: A key Mechanism that Impacts Individuals' Behavior

Social norms are defined as "rules or standards that are understood by members of a group, and guide and/or constrain social behavior without the force of law" (Cialdini and Trost, 1998, p. 152). Social norms influence people's opinions and behaviors in a variety of subjects (e.g., Teunissen et al., 2012; Lindström et al., 2018; Masur et al., 2021; Bursztyn et al., 2020; Rimal and Real, 2003). For example, norms of underage or excessive drinking influence people's opinions about acceptable drinking habits and can in turn influence their own consumption patterns (Olds and Thombs, 2001; Oostveen et al., 1996). What matters here is not the actual prevalence of underage or excessive drinking, referred to as *collective* norms. Instead, it is *perceived* norms, perceptions of others' behaviors (Rimal and Real, 2003; Cialdini et al., 1990; Lapinski and Rimal, 2005; Chung and Rimal, 2016), that influence an individual's opinions and behaviors. People's perceptions of norms (i.e., perceived norms) can diverge from the actual norms (i.e., collective norms) (Lapinski and Rimal, 2005). For example, students often held exaggerated perceptions about the prevalence of drinking

(Perkins and Berkowitz, 1986). Such perceptions influence some individuals to rationalize their own excessive drinking habits (Lapinski and Rimal, 2005).

The stronger influences of perceived norms compared to collective norms occurs for at least three reasons (Rimal and Real, 2003; Oostveen et al., 1996; Olds and Thombs, 2001). First, prior studies show that many people are often poor at estimating collective norms (Perkins and Berkowitz, 1986; McAlaney et al., 2011). For example, Perkins and Berkowitz (1986) show students tend to perceive alcohol consumption as more prevalent than it actually is. These perceived norms influence people to believe their alcohol consumption patterns are within the prevailing norms of their community (Oostveen et al., 1996; Olds and Thombs, 2001). Second, people cannot make inferences about others' actual beliefs independent of their own perceptions of others' beliefs and attitudes (Rimal and Real, 2003). That is, individuals' knowledge about others' beliefs and attitudes is influenced by their perceptions and interpretations of their social interactions with others, which is inherently subjective (Lapinski and Rimal, 2005). Third, prior work demonstrates that even if people are informed about others' actual views, still their behaviors are influenced more by their perceptions of others' behaviors compared to others' actual behaviors (Rimal and Real, 2003).

There are three types of perceived norms, which refer to people's perceptions of what others do (i.e., descriptive norms), what others approve of and what they condemn (i.e., injunctive norms), and how an individual thinks they are expected to behave (i.e., subjective norms) (Rimal and Real, 2003; Cialdini and Trost, 1998; Cialdini et al., 1991; Chung and Rimal, 2016). All three types of perceived norms can influence people's behaviors via different mechanisms (Rimal and Real, 2003; Smith et al., 2007; Park and Smith, 2007). For example, observing a behavior repeatedly within a community leads people to perceive the behavior as part of the community's norms (i.e., descriptive norms) (e.g., O'Gorman and Garry, 1976; Schroeder and Prentice, 1998), and a desire for social conformity increases the likelihood of people exhibiting that behavior themselves (Rimal and Real, 2003; Rimal et al., 2005). As another example, when people perceive a certain behavior as approved by a community (i.e., injunctive norms), they are more likely to approve of the behavior and adopt it themselves (Cialdini et al., 1991; Lapinski and Rimal, 2005). Similarly,

people's behavior is influenced by their perceptions of others' expectations (i.e., subjective norms), due to their desire to avoid risking interpersonal harmony by going against others' expectations (Chung and Rimal, 2016).

In online communities, perceived norms can similarly influence people's impressions about popular opinions and behaviors, and consequently impact their opinions and behavior (e.g., Matias, 2019; Masur et al., 2021; Rashidi et al., 2020; Dym and Fiesler, 2018). For example, people's opinion about acceptable language in online discussion (Matias, 2019; Hovy and Yang, 2021), privacy concerns (Masur et al., 2021; Rashidi et al., 2020; Dym and Fiesler, 2018), reliability of different pieces of news (Colliander, 2019; Gimpel et al., 2021; Park et al., 2012) and views of different events (Wojcieszak et al., 2020; Phadke et al., 2021) are shaped based on what they perceived as prevalent among (i.e., descriptive), and approved of (i.e., injunctive), by others within their community.

While numerous studies acknowledges the influence of perceived norms online, the mechanisms by which norms have their influences online may have some differences with those in offline settings. For example, anonymous online settings (Deutsch and Gerard, 1955; Levy, 1960), lack of nonverbal cues (Bargh et al., 2004), among others, can impact the ways by which people perceive and are influenced by social norms. Therefore, it is important to explore the mechanism by which norms are perceived and impact online communities.

3.2.2 The Role of Perceived Norms on Individuals' Response to Conspiratorial Content and Related Misinformation

Perceived norms particularly influence people's behavior around conspiratorial content and related misinformation (Andı and Akesson, 2020; Park et al., 2012; Wojcieszak et al., 2020; Colliander, 2019; Koo et al., 2021; Gimpel et al., 2021; Sloman et al., 2018). Specifically, perceived norms play a role on the types of content that people share within their community (Olson and Zanna, 1979; Atkin, 1985; Andı and Akesson, 2020; Park et al., 2012; Wojcieszak et al., 2020), and the way they identify misinformation and respond to it (Colliander, 2019; Koo et al., 2021; Gimpel et al., 2021). For example, Colliander (2019) demonstrates that many people's perceptions of news credibility are influenced by the responses of others. That is, if others point out the news is fake, they are more likely to consider it as fake news

as well, and less likely to share it (Colliander, 2019). In an experimental study, Andı and Akesson (2020) design an intervention based on descriptive norms, wherein participants are shown a message that suggests most people think twice before sharing news. They show that displaying this normative information can make people less likely to share fake articles and illustrate this influence occurs due to norm conformity to descriptive norms. In another experiment, however, Gimpel et al. (2021) show that injunctive norms most strongly influence people's response to fake news compared to descriptive norms. Specifically, they show that highlighting reporting fake news as a socially desired behavior using an injunctive norm intervention leads to higher reporting rates for fake news, while descriptive norms do not have such an effect. Koo et al. (2021), instead of manipulating perceived norms and investigating their effects, directly ask participants about perceptions of norms of correcting misinformation. Specifically, they ask, "If a typical American has posted information that was made-up, how likely is it that they will correct it?". Their analysis reveals that perceived norms around self-correct influence individuals to self-correct themselves.

The impacts of perceived norms are particularly significant when people try to make sense of conspiracy theories (Xiao et al., 2021; Phadke et al., 2021; Cookson et al., 2021). Indeed, conspiracy theories are often associated with incomplete, uncertain data (Xiao et al., 2021; Phadke et al., 2021). To deal with such uncertainty, many people tend to look to others' behavior for guidance about appropriate behaviors (Moravec et al., 2018; Walther et al., 2002; Smith et al., 2007). To deal with uncertainly involved in conspiracy theories, people similarly look at how others respond to such arguments as guidance (Xiao et al., 2021; Cookson et al., 2021). Given the role of perceived norms around conspiracy theories, it is important to explore the mechanisms by which these norms are perceived.

3.3 Mechanisms of Perceiving a Community's Norms in an Online Context

While a growing body of work acknowledges the role of perceived norms on individuals' response to misinformation (e.g., Colliander, 2019; Andı and Akesson, 2020; Gimpel et al., 2021; Koo et al., 2021), relatively less work has explored the mechanisms by which norms

around conspiracies and misinformation are perceived.

Studies in other contexts have provided insights into how individuals might perceive norms in offline and online settings. In offline settings, observing a behavior as common among others is a key source of information that influence individuals' behaviors around various topics (Bicchieri, 2005; Cialdini et al., 1991; Cialdini and Trost, 1998). This influence could potentially occur via influencing the perceived norms (Cialdini et al., 1991). Indeed, recent work has explored and acknowledged that observing common behaviors of others regarding privacy practices influences people's privacy settings in online settings (Baumer et al., 2017; Masur et al., 2021; Trepte and Reinecke, 2011). For example, if others within the community have private profiles, participants are more likely to have private profiles as well (Masur et al., 2021). Masur et al. (2021) show that this influence occurs through influencing perceived norms around privacy concerns. We hypothesize that this relationship between observing exhibitions of common behavior and perceived norms will similarly hold in the context of conspiratorial content and related misinformation. More precisely, we posit that the prevalence of posts that exhibit anti-vaccine behaviors in an online community results in a higher perception of norms around anti-vaccine behavior.

H1: Seeing a higher prevalence of posts with anti-vaccine content will result in higher participants' perception of norms about anti-vaccine beliefs and behaviors. This includes (H1.a) descriptive, (H1.b) injunctive, and (H1.c) subjective norms.

Another element that influences individuals' response to various issues is how other people respond to it (Colliander, 2019; Koo et al., 2021; Gimpel et al., 2021). For example, as discussed in Section 3.2.2, many people tend to read others' comments to a piece of content when deciding whether or not to share it, and how to respond to it themselves (Colliander, 2019). This influence becomes strong when the situation involves uncertainty (Smith et al., 2007; Moravec et al., 2018; Walther et al., 2002), which is particularly true regarding conspiracy theories and related misinformation. This work aims to investigate whether this influence of others' responses occurs via influencing perceived norms.

H2: When users respond to anti-vaccine content with support, participants' perceptions of norms about anti-vaccine beliefs and behaviors will increase, and when users respond to anti-vaccine content with opposition, participants' perceptions of such norms will decrease.

This includes (H2.a) descriptive, (H2.b) injunctive, and (H2.c) subjective norms.

Another source of information that informs people about community's norms can be through the expectations established by institutions (Cialdini and Goldstein, 2004), especially when institutions indicate the ways by which those expectations will be enforced (Reiter and Samuel, 1980; De Kort et al., 2008). For example, Matias (2019) demonstrates that providing normative information in the form of community rules decreases unruly and harassing conversations. Matias (2019) argues that these effects occur by influencing perceptions around what is normative in a community. In this study, we explore whether displaying a community's rules actually influences perceived norms, and will similarly hold in the context of anti-vaccine content.

H3: Seeing a community's established rules about misinformation will result in lower participants' norm perceptions of anti-vaccine beliefs and behaviors. This includes (H3.a) descriptive, (H3.b) injunctive, and (H3.c) subjective norms.

3.4 From Perceived Norms to Broader Perceptions about a Community in an Online Context

3.4.1 Perceived Norms and Social Tolerance

Perceived norms can lead people to tolerate events that they used to prohibit (Chong, 1994; Oyamot Jr et al., 2017). An individual may disapprove of a particular behavior, but if they perceive that the community thinks that behavior is normative they may be willing to tolerate it. For example, if LGBTQ+ behaviors are perceived as normative, individual people who do not approve of such identities may nonetheless be willing to tolerate them because of those norm perceptions (Findor et al., 2021).

Such social tolerance is distinct from injunctive norms, which refer to people's perceptions of what others approve of. In contrast, perceived social tolerance refers to what people believe others are willing to tolerate. These perceptions can be as important as perceived norms, as they can affect how people choose to behave, and how they react to the behaviors of others (Ford et al., 2001). For example, when disparaging humor (e.g., racist or

sexist humor) is observed as socially tolerated, regardless of whether or not this behavior is approved by others (i.e., injunctive norms), people high in hostile sexism or those high in racism may be more likely to exhibit such behaviors. In addition, observing these behaviors as socially tolerated may make other people less likely to speak up against such behaviors.

In online communities, perceptions of social tolerance can similarly affect people's behaviors (Findor et al., 2021; Anderson et al., 2018). The effects include both positives and negatives. For example, social tolerance of marginalized groups such as LGBTQ+ individuals can reduce prejudiced attitudes towards them on social media (Findor et al., 2021). On the other hand, social tolerance of toxic language online can increase instances of toxic conversations (Anderson et al., 2018; Matias, 2019). Social tolerance of conspiratorial content and misinformation can similarly have negative consequences. Examples include increasing the spread of conspiracy theories and expanding their reach and influencing perceptions around the popularity of such misleading theories (Kjeldahl and Hendricks, 2018; O'GORMAN, 1975; Prentice and Miller, 1996).

This study examines how different elements of an online community influence perceptions of social tolerance of anti-vaccine behaviors. It asks:

RQ1.a: What elements in an online community (i.e., prevalence of anti-vaccine content, the response of other users, and community rules) play a role on participants' perceptions of social tolerance of anti-vaccine behaviors in the community?

In addition, this study examines whether perceptions of different types of perceived norms around anti-vaccine behaviors influence perceptions around social tolerance of such behaviors in a community. It asks:

RQ1.b: Do participants' perceptions of a community's norms (i.e., descriptive norms, injunctive norms, subjective norms) online influence their perceptions around social tolerance of anti-vaccine behaviors in that community?

3.4.2 Perceived Norms and Escalated Behaviors

As discussed in Section 3.2.1, people's perceptions of social norms are often inaccurate and sometimes even deviated from the actual norms (i.e., collective norms) (Rimal and Real, 2003; Rimal et al., 2005). In particular, people might form exaggerated perceptions about a

community's norms, thereby consider escalated behaviors as normative as well (Rimal and Real, 2003; Rimal et al., 2005). For example, people sometimes develop exaggerated beliefs about others' alcohol consumption, thus consider their excessive alcohol consumption as normative based on their inaccurate perceptions of norms (Perkins and Berkowitz, 1986). Put differently, the (inaccurate) perception of excessive alcohol consumption as normative contributes to an escalation of an individual's alcohol consumption.

In online contexts, exaggerated perceptions of social norms can similarly lead people to consider escalated behaviors as normative. For example, people might hold exaggerated perceptions of norms of toxic language (Matias, 2019). As a result, they may consider escalated toxic conversations as normative as well, while such escalated conversations are neither pervasive (i.e., descriptive norms) nor approved by the majority (i.e., injunctive norms), or expected by the majority (i.e., subjective norms). Perceiving escalated behaviors of toxic language as normative can lead people to tolerate, approve of, or even engage in toxic conversations themselves (Matias, 2019; Cheng et al., 2017). In addition, such perceptions can discourage many people from engaging in discussions to avoid potential consequences (Munn, 2020; Clapp et al., 2016).

Recent work suggests that people might similarly hold exaggerated perceptions of norms around conspiracy theories, and overestimate the extent to which others endorse conspiracy theories (Cookson et al., 2021). Do such exaggerated perceptions of norms around conspiracy theories lead to expectations around escalated behaviors around this topic? This study examines whether and how perceptions of norms around misinformation can guide expectations around escalated behaviors in the context of anti-vaccine behaviors. It asks:

RQ2.a: Do participants' perceptions of a community's norms, and/or their perceptions of social tolerance of anti-vaccine behaviors, guide their expectations of escalated behaviors within the community?

3.4.3 Perceived Norms and Perceived Escalated Behaviors Beyond a Community

In addition to the influence of norms and social tolerance of expectations around how behaviors might escalate within a community, these perceptions may similarly have influence

on expectations about escalated behavior outside a community, such as in offline settings. Thus, this chapter also investigates this link in more detail, and asks the following questions:

RQ2.b: Do participants' perceptions of a community's norms, and/or their perceptions of social tolerance of anti-vaccine behaviors, guide their expectations of escalated behaviors outside the community?

3.4.4 Perceived Norms and Perceived Beliefs in other Conspiracy Theories

Above, we suggest that expectations about escalated behaviors online may differ from those offline. In some ways, offline escalation represents the participant inferring other behaviors that they have not directly observed as being likely of community members. Similarly, participants may make inferences about community members' beliefs based not on direct observation of content about those beliefs, but based on perceived norms around other content.

For example, norms of online communities can play an important role on the way community members perceive and respond to conspiracies and related misinformation (Phadke et al., 2021; Sunstein and Vermeule, 2009; Cookson et al., 2021). Prior work has shown that belief in one conspiracy is predictive of belief in other conspiracy theories (Goertzel, 1994; Xiao et al., 2021; Scheffer et al., 2022). Scheffer et al. (2022) argue that the tendency for dichotomous or black-and-white thinking as one of the reasons that can explain why people who believe in one conspiracy theory can be prone to believe in other conspiracy theories.

Thus, if members in a community are perceived as believing in one conspiracy theory, then that perception may increase the expectation that community members similarly believe other conspiracy theories not directly related. These perceptions can then attract people with pre-existing beliefs in such theories, leading to a homogeneous, like-minded group. The interplay between people's prior beliefs and the interactions with other people with similar beliefs in such theories can contribute to reinforcing and making such beliefs stronger, and can potentially lead to polarization of beliefs (Scheffer et al., 2022; Rodriguez et al., 2016).

Thus, this work first asks:

RQ3.a: What elements of an online community (i.e., prevalence of anti-vaccine content, the response of other users, and community's rules) influence participants' expectations around a community's beliefs in other conspiracy theories?

This work then asks whether different types of norm perceptions and/or social tolerance of anti-vaccine behaviors guide people's expectations around a community's beliefs in other conspiracy theories, not directly observed in the community?

RQ3.b: Do participants' perceptions of a community's a) descriptive, b) injunctive, and c) subjective norms and/or social tolerance of anti-vaccine behaviors influence their expectations of the community's beliefs in other conspiracy theories, not directly observed in the community?

Figure 3.1 depicts the hypotheses and the research questions discussed in this section.

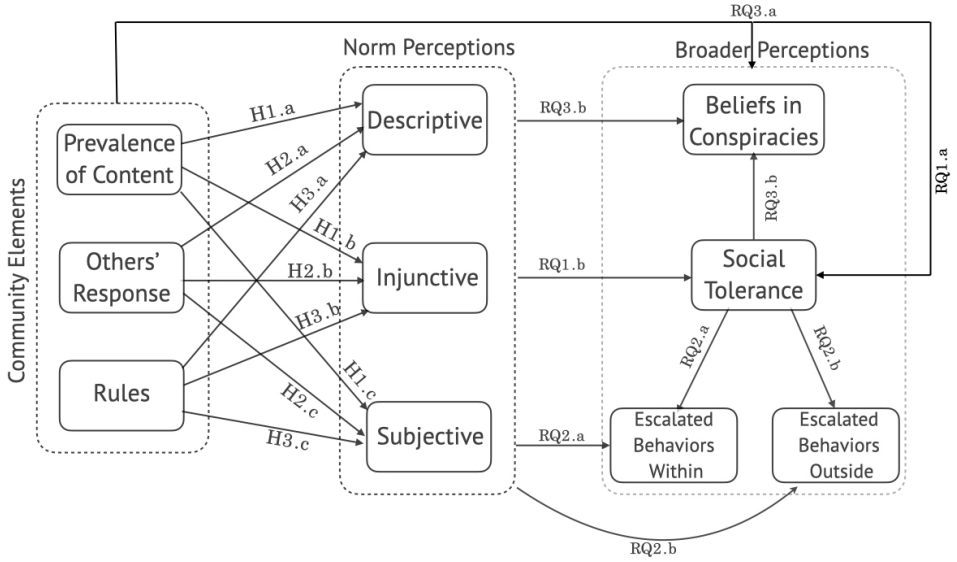


Figure 3.1: We hypothesize that the prevalence of conspiratorial content (e.g., anti-vaccine content), the response of community members to such content, and the community's established rules impact different types of perceived norms (i.e., descriptive, injunctive, and subjective), as well as broader perceptions about the community (e.g., social tolerance, escalated behaviors, and beliefs in other conspiracy theories).

3.5 Methods and Experiments

To explore the aforementioned research questions and hypothesis, two complementary between-subject studies have been designed. The first study is conducted using a screenshot study

to provide initial insights around whether and how the chosen mechanisms will have any impacts on norm perceptions, and broader perceptions about the community, including social tolerance, and expectations of how behaviors might escalate in this context. Next, the second study is conducted using a social media simulation approach to both address some of the limitations of the first study in terms of experimental realism and the length of the experiment, and also to build on the insights from the first study. To ensure concision, this dissertation centers primarily on the more comprehensive study, the simulation based study (i.e., study 2), given its alignment with the findings of screenshot study (i.e., study 1). Please refer to Aghajari et al. (2023c) for a detailed description of study 1. This section explains the methods for the simulation based experiment, including the procedure, experimental design, measures to operationalize the key variables, and the data analysis methods. The key findings are highlighted in the discussion section, followed by the implications thereof.

3.6 Simulation-based Experiment

This section presents the study's methods, including the employed social media site for the experiment (i.e., the EatSnap.Love site), procedure, participant recruitment, the experimental design, measures to operationalize our key variables, and the data analysis methods, followed by the results and discussion.

3.6.1 EatSnap.Love - Social Media Platform for Experimental Social Media Studies

EatSnap.Love is an online web application, which enhances the design and conduction of experimental research related to social media (DiFranzo et al., 2018). This platform was originally created as an "Instagram for food", and is similar to other popular social networking sites, such as Facebook and Instagram (DiFranzo et al., 2018). The EatSnap.Love platform provides the basic functionality of other popular social media platforms. Specifically, the participants are able to sign up, create an account, (optionally) fill in their profile information, browse the site and the activities of other users (e.g., check their posts, and

their comments to others' posts), and check the profile of other users on the site. In addition, the participants can interact with other users by liking and commenting under their posts. This setting on EatSnap.Love achieves a high degree of experimental realism, as reported in prior work (Masur et al., 2021; Taylor et al., 2019).

In addition to enhancing experimental realism, this experimental setting in EatSnap.Love achieves a high degree of experimental control. In particular, within the EatSnap.Love platform, researchers can simulate social interactions of online users. To do so, all the posts, actors, and interactions on the platform are controlled by the research team. Specifically, the other users that each participant sees and interacts with are all designed bots. This way, in each experimental condition, every participant experiences the same social media experience. Yet, to make the experiment look realistic, the posts are displayed dynamically, and the replies to them also appear dynamically based on a pre-defined schedule. In addition, the programmed bots [1] respond to the participants' posts after the participants share a post, again based on a pre-defined schedule. These responses are generic and defined by the authors of this paper (e.g., "nice", "love it!"). However, future research can design conversational-AI that reply back to the particpants' posts based on the content or their posts, and further improve the experimental realism of this study.

In summary, on EatSnap.Love, researchers have the full experimental control to design various social media scenarios and test their research questions, while achieving a high degree of experimental realism. (Please check more detail on the implementations of this platform on this Github repository). This simulation-based approach is employed to investigate our research questions based on a realistic social media experience.

3.6.2 Procedure

The participants were told they were going to help with beta testing a new social network site, called EatSnap.Love, for one day. They were instructed that they would be asked to share their experience with the platform designers once the experiment was over. After consenting to participate, participants completed a pre-survey with demographics. Next,

[1] The bots that respond to the participant's posts are not chosen from the actors who are part of the stimuli to avoid introducing any potential, unknown effects.

they were randomly assigned into one of the experimental conditions, and given onboarding instructions about the use of the EatSnap.Love website. Participants were asked to log onto the site for a minimum of five minutes at least twice during the experiment and to post content at least once. Participants were sent an email reminder every eight hours about the requirements of the study. At the end of the study, participants were sent a link to a post-study survey, and their account was deactivated.

Last, the participants were debriefed about the actual goal of the study. Specifically, given that the study involved deceptions about how the other actors on the site were real people, we also informed our participants that all the actors on the site were actually programmed bots designed to post content and reply to other actors' posts based on a pre-defined program. We described that the purpose of this deception was to ensure that the study was realistic, and resembled an actual social media experiment. This study was approved by the IRB 1734046 at Lehigh University.

3.6.3 Recruitment and Participants

3204 participants were recruited from Amazon Mechanical Turk, from October 2021 to Jun 2022. Participants were compensated for their time as follows: 50 cents for attending pre-survey, $3 for completing the activities on the site, and $2 for completing the post-survey. Of 3204 participants who submitted the pre-survey, 1770 created an account on EatSnap.Love website. Among them, 1579 participants attempted the post-survey. This way, of the 3204 number of participants recruited for the study, 1579 participants completed all parts of the study. The attrition rate was therefore 49.28%.

After removing the responses where participants did not finish all the steps of the experiment, and the participants who did not pass a simple attention check test, the data of 1096 participants was used for our analysis. Given the small effect size observed in the screenshot study, the same priori power analysis is used for this simulation based study. According to this priori power analysis, we needed a sample size of 1093 to detect a small effect ($f = 0.15$) with 0.95 power. The participants reported a mean age of 37.12 (SD = 10.79). Sixty four percent of the participants (n = 706) were female, thirty three percent were male (n = 365), less than one percent were non-binary (n = 3), less than one per-

cent of the participants (n = 4) preferred not to disclose their gender, and less than one percent of the participants (n = 5) preferred to self describe their gender. The majority of our sample was white (75.15%), followed by Black or African American (9.23%), Hispanic, Latino, or Spanish (6.69 %), Asian (5.79%), all other races and ethnicity were less than 1%. The distribution of education among participants was as follows: less than one percent high school incomplete or less, 13.86% high school graduate, 38.60% some college degree, 28.92% four year college degree or bachelor's degree, 4.11% some post graduate school but not graduate degree, and 14.87% postgraduate or professional degree. Neither gender ($\chi^2 = 6.94, p = 0.803$, nor age ($F(1288, 11) = 0.97, p = 0.465$), race ($\chi^2$=90.19, p=0.38), and education (χ^2= 63.11, p=0.21) differed by experimental condition.

3.6.4 Experimental Design

This study follows a 3 x 2 x 2 between-subjects experimental design. The prevalence of anti-vaccine content are set to 5%, 30%, or 60% of the posts in the experiment depending on the experimental condition. The community response to anti-vaccine content is manipulated to either support or oppose the anti-vaccine posts. The community's established rules are present or absent depending on the experimental condition as well. The experimental setting is designed using a simulation-based approach. Specifically, this study is conducted using a social media site, EatSnap.Love, described in Section 3.6.1.

All the posts, actors, and interactions on the platform are controlled by the research team. In particular, the content of this experiment (e.g., posts, comments) are crafted by the lab members at Social Design Lab, at Lehigh University. The control posts are related to every day life (without being political), and the anti-vaccine content is gathered from several anti-vaccine communities online, as well as from general feel on social media. Once all the posts are created, the research team created comments for the posts. Each post receives 3 to 5 replies, where they are displayed dynamically during the experiment to make the experiment looks live to our participants.

Figure 3.2 shows the examples of the anti-vaccine posts designed for this study, in addition to different types of community response to anti-vaccine posts (either supporting the content or opposing it), and the screenshot of the community's established rules, that

is displayed on the participants' news feed during the experiment.

3.6.5 Measures

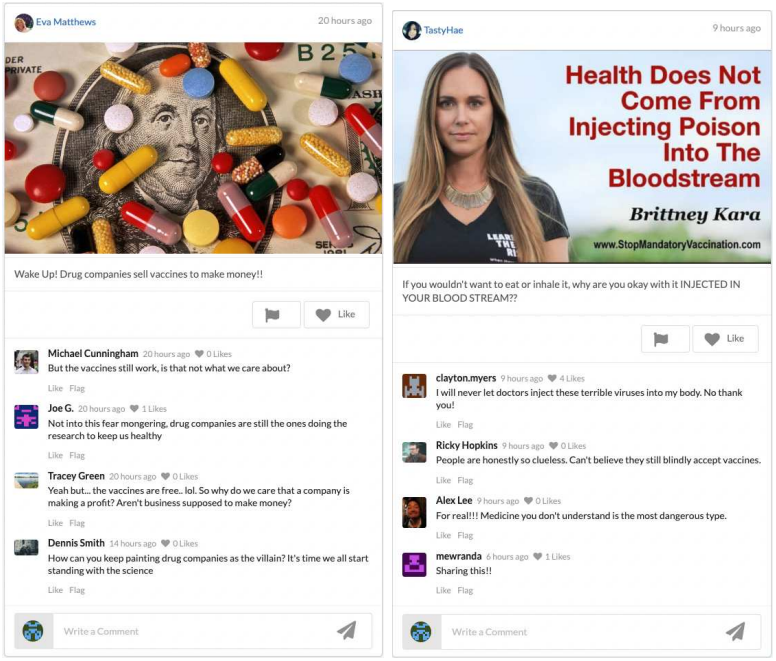
3.6.5.1 Perception of Norms

To measure participants' perceptions of norms, the scale developed by Park and Smith (2007) was adopted and modified to fit the focus of this study. This scale includes four items for descriptive norms (e.g., "The majority of people on EatSnap.Love post anti-vaccine content"), four items for injunctive norms (e.g., "The majority of people on EatSnap.Love approve of posting anti-vaccine posts."), and four items for subjective norms (e.g., "The majority of people on EatSnap.Love expect others to share anti-vaccine posts."). All of these items are measured using a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree)[2] (Park and Smith, 2007). This scale demonstrated a high degree of reliability for perceived descriptive norms ($\alpha = 0.97, M = 4.76, SD = 0.121$), injunctive norms ($\alpha = 0.98, M = 4.66, SD = 0.08$), and subjective norms ($\alpha = 0.91, M = 3.39, SD = 0.66$). Since this study aims to investigate the mechanisms of perceiving each type of perceived norms, different types of perceived norms are considered as distinct factors in the analysis.

3.6.5.2 Perceptions of Social Tolerance

To measure perceptions of social tolerance of anti-vaccine behaviors, a two-part scale is developed. The first part includes items identical to the scale used in the screenshot study ($\alpha = 0.85, M = 4.77, SD = 0.11$). The second part displays a post with anti-vaccine content and asks participants' perceptions about how EatSnap.Love community would tolerate the post (e.g., "The majority of people on EatSnap.Love are willing to tolerate this post in their feed.") ($\alpha = 0.88, M = 4.74, SD = 0.10$). Since these items were correlated strongly (r between 0.68 to 0.76), we continued our analyses with a single factor to present social tolerance of anti-vaccine behaviors $\alpha = 0.91, M = 4.75, SD = 0.15$)).

[2]Across all measures that pertain to vaccine attitudes, higher numbers indicate greater anti-vaccine norms.



((a)) ((b))

((c))

Figure 3.2: Figures (a) and (b) show two examples of posts with anti-vaccine content. Figure (a) shows examples of community members' responses to anti-vaccine content with opposition, and figure (b) shows examples of responses to anti-vaccine content with support. Figure (c) shows a screenshot of the community's established rules, that is displayed on the participants' news feed during the experiment.

3.6.5.3 Expectations of Escalated Behaviors

To separate escalation within the community from escalation outside the community, we showed the participants the posts in Figure 3.3. To capture expectations around escalated behaviors within and outside the EatSnap.Love community, we developed two distinct measures, as follows.

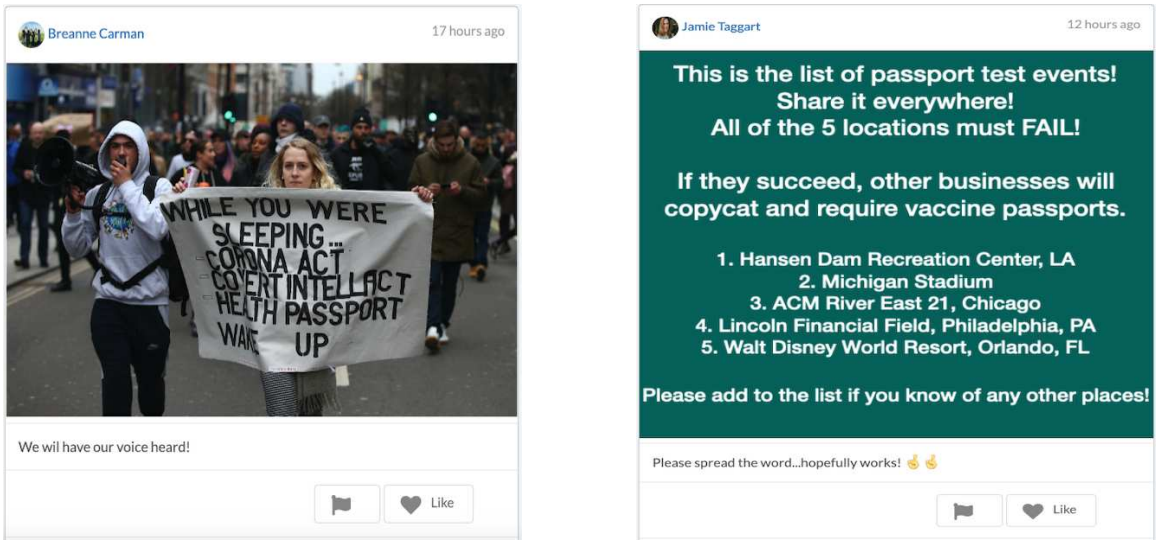


Figure 3.3: Examples of posts that show escalated behaviors regarding anti-vaccine behaviors (i.e., protesting to fight against vaccinations, and spreading a message to support an anti-vaccine movement).

3.6.5.4 Expectations of Escalated Behaviors Within the Community

Participants were asked about the likelihood of escalated behavior online (e.g., sharing posts that encourage the spread of an anti-vaccine movement).A six-item scale is used to measure this factor (e.g. "The majority of people on EatSnap.Love would share this post.", "The majority of people on EatSnap.Love would like this post.") ($\alpha = 0.93, M = 4.36, SD = 0.083$).

3.6.5.5 Expectations of Escalated Behaviors Outside the Community

Participants were asked about the likelihood that EatSnap.love community members take escalated behaviors outside their community (e.g., in physical world) to support anti-vaccine activities. To do so, a six-item scale was developed (e.g., "The majority of people on

EatSnap.Love are likely to take actions encouraged by this post.", "The majority of people on EatSnap.Love are likely to support this cause with their time, their money, or other resources."). The responses are recorded based on a 7-point likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). The scale shows a high degree of reliability ($\alpha = 0.93, M = 4.11, SD = 0.12$)

The expectations about escalated behaviors beyond the community were highly correlated to perceptions of escalated behaviors online ($r > 0.80, p < 0.001$). However, given that this study aimed at examining whether any differences occur between the expectations about escalated behaviors within vs. outside a community, the analysis treats these expectations as distinct factors in the analysis.

3.6.5.6 Perceived Beliefs in Other Conspiracy Theories

The initial screenshot study shows that perceptions of a community's norms guide people's perceptions around the community's behavior regarding escalated behaviors that they have not directly observed in the community. This simiulation based study aims to examine whether perceived norms can similarly lead to inferring the community's beliefs regarding content not directly observed in the community. Specifically, it examines whether perceptions of norms around anti-vaccine content result in perceiving beliefs in other conspiracy theories. To do so, participants are shown two posts around two other conspiracy theories around climate change, and the U.S. 2020 presidential election. They are asked about their perceptions around how members of the EatSnap.Love community would respond to this post using a 6-item measure (e.g., "The majority of people on EatSnap.Love would share this post."). The reliability of responses to both conspiracies posts was high (i.e., $\alpha = 0.93$ for the conspiracy on climate change, $\alpha = 0.94$ for the conspiracy on the election). Given that the responses about both these conspiracies were strongly correlated (r¿0.80), and that we had little reason to interrogate separately perceptions about beliefs in different kinds of conspiracy theories, we continued our analysis using a single factor based on the responses. The reliability was high for the single factor scale as well ($\alpha = 0.92, M = 4.82, SD = 0.04$).

3.6.5.7 Satisfaction and Enjoyment

The participants' satisfaction about the use of EatSnap.Love was measured using a single-item (i.e., "How satisfied were you using EatSnap.Love?"). The scale was measured on a 7-point likert scale, ranging from 1 (extremely dissatisfied) to 7 (extremely satisfied) (M = 3.297, SD = 1.67). Their enjoyment with EatSnap.Love was measured using a self-developed scale with six items (e.g., "I found using EatSnap.Love entertaining"). The scale was measured on a 7-point likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree), and demonstrated a high degree of reliability (α= 0.90, M = 5.17 , SD =0.61).

3.6.5.8 Participants Perceptions of Vaccinations

Participants were asked about how they think of vaccines and vaccinations using a self-developed, 3-item scale (e.g., "I think vaccines are generally safe.", "I think vaccines are generally effective".) The responses are recorded based on a 7-point likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). The scale shows a high score of reliability ($\alpha = 0.89, M = 4.60, SD = 0.60$).

3.6.5.9 Participants' posts

The participants made a total of 1979 posts on the platform during the experiment. Only 6 of these posts were related to vaccines. Five posts had a pro-vaccine stance, and one post had an ant-vaccine stance. Given the very few vaccine-related posts, and since these posts were scattered across the conditions, we were not able to draw any conclusive conclusions about potential links between our independent variables, and people's posting behaviors related to vaccines.

3.6.6 Data Analysis

The effects related to our hypotheses and research questions were investigated using structural equation modeling (SEM) (Hox and Bechger, 1998; Tarka, 2018). This approach is chosen for two reasons. First, the dependent variables of this study are highly correlated, requiring a model that takes into account the variables' correlations. Second, based on

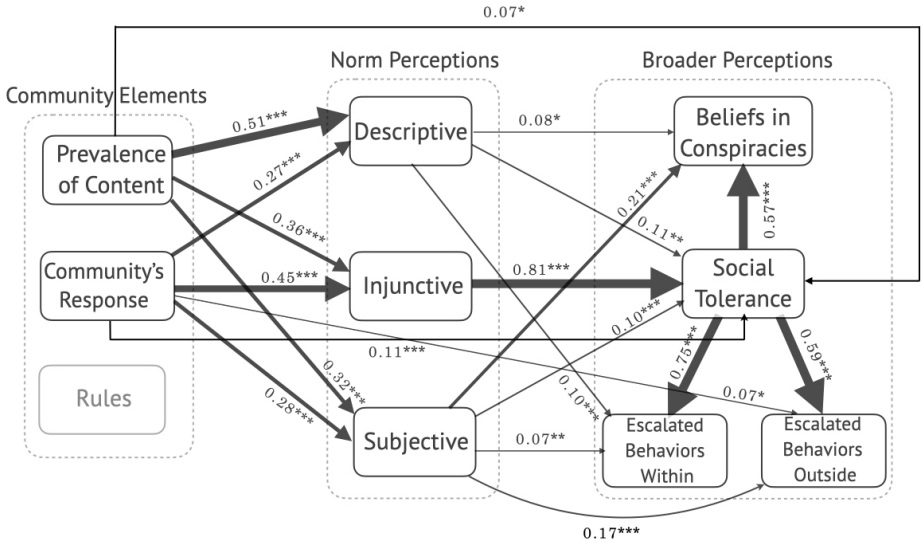


Figure 3.4: The SEM results show that, of the community elements tested, both the prevalence of content and the community's response to such content had a strong effect on different types of perceived norms, as well as perceptions around social tolerance of anti-vaccine behaviors. These norms perceptions and perceptions around social tolerance in turn lead to broader perceptions about the community, such as expectations about escalated behaviors both within and outside the community, and perceptions of beliefs in other conspiracy theories.

our hypotheses, the perceptions of social tolerance and escalated behaviors can depend on the latent factors of perceived norms. SEM helps to estimate the hypothesis of a hierarchical structure, required for modeling perceived social tolerance and perceived escalated behaviors.

The SEM analysis is conducted in `R` using `Lavaan` package (Rosseel, [n. d.]). Figure **??** depicts the results of this analysis. The multi-dimensional model fitted the data well ($\chi^2(367) = 3088.142, p < .001, CFI = 0.96, TLI = 0.95; RMSEA = 0.074, SRMR = 0.164$). Tukey's WSD (Wholly Significant Difference) post-hoc tests using simultaneous 95% confidence intervals is used for multiple comparisons between experimental groups (Pornprasertmanit et al., 2013). Specifically, we used the function `tukeySEM` in `R` to conduct Tukey's WSD post-hoc analysis (Maxwell et al., 2017). Demographics are included in the models as between-subjects covariates, and are reported when significant.

While SEM is chosen for the aforementioned reasons, the Lavaan package in R that is used to conduct the SEM analysis does not support analysis of interactions [3] (lav, 2022;

[3]While the Lavaan package allows for specifying the interaction terms in the SEM model, when fitting the model it does not include those interaction terms.

Rosseel, [n. d.]). Therefore, the potential interaction effects are examined using multivariate analysis of variance (MANOVA). Tukey's honestly significant difference (HSD) tests using simultaneous 95% confidence intervals are used for investigating the details of the interaction effects. Due to the hypothesis of a hierarchical structure between perceived norms and the broader perceptions about the community (i.e., perceptions of social tolerance, and expectations of escalated behaviors), the potential interaction effects of the manipulated variables are only investigated for the perceived norms, and they are only reported when significant.

3.6.7 Results

SEM is employed to test the causal relationships between the elements of the community, and perceptions of norms and social tolerance of anti-vaccine, as well as broader impressions about the community (i.e., expectations of escalated behaviors and beliefs in other conspiracy theories). Figure 3.4 depicts the results of the constructed SEM. The model fitted the data well ($\chi^2(717) = 4207.610, p < 0.001, CFI = 0.93, TLI = 0.92, RMSEA = 0.067, SRMR = 0.047$). This section uses the results of this SEM to test the proposed hypotheses and to investigate the study's research questions.

For the reasons mentioned in the data analysis Section 3.6.6, this study examines the potential interaction effects of the manipulated variables on the perceived norms using multivariate analysis of variance (MANOVA). In addition, the details of the interactions are examined using post-hoc Tukey HSD analysis, and reported when significant.

This section discusses the results based on the dependent variables (i.e., perceived norms, perceived social tolerance, expectations of escalated behaviors within and outside the community, and expectations of beliefs in other conspiracy theories not stated in the community).

3.6.7.1 Perceived Descriptive Norms

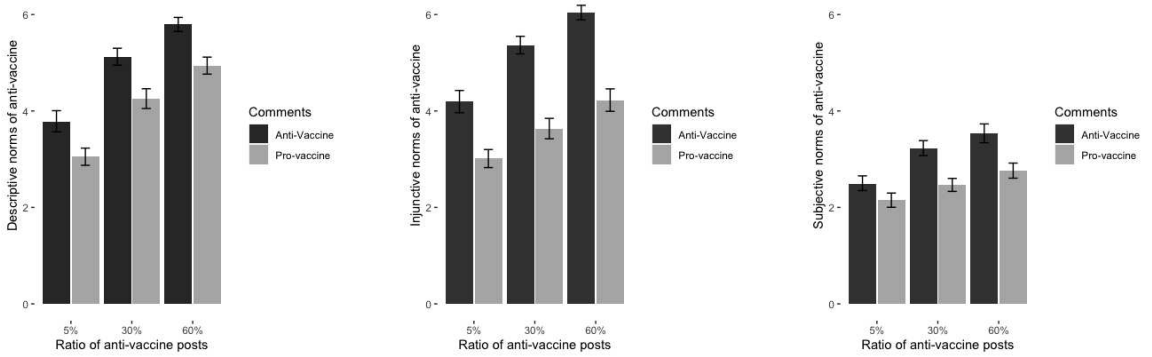
In line with our prediction in **H1.a**, prevalence of content with certain behaviors (i.e., anti-vaccine content) had a large, and significant effect on perceived descriptive norms ($B = 0.51, p < 0.001$). Tukey's WSD post-hoc tests demonstrate a significant difference

across all levels of posts with anti-vaccine posts (i.e., 5%, 30%, and 60%). However, the difference in perceived descriptive norms was greater between the conditions with 5% vs. 30% anti-vaccine posts on the feed, compared to the conditions with 30% vs. 60% anti-vaccine posts. These results again suggest that a small occurrence of content that stated certain behaviors (e.g., anti-vaccine content) can provide enough signals for participants to perceive the attitudes as normative. Others' response to anti-vaccine content also greatly and significantly influence people's perceptions of descriptive norms ($B = 0.27, p < 0.001$) (supporting **H2.a**). Therefore, the results show that the effects of observing posts with anti-vaccine content on perceived descriptive norms was relatively larger compared to the effects of others' reactions to such content (B=0.51 vs. B=0.27, p¡0.001). However, given the noticeable role of community members' responses, they can still mitigate the impacts of the spread of misleading content on perceived descriptive norms. Figure 3.5(a) depicts the impacts of the prevalence of content on perceived descriptive norms, and how these impacts are mitigated by the response of community members respond to such content.

On the other hand, displaying a community's rules that forbid posting of misinformation did not have a statistically significant effect on perceived descriptive norms ($B = 0.03, p = 0.24$) (i.e., there is not sufficient evidence to support **H3.a**). The community's rules did not have any interaction effects with the other manipulated variables of the experiment either.

3.6.7.2 Perceived Injunctive Norms

Prevalence of posts with anti-vaccine content ($B = 0.36, p < 0.001$) greatly and significantly influenced people's perceptions of injunctive norms (supporting **H1.b**). Based on Tukey's WSD post-hoc analysis for SEM, there was a significant difference across all levels of prevalence of posts with anti-vaccine content in the feed. However, the difference of perceived injunctive norms between conditions with 5% and 30% anti-vaccine posts was greater than the difference between the conditions with 30% and 60% anti-vaccine posts. In addition, the results shows that the way other users respond to anti-vaccine content had a large and significant effect nor perceived injunctive norms as well ($B = 0.45, p < 0.001$) (supporting **H2.b**). In fact, the effects of others' responses on perceived injunctive norms was relatively larger than the effects of posts with anti-vaccine content ($B = 0.45$ vs. $B = 0.36, p < 0.001$).



((a)) Descriptive Norms ((b)) Injunctive Norms ((c)) Subjective Norms

Figure 3.5: The effect of prevalence of anti-vaccine content, and the community members' response to anti-vaccine posts on perceived A) descriptive, B) injunctive, and C) subjective norms about anti-vaccine behaviors. Greater values indicate perceptions of anti-vaccine beliefs and behaviors as more normative. In contrast with the results of the screenshot study, community's response mitigates, but does not completely eliminate, the effect of prevalence on norm perception.

These results are in line with the definition of injunctive norms (See Figure 3.5(b)). In particular, injunctive norms refer to perceptions around what members of a community approve of (Discussed in more detail in Section 3.2.1). Here, the response of other users to anti-vaccine posts can be a proxy of whether or not the community as a whole would approve or disapprove of the anti-vaccine content.

The interaction effects of the prevalence of anti-vaccine content and the community's response to such content on perceived injunctive norms was small, yet statistically significant ($F(1089, 2) = 5.39, p < 0.01$). A post-hoc Tukey HSD test shows that when the community members respond to anti-vaccine content via opposing comments (i.e., comments with a pro-vaccine stance), they can mitigate the effects of such content on the perceived injunctive norms. For example, the results show that when the anti-vaccine content constitutes 60% of the posts on the feed, but the community members respond to such content with opposition in the comments, the participants on average reported lower perceived injunctive norms compared to when these posts are less prevalence (i.e., constitute 30% of the posts) yet supported by the community members in comments (M = 4.40 vs. M = 5.59, p ¡ 0.001).

The effect of displaying community's rules on perceived injunctive norms was neither large nor statistically significant (i.e., there is not sufficient evidence to support **H3.b**). The

community's rules did not have any interaction effects with the other manipulated variables of the experiment either.

3.6.7.3 Perceived Subjective Norms

Prevalence of posts with anti-vaccine content ($B = 0.32, p < 0.001$), as well as the way other members of the EatSnap.Love community responds to such content ($B = 0.28, p < 0.001$) greatly and significantly influenced people's perceptions of subjective norms (supporting **H1.c** and **H2.c**) (See Figure 3.5(c)). The interaction effects of the prevalence of anti-vaccine content and the response of the community to such content were small, yet statistically significant ($F(1089, 2) = 4.49, p < 0.05$). In particular, the results show that the community's response to anti-vaccine content, if opposing such content, can mitigate the effects of the prevalence of anti-vaccine content. For example, in the condition wherein 60% of the posts exhibit anti-vaccine content but the community oppose such content in the response, the participants on average reported lower perceived subjective norms compared to when 30% of the posts exhibited anti-vaccine content but the content was supported in the community's response (M = 2.76 vs. M = 3.23, p ¡ 0.001)

However, displaying community norms did not have a statistically significant effect on perceived subjective norms ($B = -0.009, p = 0.83$) (i.e., there is not sufficient evidence to support **H3.c**). The community's rules did not have any interaction effects with the other manipulated variables of the experiment either.

3.6.7.4 Perceived Social Tolerance

In response to **RQ1.a**, prevalence of posts with anti-vaccine content had a very small, yet significant effect on perceived social tolerance of anti-vaccine content (B = 0.07 p¡0.005). The effects of the response of other users to such content on social tolerance of anti-vaccine behavior were also significant, but relatively larger ($B = 0.11, p < 0.001$).

To investigate **RQ1.b**, the constructed SEM includes a regression model that explores the effects of different perceived norms on perceptions of social tolerance of anti-vaccine attitudes. Perceived injunctive norms most strongly and significantly predict social tolerance of anti-vaccine behaviors ($B = 0.81, p < 0.001$). Perceived descriptive ($B = 0.11, p < 0.001$)

and subjective norms ($B = 0.10, p < 0.001$) both had small, yet significant effects on perceived social tolerance as well.

3.6.7.5 Expectations of Escalated Behaviors Within the Community

In response to **RQ2.c**, people's perceptions of descriptive norms ($B = 0.10, p < 0.001$), subjective norms ($B = 0.07, p < 0.001$), as well as their perceptions of social tolerance ($B = 0.75, p < 0.001$) influenced expectations around escalated behaviors within the community. In particular, perceived social tolerance of anti-vaccine behaviors most strongly predicted expectations around escalated behaviors within the community.

None of the elements of the community (i.e., the prevalence of anti-vaccine posts, other responses of other users to such content, and the community's rules) directly and significantly predicted participants' expectations around escalated behaviors within the community.

3.6.7.6 Expectations of Escalated Behaviors Outside the Community

With respect to **RQ2.d**, perceived subjective norms ($B = 0.17, p < 0.001$) had a significant effect on perceptions of escalated behaviors offline. Perceptions of social tolerance, however, had a greater impact on expectations of escalated behaviors outside the community ($B = 0.59, p < 0.001$). As discussed in Section 3.6.7.5, perceived social tolerance most strongly influenced perceptions of escalated behaviors within the community as well.

Among the elements of the community, only the response of other users to anti-vaccine content directly and significantly influenced perceptions of escalated behaviors outside the community. However, the effect of this element was very small ($B = 0.07, p < 0.05$).

3.6.7.7 Expectations of Beliefs in other Conspiracy Theories

None of the manipulated elements of the community (i.e., the prevalence of content that exhibits certain behaviors, the response of other members to such content, and the community's established rules) had neither a large nor statistically significant effect on expectations around the community's beliefs in other conspiracy theories that were not directly stated

in the community (responding to **RQ3.a**). However, in response to **RQ3.b**, perceived descriptive norms ($B = 0.08, p < 0.05$), perceived subjective norms ($B = 0.21, p < 0.001$), and perceived social tolerance ($B = 0.57, p < 0.001$) of anti-vaccine behaviors significantly predicted perceptions of beliefs in other conspiracy theories. As repeatedly observed for other perceptions about the community, people's perceptions of social tolerance had the strongest effect on expectations around the community's beliefs in other conspiracy theories not directly stated in the community as well.

Participants' opinions on the safety and efficacy of vaccination had a small, yet significant effect on perceptions of descriptive norms. Specifically, participants who reported higher trust in vaccines' safety and efficacy on average reported slightly higher perceived descriptive norms of anti-vaccines ($M = 4.624$ vs. $M = 4.383$). However, this factor did not affect perceptions around any other types of norms, social tolerance, or escalated behaviors. These results may be partially because of the small variation in the participants' reported opinions on the safety and efficacy of vaccination ($M = 4.60, SD = 0.60$). Future work can clarify the potential effects of opinions about vaccinations on perceptions of norms around vaccines. In addition, none of the study's outcome variables varied based on participant demographics.

3.7 Discussion

Both the conducted experiments demonstrate that observing the prevalence of anti-vaccine content influences perceived norms around anti-vaccine beliefs and behaviors. In particular, different degrees of the prevalence of anti-vaccine content lead to different perceptions of norms around anti-vaccine behaviors. However, the relationship between the prevalence of content and perceptions of norms is unlikely to be linear. Specifically, the difference in perceived norms was larger between small (i.e., 5%) and moderate (i.e., 30%) prevalence of posts with anti-vaccine content in comparison with the difference between moderate (i.e., 30%) and large (i.e., 60%) ratio of posts with anti-vaccine content. This finding suggests that even when a minority of posts in the feed (i.e., 30%) exhibit a certain behavior, that behavior can be perceived as normal, both in terms of norms of what others do (i.e., descriptive norms) and what others approve of (i.e., injunctive norms). This result is also in line with

prior work on how people often perceive exaggerated norms in various contexts (Rimal and Real, 2003; Masur et al., 2021).

The results also demonstrate that the response of the community members to anti-vaccine content has a relatively greater effect than the effect of the content itself on perceptions around what the community approves of (i.e., injunctive norm perceptions) and what the community is willing to tolerate (i.e., social tolerance). These results connote the significant role that community members can play in mitigating the impacts of the prevalence of conspiratorial content on impressions about the community as a whole. However, while displaying a community's rules might influence individuals' behaviors (e.g., Matias, 2019), in our results, in the conducted studies the effects of this intervention on perceptions about a community's norms was *not* observed. The community's rules might appear to be established by the moderators or platform rather than by its members. Future work should investigate the impacts of other ways of communicating a community's rules. In addition, it is possible that the presence of rules around misinformation in our experiment has not been salient enough. In particular, given that the rules about misinformation are displayed in conjunction with other community's rules (e.g., rules around respectful interactions), their presence may not be as noticeable to our participants. Future work should investigate the impacts of other ways of communicating a community's rules.

In addition, the effects of the prevalence of certain content and the community members' response to such content are not limited to influencing perceived norms. For example, this study demonstrates that perceived norms of anti-vaccine behaviors impact expectations around escalated behaviors, both within and outside the community. Moreover, perceived norms around one conspiracy theory can guide expectations around conspiratorial thinking in the community. At least two possible explanations could account for this effect. First, many conspiracy theories are based on a broad distrust in science (orević et al., 2021; Aupers, 2012; Pasek, 2018; Douglas and Sutton, 2015). Perceiving one such anti-scientific conspiracy (here, opposition to vaccines) as popular in a community may lead to perceptions of belief in other conspiracy theories founded on a distrust of science. Second, conspiracy theories more generally involve certain psychological similarities. They tend to be fairly speculative, to involve significant complexity, to provide a sense of control over uncertainty, etc. (Douglas

et al., 2017). Thus, participants who perceived a community as believing in one conspiracy theory may view that belief as stemming from underlying cognitive proclivities that make community members more likely to believe in other conspiracy theories.

3.8 Broader Implications

The remainder of this section considers how this chapter's findings offer implications, not only in terms of designing platforms for online communities but also in terms of community policies and governance.

3.8.1 Implications for Designing Platforms for Online Communities

Although the prevalence of content influences perceptions about a community's norms, **the community's response** to such content impacts perceived norms as well. Indeed, the response of community members to such content can have an even more substantial impact on perceptions around what the community approves of and what it will tolerate. Thus, with their responses, community members can mitigate some of the impacts of misleading content on perceptions about the community. Therefore, the response of community members to misleading content not only can have individual level impacts (e.g., influence people's view of and responses to the immediate issues) (Colliander, 2019), it can also address some of the community level impacts of misleading content, such as its effects on perceptions about a community. These findings suggest the importance of foregrounding the way members of a community respond to different content, especially in the case of conspiratorial content and related misinformation.

Platform designers and academic researchers can thus consider design solutions that spotlight responses from the community members. For example, instead of relying only on highlighting the numbers of downvotes or question marks to misleading pieces of content (e.g., Bhuiyan et al., 2021a), future research could investigate the impacts of highlighting the authentic comments from the community members. Could such highlighting influence perceived norms? Future work can also examine designing interventions that encourage people to join this effort of addressing misleading content. Such interventions

might, for instance, ask a user whether there is information in a given post that the user believes should be corrected for other community members. Doing so via selective notifications could provide just-in-time responses to misinformation, not from an automated fact checker but from other human community members. Indeed, these community-oriented interventions should be designed in a way to ensure that the involvement of the community members is effective and will not result in disrupting the community and its interactions. Similar to unintended side effects of some content moderation practices (e.g., Eslami et al., 2019), community-oriented approaches can potentially result in deleterious effects as well. For example, if not provided with proper guidance, some people may attempt to correct misinformation using uncivil comments (Masullo and Kim, 2021). Not only are such involvements less favorable (Masullo and Kim, 2021) and less effective (Coe et al., 2014; Scheufele and Krause, 2019), they can even negatively impact perceptions of norms around toxic conversations. In addition, the proposed community-generated responses can potentially be effective in situations wherein there are agreed-upon facts (e.g., scientific topics, such as climate change, and vaccines). However, there exist cases where there is no clear objective truth (Yong, 2004; Detmer, 2003). For example, discussions around religious issues, or cultures, are not objective, but subjective and dependent on social constructs, which vary from one culture to another, and/or from one religion to another (Detmer, 2003). In addition, there are situations of uncertainty, for example just after a sudden, shocking event takes place (e.g., the Boston Marathon bombing), where there are no verifiable facts about the event, thus no one still knows the truth. In such cases, promoting community-centered may result in privileging certain points of view over others. Thus, future researchers should consider carefully such possible side effects when designing community-oriented interventions.

In addition, users could be prompted about the damages of misleading content on their community as a whole. Users could be told that the goal in responding to such content is not necessarily to correct the individual person who posted the misinformation but rather to ensure that others are not misled by it. In fact, such policies could be established explicitly in community rules or governance procedures. Put differently, just as researchers should consider community-level effects and interventions, so, too, should community members be encouraged to think about the effects of misinformation and their responses to it at the

community level.

3.8.2 Implications for Designing Policies and Governance for Online Communities

While prior work suggests that displaying **explicit community rules** can impact individuals' behaviors (e.g., Matias, 2019), our results did not support the direct effects of this intervention on perceptions about community norms. Indeed, these results do not imply that online communities should abandon explicitly displaying their rules. However, it suggests that displaying rules *may* not fully address the broader impacts of misleading content in terms of impressions about the behaviors and opinions of a community. Thus, online community leaders can expect explicit posting of rules to influence some behaviors, but if they want to shape perceptions about a community's norms, they will likely need to pursue other strategies.

The **prevalence of certain types of content** shared in a community, however, strongly impacts what people perceive as normative within a community. For example, when anti-vaccine posts are shared frequently, the behavior exhibited in the content can influence perceptions about what people of a community do and believe in (i.e., descriptive norms), what they approve of (i.e., injunctive norms), and what they expect others to do regarding vaccinations (i.e., subjective norms). In addition, the impacts of content are not limited to only influencing perceived norms regarding the immediate, observed topic (e.g., anti-vaccine views). The frequent sharing of content with certain behaviors impacts perceived norms, through which it contributes to the broader perceptions about the community as a whole (e.g., how the community as a whole tolerates anti-vaccine behavior, the community's beliefs in other conspiracy theories).

Therefore, the findings suggest that to address the broader impacts of conspiratorial topics and related misinformation, social media companies, as well as community moderators and administrators, still need to account for the role of misleading content. Put differently, addressing misleading content is not only required to address the spread of misleading content on individuals but also vital for addressing the broader impacts of misleading content at the community level (e.g., perceptions about normative behaviors within

an online community). Therefore, while we do need to move towards community-oriented aspects of misleading content, as suggested by prior work (Aghajari et al., 2023b), we should not abandon approaches that address individual pieces of content or the individual actors involved in its spread.

3.9 Contributions and Future Work Directions

In line with the research focus of this dissertation to understand the scope of misinformation beyond false and misleading content, this chapter contributes to expanding knowledge around the broad impacts of misinformation at the community level. Specifically, it demonstrates the significant role of misinformation in shaping perceptions about online communities. These perceptions include perceptions about a community norms, the types of behaviors that the community is likely to tolerate and engage in, as well as the broader community's beliefs about events not directly discussed in the community. In addition, this chapter contributes to understanding around the mechanisms by which these impacts occurs. In particular, it highlights the way false and misleading content, even if constitutes the minority of the posts, can significantly influence perceptions about online communities. Therefore, while it is important to move towards community-oriented aspects of misinformation and account for the community factors that contribute to this phenomenon, as advocated in Chapter 2, we should not discard approaches that address individual pieces of false and misleading content. Indeed, such individualistic interventions play an important role in mitigating not only the spread of misinformation, but also its impacts on the broader social context.

In addition, this chapter offers valuable insights on designing strategies that leverage the role of online communities in addressing the broad impacts of misinformation at the community level. Specifically, the findings of this chapter underscores the significance of community response in mitigating the broad effects of misinformation on shaping perceptions about a community. These results have important implications for designing platforms for online communities, as well as for designing community policies (Discussed in Section 3.8).

While this chapter provides important insights about the role of community response on the impacts of misinformation at the community level (e.g., perceptions about community norms, and the broader impressions taken about a community), it examines community response only as the expression of agreement and disagreement with an event (e.g., conspiratorial content). However, there are more nuances and complexities involved in the response of online communities to an event beyond simply expressing agreement or disagreement with a post through comments. For example, the way people express their stance towards different content shared in their community, discuss its various aspects, express their opinions about the events, and share their personal experiences related to those events all involve nuances and complexities that go beyond expressions of agreement and disagreement with a certain post. Accounting for these nuances in community response to an event allows us to better understand how a community views an event and responds to its various aspects. In this way, we may be able to gain a deeper understanding of how false and misleading content interplays with community response, and the way these elements together contribute to the impacts of misinformation at the community level, from its impacts to perceptions about a community, to contributing to the processes involved in meaning constructions about the world's events. Therefore, to account for the nuanced nature of community responses and the interplay between these responses and false or misleading content, the following chapter employs the concept of framing.

Chapter 4

Developing a Computational Technique to Enable Exploring Misinformation Manifestation as a Broader Phenomenon

4.1 Introduction and Motivations: Framing as a Conceptual Framework to Take an Ecological Approach to Misinformation

This section employs the concept of framing to examine misinformation as a broad societal phenomenon that can transcend any individual piece of content, and impacts individuals and society beyond misleading them about a particular piece of content. While there are a variety of societal concepts that may be relevant (e.g., social norms, network structure), this section explains how the concept of framing most readily corresponds to the goal of examining misinformation as a societal phenomenon.

In particular, this dissertation focuses on how misinformation plays a role in the way people assemble, organize, and respond to events, beyond impacting their response to in-

dividual pieces of content. This focus indeed aligns with what the concept of *framing* essentially is about. Put precisely, framing refers to the processes by which people construct their understandings of events around them, and respond to the circumstances surrounding them. This concept provides a framework to explore how the interplay between false and misleading content and the way communities respond to such content around an event together contribute to people's understanding of an event and their response to it (described in more detail in Chapter 5). In this way, framing allows us to move beyond considering misinformation as false and misleading content. Rather, it enables us to understand misinformation as the interaction among false content, community response, and the processes of meaning constructions (i.e., framing), whose impacts go beyond misleading individuals about a single piece of content.

For the focus described above, this work needs an approach to framing that positions it as a dynamic process, which evolves based on not only a statement but an array of content that people encounter, and generate themselves. While the concept of framing is studied in different fields — from political communication research (Froehlich and Rüdiger, 2006; Scheufele, 2000), to psychology (Tversky and Kahneman, 1985), to sociology (Goffman, 1974), to behavioral economics (Kahneman and Tversky, 1984) — this dissertation will take an approach to framing that is in line with the perspective from sociological studies (Gamson, 1989; Scheufele, 1999; Benford and Snow, 2000). These studies treat framing as a dynamic process by which people interpret and make sense of events around them and construct their understanding of those events. Within this sociological paradigm, processes of framing involve many aspects of interpretation and meaning making, including how people define problems, diagnose causes, make moral judgments related to those problems, and suggest remedies (Entman, 1993; Gamson and Modigliani, 1989). Understanding these functions gives insights into how people view an event and respond to its surrounding aspects, and may provide a framework to examine the role of false and misleading content and community response to such content in these processes.

To examine framing with this kind of dynamic, processual orientation, employing conventional methods, such as manual content analysis, presents several challenges. First, manual content analysis, due to its labor-intensive nature, limits our ability to study fram-

ing processes on a large scale. However, the framing processes illustrated here by definition occur across heterogeneous entities within a diverse media ecology. Thus, to understand these processes, we would need to account for various sources in the wide, diverse information ecosystem. In theory, conducting such an analysis through a close reading of numerous documents across multiple sources may seem achievable. However, in practice, it is very challenging and almost impossible to process such a large scale using close reading of the documents. In addition, focusing on a single document at a time via a close reading of content may limit our horizon to observe the hidden patterns across the documents in the corpus (Underwood, 2019). Second, a close reading of content by human researchers in the content analysis may bring in concerns related to prior knowledge and certain perspectives, which may lead to observing certain framing but not others. An approach that moves beyond what human researchers might provide different perspectives related to framing that could potentially remain unnoticed by human researchers.

One viable means of meeting these criteria is by using computational models. A computational model addresses the challenges with scalability concerns of manual coding, allows us to account for the wider, diverse information ecosystem, and can reach alternative perspectives that may not be noticed using a manual coding approach. However, no prior computational approach has been designed to analyze framing processes. Specifically, most computational techniques for analysis of framing are designed and developed to examine frames *per se,* (i.e., distinct entities that can be invoked in a given or across multiple pieces of content) (e.g., Card et al., 2016; Walter and Ophir, 2019; Naderi and Hirst, 2017; Morstatter et al., 2018). While there are studies that look at framing as a process of meaning constructions (Goffman, 1974; Gamson and Modigliani, 1989; Benford and Snow, 2000; Scheufele, 1999), these studies employ content analysis, and no prior approach has been developed to conduct such analysis computationally.

To address this gap, and to be able to examine framing processes as dynamic processes of meaning construction for the goal of this dissertation, this chapter explores different linguistic features to design computational models. It examines the efficacy of these models in helping researchers with analyzing the processes of framing at a large scale, across a heterogeneous media ecology. To clarify, the computational techniques examined in this chapter

do not intend to identify framing processes directly. Rather, they are intended to identify text-oriented intuitions that might be indicative of framing and can assist researchers when examining the processes of framing.

With this goal, the rest of this section describes the work presented in this chapter in the following way. It starts with clarifying the goal of these proposed models, i.e., finding an unsupervised technique that can help researchers with analyzing the processes involved in framing analysis. Then, it describes three models that could potentially be useful for this goal and explains the motivations behind their design. Specifically, it explains how each of these proposed models focuses on identifying patterns in language that might be related to framing and may provide insights into the evidence of framing. It trains the models on data in the context of the COVID-19 pandemic. To examine the utility of these models, this chapter provides an evaluation of the three proposed models in terms of their effectiveness in assisting researchers examine processes of framing across large corpus of documents.

The findings of this evaluation study provide valuable insights into the criteria utilized by researchers to determine the efficacy of computational models in supporting framing analysis. In addition, this human subject assessment identifies one of the models that are designed and developed in this work, and its affordances in capturing linguistic patterns indicative of framing language, to be the most effective in assisting researchers with their framing analysis.

4.2 Linguistic Attributes Relevant to Framing Language

This section outlines linguistic attributes that might pertain to framing. These attributes are then employed in designing the models in Section 4.3.

4.2.1 Word Choice

The very definition of framing highlights the importance of word choice. Entman (1993) suggests framing is often manifest through particular "keywords", and "stock phrases." Gamson and Modigliani (1989) also suggest the importance of "catchphrases" and "exemplars" as evidence of framing. In a computational approach, Baumer et al. (2015) suggest

the role of lexical features that capture the specific words used, as one of the most important indicators of framing. While these prior work focus on framing based on distinct entities (i.e., frames) that are either present or absent in documents, this proposed work hypothesizes that word choice could also be similarly significant in the language of framing (i.e., as a dynamic process of meaning constructions). In particular, word choice can help infer the events under discussion in a given corpus, as well as the issues highlighted around those events.

Furthermore, the word choice within a corpus can provide insights around how events and their associated issues are *labeled*. Such labeling is an important component of framing (Lau and Schlesinger, 2005). The way events are labeled in turn is an important component of how people make meaning of world events (i.e., framing) (Lau and Schlesinger, 2005). For example, in the case of the COVID-19 vaccines, this matter might be labeled in different ways. For instance, vaccines might be labeled as a societal right, or as a marketable commodity, among others. As a societal right, vaccination is often discussed as a basic human right that the government should provide to all citizens regardless of their socioeconomic status. With this labeling, people might discuss that vaccines benefit the entire society. Certain words, such as fundamental rights, societal rights, basic needs, universal distribution, and equality might be more likely to be observed. However, when labeling the same issue of vaccination a marketable commodity, vaccines are considered as a product, which its value, similar to any other goods or products, may be determined by supply and demand. In such discussions, certain words such as consumer choice, private insurance, healthcare investment, marketing strategies, etc. are more likely to be observed. Therefore, the choice of words can provide valuable insights into not only the events discussed but how they are labeled in a given corpus as well.

Most prior computational work on framing also focuses on word-based features (Walter and Ophir, 2019; Liu et al., 2019). In comparing different types of features, Baumer et al. (2015) found that lexical features (unigrams, bigrams, trigrams, etc.) were one of the most important indicators of language that human participants saw as related to framing. While Baumer et al. (2015) focus on frames as distinct entities that are either present or absent in documents, the work presented in this chapter posits that word choice could also be

similarly relevant to the language of framing, as defined in this work (i.e., as a dynamic process of meaning constructions).

4.2.2 Latent Themes (i.e., topics)

Beyond the choice of individual words, patterns of word co-occurrence within a corpus may also be indicative of framing. Specifically, examining the ways words co-occur together may inform us about the latent themes that are discussed in a given corpus. Word co-occurrence patterns representative of latent themes are often analyzed using topic modeling techniques (Blei, 2012; Roberts et al., 2014; Lucas et al., 2015; DiMaggio et al., 2013).

In fact, different computational models have utilized the co-occurrence of words to capture the latent themes (also known as latent topics) that exist within a corpus of documents (Blei, 2012; Roberts et al., 2014; Lucas et al., 2015). Such approaches have been used to determine the prevalence of frame elements in a given corpus (Walter and Ophir, 2019). For example, Walter and Ophir (2019) suggest that utilizing such themes can help in the process of conducting inductive framing analysis and identifying *framing packages* (Gamson and Modigliani, 1989). These packages are defined as organized structures of symbolic devices (e.g., metaphors, catchphrases, moral appeals) that offer interpretations and meanings for events and their surrounding issues. Specifically, Walter and Ophir (2019) employ community detection algorithms to cluster together the themes based on their co-occurrence over documents and consider each cluster of themes as representative of a framing package. This proposed work similarly hypothesizes that identifying latent themes based on words' co-occurrence may carry intuitions around interpretive packages that give meaning to an event (i.e., framing processes).

However, the work presented here remains agnostic as to whether topics or clusters of topics constitute a frame *per se*. Instead, this work posits that examining topics in a corpus might provide *evidence* that can help attend to interpretive packages (i.e., frames). To do so, instead of labeling latent topics as in the work presented by Walter and Ophir (2019), this work utilizes these topics to explore insights into framing processes, defined in Section 4.1).

4.2.3 Grammatical Relationships

While knowing which groups of words co-occur can be informative, framing may also be indicated by the relationships among those words. The grammatical structure of sentences may help indicate those relationships (Pan and Kosicki, 1993; Hallahan, 1999).

Indeed, few prior computational work demonstrates that grammatical structures are important indicators of frame evoking language (Baumer et al., 2015; Recasens et al., 2013). For example, Baumer et al. (2015) show that the grammatical relations in which words appear within a document, and their role in those relations are important features when inferring the prominent frames in the document. While Baumer et al. (2015) focus on identifying frames in a classificatory approach, the work presented in this paper posits that grammatical relationships may similarly help with analyzing framing processes in an exploratory approach.

This assumption is motivated by the perspective towards framing focused in this work, i.e., the functions of interpretation in which framing is involved (e.g., what counts as a problem, how the causes are diagnosed, and what remedies are suggested). Specifically, with this perspective towards framing, in addition to capturing *what* people discuss, captured via word choice and latent themes, it is also important to also explore *how* people discuss an event. Accounting for grammatical structures helps understand relationships between words, offering more insights into how an event is discussed, to facilitate the exploration of framing processes. For example, grammatical relationships can offer insights into which words describe the issues, what roles different entities play within documents, and help understand who are given the agency of the discussed issues and who are considered the victims.

4.3 Model Designs for Framing

4.3.1 The Latent Dirichlet Allocation Model (i.e., LDA)

4.3.1.1 Motivations

Prior work leverages different topic modeling techniques to infer dominate frames (e.g., Walter and Ophir, 2019; Ylä-Anttila et al., 2022). However, it remains unclear whether and how topic modeling can be employed to identify *framing*, as the processes by which people interpret and construct meaning around the world's events (Described in Section 4.1).

The proposed work in this chapter postulates that the information that probabilistic topic modeling approach, in particular the the latent Dirichlet allocation model, i.e., LDA, provides might similarly offer valuable insights into interpretative packages of framing processes. Specifically, the LDA model provides insights into the *word choice* and the *underlying themes*, which as discussed in Section 4.2, are potentially relevant to language of framing.

4.3.1.2 Model Description

The latent Dirichlet allocation model, also referred to as LDA [1], is used to identify latent themes in large corpus of documents (Blei et al., 2003; Blei, 2012). In this model, topics (i.e., themes), are captured based on latent probability distributions over all the words in the vocabulary of a given corpus. In this way, the LDA model allows for multiple memberships of words in various topics. Therefore, the same word can be interpreted differently (implicitly, by a human reader) depending on the context (i.e., the probabilities of other words in the topic) (Blei, 2012; DiMaggio et al., 2013; Walter and Ophir, 2019). Figure 4.1 demonstrates the plate diagram of this model (Blei, 2012).

[1] The details of this model are omitted in this work. Refer to Blei (2012) for greater details on this model.

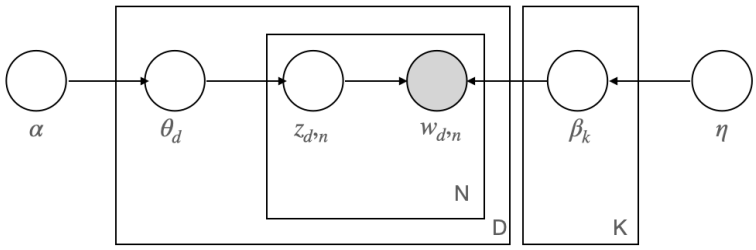


Figure 4.1: The graphic model for latent Dirichlet allocation, LDA (Blei, 2012). There are K topics $(\beta)_K$, wherein topic $(\beta)_k$is a distribution over vocabulary of all words in the corpus. θ_d is the topic probability for topic k in the document d. Finally, $z_{d,n}$ is the topic assignment for the nth word in the document d.

4.3.2 The Latent Dirichlet Allocation Grammatical-Relationship Model (i.e., LDA-GR)

4.3.2.1 Motivation

The latent topics, as defined and operationalized in prior topic modeling techniques (e.g., the LDA topic modeling (Blei et al., 2003)), are based on words' co-occurrence in a given corpus. While these topics account for the word choice, and the latent themes, they do not account for the grammatical relationships in which topic's terms occur.

To account for grammatical relationships in which topic terms occur, this work designs and develops an extension of the LDA model, named the Latent Dirichlet Allocation-Grammatical Relationship model, i.e., LDA-GR, described below.

4.3.2.2 Model Description

To capture grammatical relationships in which words appear, LDA-GR incorporates a one-to-one correspondence between each word tokens and grammatical relationships. To do so, it employs Stanford coreNLP to parse each document (De Marneffe et al., 2006; Manning et al., 2014). For each word token w in document d, a tuple of $< w, reln.role >$ is created, wherein $reln$ is the typed dependency of the word w in the document d, and $role$ specifies the role of the word w in the typed dependency of $reln$. A word could either play the role of the *governor* or the *dependent* in a dependency relationship. The governor, referred to as "gov" in the coreNLP parser and in our work, is the head of a dependency, and

the dependent, referred to as "dep", is the word that is governed or controlled by the governor. For example, in the sentence "Science defeated COVID-19.", there is an *nsubj* (i.e., nominative subject) typed dependency between the words "Science" and "defeated", nsubj(defeated, science), wherein defeated is the governor of this relation, and science is dependent.

As shown in Figure 4.2, this model's structure is almost identical to the LDA model (Blei et al., 2003). Similar to the LDA model, the topic allocations are based on a generative probabilistic model of the corpus of documents. The only difference of LDA-GR with LDA is how an input token is a tuple of $< w, reln >$ in LDA-GR as apposed to a word token w in the LDA model. As in LDA, the documents are presented as a random mixture over the latent topics. Consequently, each topic is characterized by a distribution over tuples of $< w, reln.role >$ described above.

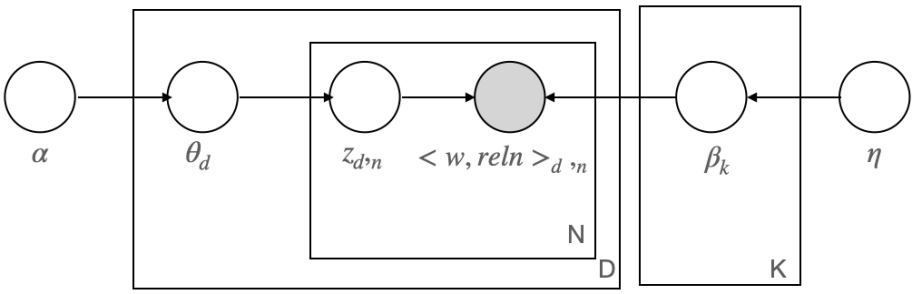


Figure 4.2: The plate diagram for the LDA-GR model.

Although LDA-GR is able to incorporate the grammatical relationships while generating the topics, it has some drawbacks as well. First, it enforces a strict one-to-one mapping between each individual word and each grammatical relationship. While this technique allows words to appear in different grammatical relationships with different probability, it may result in losing some grammatical relationships in which a word token occurs less frequently. These less frequently observed pairs of words and grammatical relationships might be still potentially important when looking for evidence of framing. Second, two sentences may be written with different syntactic structures and still convey similar semantic. For example, the two sentences "Science defeated COVID-19" and "COVID-19 is defeated by science" carry similar messages. However, they could (or could not) carry similar weights in

framing processes. LDA-GR does not allow to capture different grammatical relationships in which a term occur if it is not highly observed. Third, including all the potential grammatical relationships in which a given word could appear results in substantially expanding the number of dimensions within the topic distributions of the previous model (i.e., the β Matrix), while simultaneously increasing its sparsity. This sparsity could in turn impact the quality of topics (Blei and Lafferty, 2006; Popescul et al., 2013).

4.3.3 The Linked Latent Theta Role Model (i.e., LLTR)

4.3.3.1 Motivation

Consider the following two sentences. "Science defeated COVID-19.", "COVID-19 is defeated by science." These sentences convey semantically similar messages. However, in these examples, same words appear in different grammatical relationships (E.g., "Science" appears as a nominal subject and a direct object in these examples, respectively.). Despite these different grammatical relationships, it is possible that these same words with different grammatical relationships (or these sentences) play similar role in framing processes. LDA-GR, however, treats these appearance of the word "Science" as completely unrelated, repeated tokens. Alternatively, capturing different grammatical relationships in which topic terms occurs and might convey similar roles could indeed be beneficial in investigating framing evidence. In addition, with such an approach, we would not loose information about less frequent grammatical relationships in which a topic term might appear.

To capture different grammatical relationships in which topic terms occur, we propose the linked latent theta role model, i.e., LLTR. To do so, instead of capturing one-to-one link between words and grammatical relations, LLTR learns the distribution over grammatical relations in which topic terms occur. In addition, to help connect information about grammatical relationship to the topics' words, it also captures distributions related to the second argument by which topic terms manifest within a grammatical relationship.

The intuition to learn grammatical relationships in the form of distribution is motivated based on the design of the LDA model. Specifically, we posits that much in the same way that topics are captured based on words' probability distributions in LDA (Blei et al., 2003),

there exit some latent variables that can capture the probability distributions on a set of grammatical relationships in which topic words occur. Similarly, there exist other latent variables that capture the probability distributions pertaining to the second arguments by which each topic term appears in a grammatical relationship.

Intuitively, these latent variables, similar to the concept of theta roles in English (Aronow, 2016), are captured based on syntactic structures. However, these latent variables do not directly map to any syntactic nor semantic structures. Indeed, similar to how Blei et al. (2003, p. 996) make no epistemological claim regarding topics, the work presented in this paper designs theta role latent variables only to help capture the probability distributions on a set of grammatical relationships within each topic, without making any claims regarding a mapping between these theta role latent variables and any syntactic nor semantic constructs.

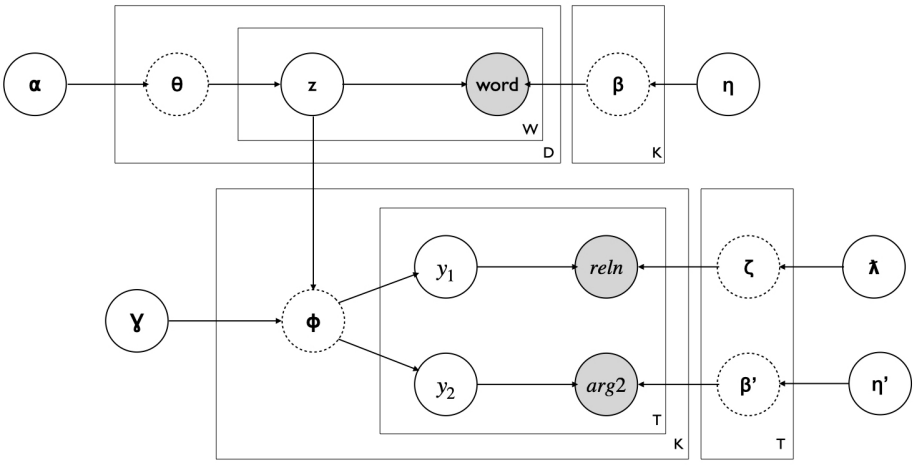


Figure 4.3: The plate diagram for the LLTR model.

4.3.3.2 Model Description

The LLTR model (Figure 4.3) simultaneously learns a two-component latent variable, theta role, for each topic. These components include (a) distributions over grammatical relationships, and (b) distributions over all the other arguments by which a word appears in grammatical relationships within the context of each topic. Analogous to the model by Ritter et al. (2010), known as LinkLDA, LLTR employs a linked latent variable to enable

learning associated pairs of grammatical relationships and arguments that appear in those grammatical relationship ($reln$, arg). To do so, rather than requiring both components to be generated from one possible pairs of $|T|$ multinomials (ζ_t, β'_t), this model allows these component (i.e., the grammatical relationships and the associated arguments) to be drawn from $|T|^2$ possible pairs. However, to increase the likelihood of states in which grammatical relationship *reln* and the arguments component *arg* drawn from theta role assignments that are related to each others, this model uses a sparse prior over the theta role distributions.

The following section defines the terminologies used in the LLTR model, followed by the model's generative story.

4.3.3.2.1 Definitions First, let there be K latent topics, where each topic β_k is a multinomial over the V words in the vocabulary (Blei et al., 2003), drawn from a Dirichlet parameterized by η (i.e., $\beta_k \sim Dir(\eta)$). For each topic, define T latent theta roles ϕ_t, where each theta role has a set of two multinomial ϕ_{1t} and ϕ_{2t}, corresponding to the two component of theta role (i.e., grammatical relationships $relns$, and argument components arg). Specifically, ϕ_{1t} is a multinomial distribution over the K latent topics for the first component of the theta role t, which is associated with R numbers of grammatical relationships $reln$. Each grammatical relationship is drawn from a Dirichlet distribution, parameterized by γ (i.e., $\phi_1 \sim Dir(\gamma)$). Within the ϕ_{1t} matrix, each row represents the topic distribution of the theta role t over the K latent topics. ϕ_{2t}, on the other hand, is a multinomial distribution over the K latent topics for the second component of the theta role t, which is associated with A numbers of argument components, arg. These arguments are drawn from a Dirichlet distribution parameterized by γ (i.e., $\phi_2 \sim Dir(\gamma)$). Within the ϕ_{2t} matrix, each row represents the topic distribution of the theta role t over the K latent topics. The model's generative story is as follows.

- For each document d_i:
  - Choose the length of the document $N \sim Poisson(\xi)$.
  - For each word w_j to w_N:
    - Draw a topic assignment z with corresponding multinomial distribution over

latent topics from the θ matrix, based on P(z—θ,d_i)

* Conditioned on the topic z, draw a theta role y_1 with corresponding distribution from the ϕ_1 matrix (i.e., $y_1 \sim Multinomial(\phi_1)$).
* Choose the grammatical relationship $reln$ from P(reln—y_1,ζ)
* Conditioned on the topic z, draw a theta role y_2 with corresponding distribution from the ϕ_2 matrix (i.e., $y_2 \sim Multinomial(\phi_2)$).
* Choose the argument component a from P(a—y_2,β')
* For the topic z drawn in the previous step, choose w_j from P(w_j—z,β)

4.3.3.2.2 The inference process The inference for the topic-word distribution, i.e., β, the model adopts the process in LDA (Blei et al., 2003), hence omitted for simplicity. For the inference on the probability distribution of theta role components, including ϕ_1 and ϕ_2 , collapsed Gibbs sampling is employed as follows (Griffiths and Steyvers, 2004; Geman and Geman, 1984). At each iteration, for each word w, provided topic z is selected, we sample theta role y_1 from the grammatical relationship component of theta role Φ_1 as follows.

$P(y_1|reln_i, \Phi_1, z) \propto P(reln_i|y_1) * P(y_1|\Phi_1, z)$

$P(reln_i|y_1) = \frac{Count_{reln_i,y_1}+\lambda}{\sum_{i=1}^{R} Count_{reln_i,y_1}+\lambda*R}$

$P(y_1|\Phi_1, z) = \frac{Count_{y_1,z}+\gamma}{\sum_{j=1}^{T} Count_{y_1 j,z}+\gamma*T}$

Here, $Count_{reln_i,y_1}$ is the count of all words whose grammatical relationship is $reln_i$ and the first argument of their theta role is y_1. $\sum_{j=1}^{R} C_{reln_i,y_1}$ is the same count, summing over all the R grammatical relationships $reln$. Specifically, the probability that the theta role y_1 is selected for the first component of theta role based on the Φ_1 distribution, provided that topic z is selected, is proportional to the probability of grammatical relationship $reln_i$ belong to this theta role y_1, times the contribution of this theta role y_1 for the assigned topic z.

Following the same approach, the theta role y_2 is sampled from the second argument component of theta role Φ_2.

$P(y_2|arg_i, \Phi_2, z) \propto P(arg_i|y_2) * P(y_2|\Phi_2, z)$ $P(arg_i|y_2) = \frac{Count_{arg_i,y_2}+\beta'}{\sum_{j=1}^{A} Count_{arg_j,y_2}+\beta'*A}$

$P(y_{2i}|\Phi_2, z) = \frac{Count_{y_{2i},z}+\gamma}{\sum_{j=1}^{T} Count_{y_{2j},z}+\gamma*T}$

Employing previous notation, $Count_{arg_i,y_2}$ is the count of all words that are observed in a grammatical relationship with the argument arg_i, and $\sum_{j=1}^{A} Count_{arg_j,z}$ is the same count, summing over all the possible A arguments within the corpus. Here, the probability of choosing y_2 as the second component of latent theta role for the component arg_j is proportional to the probability that the argument arg_j belongs to this theta role y_2, times the contribution of this theta role y_2 for the assigned topic z.

4.3.3.2.3 Extracting top documents After the inference process, a post hoc analysis is conducted to capture the distributions of theta role components, i.e., grammatical relationship and argument components, for each topic terms. Given these probabilities are within the context of each topic (i.e., we know a priori which topic we are calculating these distributions for), the probabilities of topic-theta role are going to be constant for all of these distributions, thus omitted for simplicity.

$P(reln_i|w_j, \Phi_1) \sim P(reln_i|w_j) * P(w_j|\Phi_1)$

$P(arg_i|w_j, \Phi_2) \sim P(arg_i|w_j) * P(w_j|\Phi_2)$

Next, for each topic term, example documents are extracted if the topic term co-occur with its top argument words in the specified grammatical relationship captured in the topic's theta role, provided that probability of the example document for the topic of exceeds a fixed threshold (e.g., 40%).

4.3.4 Data

4.3.4.1 Data Description

State health departments' news around COVID-19 is collected using the Scrapy scraping library (Zyte, 2021). To extract these documents, a set of hashtags introduced and employed in prior related work is used [2] (Wicke and Bolognesi, 2020). However, given that the departments of health do not use hashtags, the documents are collected using search terms relevant to these hashtags instead.

In accordance with this approach, a total of 5,462 news documents were collected. How-

[2] "#COVID19", "#coronavirus", "#ncov2019", "#2019ncov", "#nCoV", "#nCoV2019", "#2019nCoV", "#COVID19"

ever, due to resource constraints associated with processing such a large dataset, random subsampling was utilized prior to model execution. Thus, the models are run on a sample of size 3264 documents.

4.3.4.2 Data Preparation

Our baseline model, the Latent Dirichlet Allocation (LDA), as well as the linked latent theta role model (Described in Section 4.3.3), receive word tokens as input. Therefore, the documents are converted to word tokens, referred to as w.

The second mode (Described Section 4.3.2), referred to as LDA-GR Framing Model, requires tuples of $< w, reln.role >$ as input, wherein w is a word token, and $reln.role$ is the concatenation of the typed decency $reln$ in which the word plays and $role$ specifies whether the word is the governor or the dependent of that typed dependency relation. The word tokens are extracted using the same approach described above. To extract the grammatical relationships of words within a given document, we employ the Stanford coreNLP library (Manning et al., 2014).

4.4 The Evaluation Approach

While different metrics have been established to evaluate topics (Chang et al., 2009; Röder et al., 2015; Hosseiny Marani et al., 2022; Marani and Baumer, 2023), none of them are suitable for evaluating topics within the context of framing, due to two primary reasons. First, the presented models are not designed to identify framing directly, but rather to provide evidence of framing in a given corpus to assist researcher in conducting exploratory analysis of framing. Thus, it is relevant to evaluate the models by the human subjects for which the models are developed for (i.e., researchers) (van der Lee et al., 2021; Hoyle et al., 2021), and examine how they perceive the utility of the models' outcomes in providing useful information to assist them in exploring framing processes. Second, there is no objectively correct answers for inferring framing processes evidenced in text. Put differently, inferring framing is an inherently subjective task (e.g., Schön and Rein, 1994; Kuypers, 2010; Van Gorp, 2010), and different people might reach to different framing within an exact same corpus

of data. Similarly, different people might consider different linguistic patterns within the same corpus as evidence of framing. Therefore, instead of designing automatic evaluation metrics, it is important to explore human judgment of assess the efficacy of these models (Hoyle et al., 2021).

Therefore, this work designs and conducts a human assessment approach to examine perceived utility of each of the discussed models in facilitating exploratory analysis of framing. In line with Hoyle et al. (2021) arguments about the importance of "relevant human readers" to assess such computational models, we invited researchers familiar with the concept of framing to assess these models. To be clear, the goal of this evaluation is not to assess which of these models results in more consistent responses across the framings that different participants identify. Rather, the goal is to assess the perceived efficacy of these models in facilitating exploratory analysis of framing for whom the model is designed for. This approach was employed in a number of previous studies to evaluate models designs to study subjective concepts (Smith et al., 2017; Dinakar et al., 2015; Lee et al., 2017; Poursabzi-Sangdeh et al., 2016).

Our evaluation approach consists of a two-phase study, including survey study, and a follow up semi-structure interview study. These survey and interview studies are approved by the IRB numbers (2142707-3) and (2128233-3) at Lehigh university.

In the first part, **a survey study** is designed, in which participants engage with each of the models, and then assess the utility and the efficacy of the reviewed models in facilitating their analysis framing processes. This survey was initially designed to enable a quantitative comparison of the utility of the models as well. However, given the challenges of recruiting researchers at scale, who are the ones for whom these models are designed (i.e., "relevant human readers" (Hoyle et al., 2021)), and informed by a pilot test to attempt to use skilled freelancers with experiences relevant to framing as a participants [3], the quantitative comparison of the models deemed unfeasible for this work. Therefore, the designed survey is

[3]In a pilot study, we also explored recruiting skilled freelancers with expertise relevant to analyzing framing processes were also investigated. Specifically, freelancers with skills in areas such as technical writing, reading comprehension, and journalism (including journalistic writing) were recruited through the Upwork platform (Upwork, 2025). However, upon analyzing the responses from this pilot study, it became evident that the models, at least in their current stage and interface, remains too complex for non-researcher communities to effectively engage with or benefit from.

used to only enable engagement with the models, and to scaffold a follow up qualitative interview about potential efficacy of each of these models with the researchers who reviewed this models. The survey study is described in details in Section 4.4.2.

In the second part, the respondents are invited to enroll in a follow up **semi-structured interview.** This interview is designed to help understand a) how the patterns identified using each of these models resemble what researchers would consider when analyzing framing processes, b) how effective these patterns might be in assisting researchers to identify and analyze framing processes, and c) which of the models designed can be the most effective in helping researcher identify different functions by which framing performs (See section 4.1 for more details). The interview study is described in details in Section 4.4.3.

4.4.1 Participants:

To select participants for evaluation the models, this work follows the guidance by van der Lee et al. (2021) and Hoyle et al. (2021), and account for the "relevant" of participants to the model and the study tasks. Specifically, as argued by van der Lee et al. (2021) and Hoyle et al. (2021), this work acknowledges the importance of involving the intended user group in the evaluation process. In the case of this study, since the computational models described here are designed and developed to assist researchers with exploratory framing analysis, the "relevant human readers" are researchers who study framing. Thus, this study recruits researchers conducting framing analysis as the study participants.

To address the aforementioned challenge faced during data collection, the convenient sampling approach (Jager et al., 2017) is adopted. Specifically, researchers from the department of communication from the department of Journalism & Communication at Lehigh University, who conduct framing analysis using qualitative approach, as well as researchers who the author knew of with expertise in framing analysis, were invited to evaluate the developed computational models. Overall, five researchers, ranging from grad students, to post-doc researcher, to associate professors enrolled in this study.

4.4.2 Phase 1: Engaging with the models and assessing them in a survey study

Upon consenting to participate, participants are randomly assigned to one of six conditions. In each condition, participants work with and evaluate two of the aforementioned models. This design facilitates a within-subject study, with a relatively smaller sample size and a reduced study duration. The survey is structured as follows:

First, participants are asked to read a passage that defines framing as conceptualized in this study (see the italicized paragraph below). This definition highlights specific aspects of interpretation and meaning-making relevant to framing processes, including how people define issues, diagnose causes, make moral judgments related to those issues, and propose remedies. The following passage illustrates the definition of framing provided to participants in this survey study.

> **Framing definition** Framing is a dynamic and constantly evolving set of processes by which people construct their understanding of the world's events. The processes of framing help organize facts and information to give them meaning. Framing influences our understanding both of major world events, such as the COVID-19 pandemic, and of our personal daily experiences, such as a visit to the doctor's office. Framing involves different processes, including the following:
>
> - Determining what counts as **an issue**.
> - Diagnosing **the causes** of those issues.
> - Making **moral judgments**, such as about what is right and wrong, or about how people ought to behave.
> - Suggesting **potential remedies** to address the issues under consideration.

In addition, participants are provided with supplementary notes on framing before proceeding to the evaluation phase. These notes clarify the concept of framing, highlighting what framing entails and what it does not. For example, these notes explain how framing does not occur based on a single document. Rather, these framing processes unfold across

a complex, heterogeneous media ecology. Moreover, these notes illustrate how within an array of documents, there could exist one or more aspects of interpretation that framing is involved with.

Second, participants are informed that they will review results generated by two computational models. For each model, they are presented with three sample topics and are instructed to review these results for at least ten minutes. Following this review, participants respond to questions regarding framing based on the model outputs. This step is designed to engage participants actively in the experiment and encourage their involvement in analyzing framing processes. The model outputs are displayed using a HTML based interface. This interface is interactive, and participants are instructed on its functionalities. A screenshot of the interface for each model is provided in Figures 4.4 to 4.6.

Third, to learn what linguistic attributes are more in line with framing evidence, we asked participants to describe the most useful aspects of the results in helping them explore framing processes. In addition, to learn the limitations of the models, participants are asked if there is anything they would be interested to know when analyzing framing and the models fall short in providing them with relevant information.

Fourth, participants reported their assessment of the utility of the model's outcomes in terms of a) being useful and b) being easy to understand for each of the four aspects of framing processes. The ease of use is reversed coded to difficulty of use (Curran, 2016). Below shows the list of items asked.

The results **were useful** in helping me understand ..

- the issues discussed.
- the causes discussed.
- the moral judgments that were discussed.
- the potential remedies suggested for those issues discussed

The results **made it difficult** to understand...

- the issues discussed.

- the causes discussed.
- the moral judgments that were discussed.
- the potential remedies suggested for those issues discussed

Fifth, upon reviewing both the assigned models and evaluating them separately, participants were asked to compare and contrast the utility of the reviewed models. This assessment utilized items adopted from metrics introduced by Hart (2006); Brooke (1996). These items are measured using a 7 Likert scale, and are listed below.

In comparing the two models, which model ..

- was **more confusing** to work with.
- helped you **more successfully** complete the task.
- required **more effort** for you to understand.
- provided information that was **easier to understand**.
- showed information in a **needlessly complex** way.

Responses are captured using a Likert scale ranging from 1 to 7, labeled as strongly disagree to strongly agree.

To clarify, the responses to these survey items are not directly analyzed to determine which model is the most effective in assisting researchers with their analysis. Such an exploration is not feasible due to the limited number of research participants. Instead, these survey items were leveraged to engage participants with the results, and to prompt researchers to reflect on their experiences using each models. These responses were then reviewed with participants in a follow-up semi-structured interview (discussed in following section) to gain deeper insights into the aspects of each model that contributed to a more or less effective experience, in terms of examining framing evidence.

4.4.3 Phase 2: Qualitative Model Assessment in a Follow-up Semi-structured Interview Study

Participants who completed the aforementioned survey study were invited to this follow up semi-structured interview study. As discussed before, this follow up interview is designed to gather greater details about how the respondents worked with each of the models, how the patterns identified by each of these models might be indicative of framing language we designed an interview study, and what are the challenges in the current models that may be useful to addressed in future work. This semi-structured interview is designed as follows.

Participants responses are shared with them prior to the interview. During the interview, participants' responses are reviewed with them, and they are encouraged to share the way they have leveraged the model's results to attend to the questions on framing in the survey. Examples of questions asked in this semi-structured interview is listed below.

- Can you describe how these patterns relate to the issues that are discussed in the corpus?
- How did you use these patterns to understand the organization of facts and information discussed in the corpus of documents?
- What aspects of these patterns help you understand how the causes of these issues? Please provide a concrete example.
- How these patterns might assist you in recognizing any moral judgments that might be expressed about the discussed issues? Was inferring moral judgment any different than other aspects of framing?
- How these patterns might be leveraged to identify or infer the potential remedies that are suggested in people's discussions?

If participants did not provide a response to a question during the survey, they were asked to describe any attempts they made, even if those attempts were unsuccessful.

The responses were analyzed using a **thematic analysis** approach (Braun and Clarke, 2006; Lofland et al., 2022) to determine how the linguistic patterns identified by each of

the discussed models most effectively assist participants in analyzing framing processes. Specifically, this analysis explores which aspects of the results may be irrelevant or confusing for the purpose of analyzing framing processes. Furthermore, these analyses examine which of the proposed linguistic attributes are most relevant for analyzing specific aspects of framing, including the identification of problems, the diagnosis of their causes, and the suggested remedies. These analysis are conducted as follows.

The interviews, which were conducted on Zoom, were transcribed using Temi platform (Temi, 2025). The author read through the transcripts and coded the key points made by the participants that developed through transcripts. After creating the codes, the author then reviewed the transcripts to ensure the identified codes are inclusive of all the key points made by the participants. The author and her advisor discussed the codes, to ensure each code is salient, and to validate the latent code identified by the author. By latent code, this work refers to codes that is inferred to capture the essence of experience that the participant reported, which may not be readily in the words used by the participants (Byrne, 2022).

The author and her advisor discussed, revised, and confirmed the code inferred by the first author. Next, in a collaborative session the author and her advisor discussed the codes, and grouped related codes under main themes that naturally emerged. Following this step, the author reviewed the themes to ensure each theme is focused on a distinct criteria, and whether the emerged themes cohere meaningfully together to respond to the evaluation purpose in this study. Lastly, the author reviewed the transcribed to find excerpt of interview that are related to each theme to be included in the following results. The results of these analysis are described in Section 4.4.5.

4.4.4 Interactive Interface for Human-Subject Model Evaluation

To facilitate analyzing the results in a human-subject study, a web-based interface has been developed. This interface is interactive, and enables researchers to explore different component of each of the models (e.g., top terms, grammatical relationships, co-occurring terms, and top document examples).

Figure 4.4 provides a screenshot of the interface of the LDA model. This interface includes topic terms, their probability, and example document in which topic terms appear.

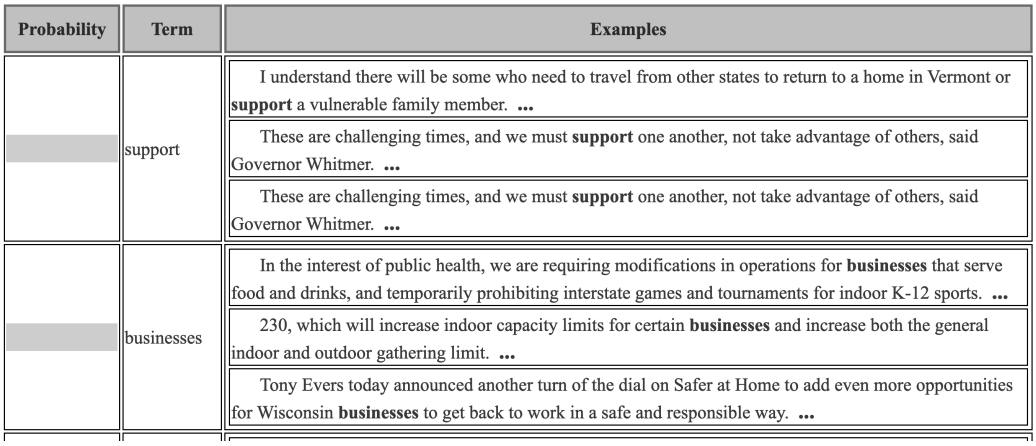


Figure 4.4: A screenshot of the LDA model's interface, which includes topic terms, their probability, and the example document in which they appear. Note: this screenshot only presents part of the topic, to give an overview of the the model components, while ensuring concision.

Figure 4.5 depicts a screenshot of the interface of the LDA-GR model. This interface includes topic terms, their probability, the grammatical relationship in which each topic term appears, as well as example documents in which topic terms appear in the captured grammatical relationship. Piloting this model with a number of researchers in our research lab, we learned that providing these grammatical relationships without their context makes it hard to interpret. Therefore, in a post-processing step, we captured example documents in which these topic terms appear in their associated grammatical relationships, provided that those documents are representative of the topic.

Figure 4.6 depicts a screenshot of the interface of the LLTR model. This model, as discussed in 4.3.3, includes topic terms, topic terms probability, their co-occurring terms, and example document in which topic term co-occurring with other topic terms within topic document. Similar to the two other models, these example documents are chosen only if their probability for the given document was above a certain threshold (i.e., threshold = 0.40).

4.4.5 Evaluation Results

To reiterate, the evaluation study is designed to assess the perceived efficacy of the models examined in this chapter in facilitating exploratory analysis of framing, rather than on the framings that researchers identified in this study. Therefore, this section does not discuss

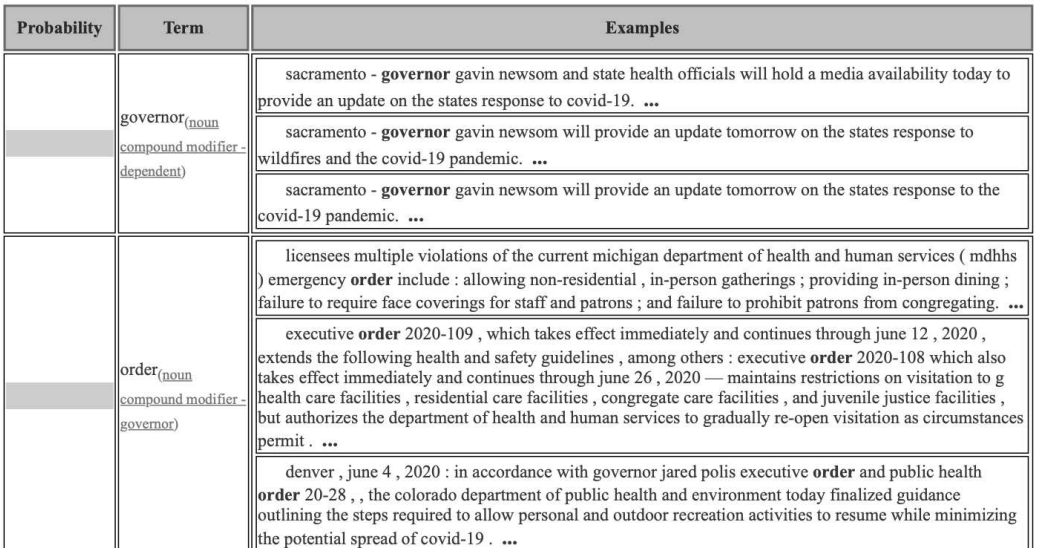


Figure 4.5: A screenshot of the LDA-GR model's interface, which includes topic terms, their probability scores, the grammatical relationship in which they appear, and the example documents of the appearance of each topic term in its associated grammatical relationship within the corpus. Note that this screenshot only presents part of the results, to give an overview of the the model' components, while ensuring concision.

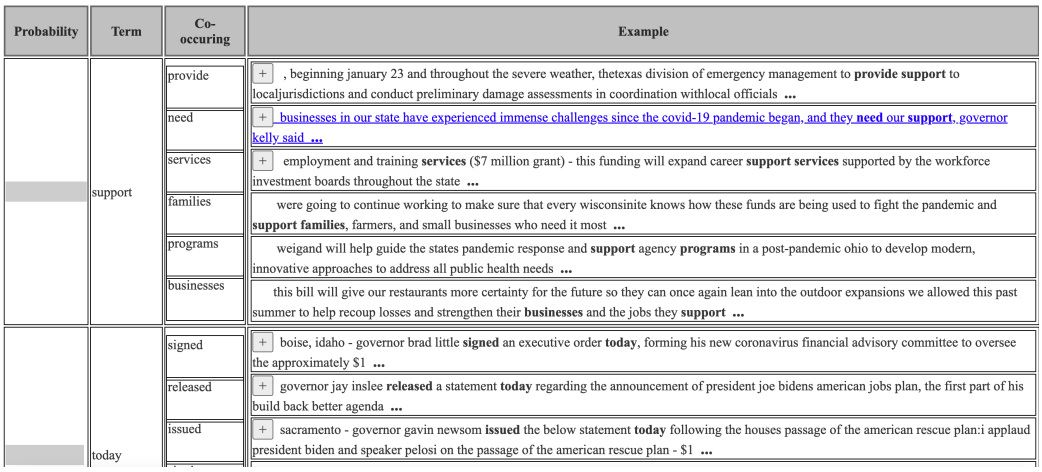


Figure 4.6: Screenshot of the LLTR model interface, including topic terms, their probability, a set of co-occurring terms for each topic term, and example documents in which each topic term appears with its co-occurring terms. Note that this screenshot only presents part of the results, to give an overview of the the model' components, while ensuring concision.

the specific framings that are identified in this study. Instead, it examines the criteria considered by our participants in assessing the efficacy and utility of the models discussed, as well as the ways in which they evaluated each model based on these criteria.

To find out these criteria, thematic analysis (Braun and Clarke, 2006; Lofland et al., 2022) was employed (Described in Section 4.4.3). In conducting the analysis, the codes that

emerged in the initial overview of the transcripts (Discussed in Section 4.4.3), were all focused on criteria about whether and how much different components offered by each model were effective in assisting the participants find evidence of framings (e.g., breadth of information, distinctiveness of topics), and whether and how much the results were consistent and sufficient for inferring framing processes by our participants (e.g., diversity of example documents, clear connections between example documents). These codes are grouped under four main themes, which all are related to the criteria by which the participants assessed the efficacy of the models in facilitating their analysis of framing processes. These criteria included *context*, *clarity*, *confidence*, and *curve*. The following passage briefly summarizes these four criteria. Next section then describes each of these criteria in greater details, followed by discussions around the degree to which our participants believed each of the three tested models met each criterion.

Context concerns the extent to which each model provides information about how each linguistic pattern (e.g., topic terms) appears within its immediate sentence, and the overall discussion in the document in which linguistic patterns occur. Participants also discussed the importance of whether and how much the each model provides diverse, yet connected contexts for each topic term, which was essential to capture broader arguments being made across topic documents. **Clarity** focuses on whether and how much the relationship between the model's results and framing language in corpus is readily clear to the participant. The clarity criterion is discussed in terms of whether and how much meaning of topic terms, as well as the broader arguments being discussed across the documents is clear, and if the participants could readily confirm the consistently of inferred framing across example documents. **Confidence** pertains to participants perceptions about the model's results being representative of the whole corpus, as well as participants' certainty about whether and how much the framings that they inferred using the provided results were through and were build upon on sufficient evidence. **Curve**, i.e., learning curve, refers to the time and efforts the participants need to spend to both learn different components of each model' results, and to then use those components to analyze framing.

4.4.5.1 Context

4.4.5.1.1 Criterion Definition: Understanding framing requires not only evidence about linguistic patterns that occur in a corpus, but also the indication about the *context* in which those patterns occur. Here, context refers both to the other specific words in close proximity of each topic term (i.e., within the same sentence) and to the overall ideas being discussed in the document where the pattern occurs (e.g., the outbreak, or vaccination, or government responsibility). For example, to figure the context of topic term "support", participants examined the immediate sentence in which the term appears, but they also needed to see the overall discussions around this term made in the document in which the topic term appeared, as well as the overall corpus, to figure what different type/types of support are considered and what entities are providing these supports, and who are benefiting from it. To figure the discussions in which topic term occurred, participants either used the co-occurring terms in the model that offered this component (i.e., LLTR), or skimmed, and sometimes needed to read the full document in the other two models (i.e., LDA and LDA-GR).

To infer framing, according to participants, understanding the context of topic terms is essential, but may not be sufficient. Specifically, participants highlighted the importance of observing different contexts in which topic terms occur to help understand the overarching arguments constructed with these topic terms, and to validate if the inferred framing is evident and consistent across multiple documents. For example, P1 noted that seeing just one example document (i.e., context) for a topic term "felt random", and they would need to observe a broader range of example documents (i.e., different contexts) to ensure that the framing evidence is consistent and not coincidental. In contrast, observing a topic term within the context of multiple documents, especially when the documents were closely related (e.g., different discussions around governmental supports for different impacted entities), allowed participants to identify evidence of framing. For example, P2 referred to the topic term "support", and explained that seeing it across different documents helped her understand the various types of supports that the government is providing to address the pandemic and its effects on various entities, including small businesses and families.

The importance of context(s), however, varied depending on the aspect of framing being addressed. Specifically, when identifying the issues discussed, participants noted that while context(s) facilitated easier inferences, they were still able to infer "general idea" (P2 and P4) about the issues within the corpus by only reviewing topic terms. However, when exploring other aspects of framing, namely causes, potential remedies, and moral judgments, participants emphasized the necessity of seeing the topic terms within their context, before making any inferences about these framing processes. For instance, P1 mentioned that "the moral judgment piece has a lot more to do with, with tone [...] without the examples, I think it's, I'm not sure there's really any way to get morality without that." Along this line, P4 emphasized the importance of seeing the context of topic terms to attend to moral judgments being made in the corpus. Specifically, P4 noted that "moral judgments are something that you need to physically read and kind of get the whole scope of the article and kind of the tone of the entire article to kind of understand.". For the moral judgment aspects, in particular, participants also emphasized the importance of "diverse contexts" to ensure if the moral assessment is consistent across documents and the captured evidence is not random (e.g., "like heroes on the front lines, like that's clearly a pretty like positive, um, appraisal, [...] but this honestly didn't stick out to me and it's hard to tell if this is indicative of, of sort of bigger things in these topics just because I think there's like one or two examples where it sticks out ").

4.4.5.1.2 Models efficacy in providing context(s) Participants noted the models differed in terms of providing effective context(s). Specifically, the LDA model is perceived as the least effective in providing the needed context to analyze framing. While this model provides multiple example documents for each topic term, offering the immediate sentence and the documents in which each topic term occurs, these examples documents were not perceived effective in providing the larger arguments being discussed in terms of how issues are discussed, and the nuances around them (e.g., what are the causes to the issues and what are potential solutions to address them) required to analyze framing. More specifically, the model did not provide supportive evidence to help connect the individual context of each topic term and form understanding around framing process. For instance, P4 mentioned

that:

> [...] the keywords or the terms that like that I encountered, um, they were really helpful kind of developing like a base frame or like a general idea of what the article was about. Um, it kind of lacked the specificity that I needed, especially for the kind of more complicated components like the moral judgments and the solutions.

Similarly, P2 mentioned while the LDA model signals the importance of words, it does not provide supports to understand why each word is important and how it is used in the documents (e.g., "I know that communities is important. I don't know exactly how they're talking about communities.").

In an attempt to figure out the broader arguments being made across the documents via checking the different contexts in which topic term appeared, participants mentioned the need to read the full documents, albeit reported to be not always effective. P2, for example, mentioned that

> [With the LDA model] I had to do a lot more reading and, um, like, it, it just kind of felt like the terms that were being pulled out, I guess they are helpful, but they didn't provide me nearly as much information as the other one. So I had to do a lot more digging and kinda reading with each of the terms.

Even when reading the documents, participants mentioned the provided documents in the LDA model lacked enough "narrative" to give them a holistic overview of the documents. More specifically, while LDA captured multiple documents in which topic terms appeared, participants noted that the captured documents tend to be "very similar" (P1), falling short in providing "different contexts" (P1, P2) in which topic term appear, thereby lacked the evidence about providing an overview of the ideas being discussed (as opposed to the LLTR that provided different ways in which topic terms appeared).

For example, P1 refereed to the topic term "emergency", and how all the example documents provided using the LDA model for this term all are focused on "emergency order", as opposed to the different context in which this exact term appeared in LLTR.

Lack of diversity of contexts in which topic term appears in the LDA model made the provided context less comprehensive, making it difficult for participants to capture framing languages across the documents. Along this line, P2 emphasizes the lack of connections across the documents, and notes that LDA seems to work as a "search function", "[that] is like somebody Ctrl+F for me, but they got to choose all the words that they looked up [Instead of the participant getting to choose]." They emphasized that apart from the need to read the whole documents captured in the LDA model, the example documents provided by LDA were not as cohesive in terms of context that these documents provided.

The LDA-GR model is evaluated very similar to the LDA model in terms of the effectiveness of the context it provides for framing analysis, with only minor differences due to offering an additional components, i.e., grammatical relationships. That is, similar to the LDA model, participants explored the immediate context of each topic term (i.e., the sentence in which topic term appeared), and the document in which topic term occurred. However, in most cases, participants needed to read the full documents to understand the context surrounding the topic terms. In addition, similar to the LDA model, the connection between the topic documents were less clear, and as a result, this model also falls short in providing the broader contexts to enable them identify the overall ideas begin discussed within the corpus.

While the LDA-GR model captured and provided the grammatical relationship in which topic terms occur, the grammatical relationships were not considered effective in capturing the contexts that participants required to analyze framing. Specifically, participants mentioned their attempts to use the provided grammars to connect the topic terms as a way to identify the broader arguments being discussed. However, these attempts were unsuccessful due to a)the high level of effort required to understand how to use the grammars, and b)participants needed to directly see the terms within their documents to understand the broader discussions being made. That said, participants mentioned the grammars influenced the way they engaged with the example documents, making their readings more intentional with respect to the grammatical relationships provided. For example, P3 mentioned that the provided grammatical relationships made her focus on the way each topic term appeared in the document, and that in turn guided the process by which she cap-

tured the context. For instance, P3 mentioned that there were time were looking at a topic term alone they would consider it a noun, but the grammar components suggests this term appear as a verb, and that would prompt them to look at the context in which this term appear as a verb and learn about how the term is being used.

In contrast, the LLTR model was the most effective both a) in providing diverse contexts for each topic terms made possible through the co-occurring terms and their associated example documents, and b) in making the connection between different contexts more readily visible, which together provided broader scope of ideas being discussed across the corpus.

Specifically, the inclusion of co-occurring terms for each topic term, as well as offering example documents in which these terms appeared together is discussed to provide a more holistic overview of the topics (P1, P2, and P4). These components together offered the needed and diverse contexts to attend to even more nuanced aspects of framing, such as moral judgment. P1, for instance, noted:

> These different words that go together [(i.e., co-occurring topic terms)] I think provides like a much, provides what feels like a more holistic overview of what is in the data. There's just like a lot of examples and there are a lot of different, different ones because they're all showing these different word combinations, [...] there is a lot more like I, language happening, like I'm thanking the governor, um, here, their work will mitigate the pandemic. Like that feels very different from almost any examples that came out in the other one [the LDA model]. [...] this is under another sort of co-occurring word thing and it still has similar, like, I am grateful we need to support. Um, so I think there's, "together", "together we will work". [...] there just seem to be like a lot of "we" and "collective language" behind this.

Participants compared the example documents in the LLTR model with those in the LDA model, highlighting how the LLTR model's focus on offering distinct co-occurring words and providing example documents for each of these distinct pairs of topic terms and their co-occuring terms resulted in more diverse and comprehensive understanding of

the corpus compared to the LDA. In contrast, while the LDA model provides example documents, these tend to be highly similar and lack the necessary diversity in context for effective framing analysis. Note that the reason for the need for seeing diverse contexts is mainly discussed as it gives participants confident about whether or not the framing they inferred is consistent across documents, or just appear in one random document (Discussed under the **Clarity** criterion).

In addition, participants noted that in addition to diversity and thoroughness of the provided context in LLTR (discussed above), this model made the connection between different contexts **readily visible**. That is, by providing the different terms by which the topic term co-occurred (i.e., the co-occurring term), and examples in which each of these pairs appears, this model facilitated a more comprehensive understanding of the overall ideas being discussed, without the need to read the documents in full. Put precisely, the co-occurring terms made the connection between the documents more easily *visible*, thereby provided the broader overview of the ideas being discussed within and across documents readily available. P2, for example, mentioned:

> I see some patterns here, of like these words that are showing up together. I know that "continue working" is showing up. I know that "continue to provide" is showing up, "continue to support", "continue to work". And so by seeing those patterns of, of these things are coming up together, that these words are coming up together, even just looking at this list, like work support, provide, take, working, keep serve efforts, ensure like all of these words are, uh, you know, linked to like a work ethic or hard work or like, um, consistency. And so it's definitely like this value that I'm seeing around, um, work ethic basically around hard work around, um, doing. And so I got that just from this model because I didn't even really need to look very hard at the examples. I didn't need to read very much. I could just tell from even this list of words, um, of the kinds of values that are coming up in this text and or in these texts.

Furthermore, participants mentioned seeing *different contexts* in which topic terms occurred assist them to make use of terms they would otherwise ignored in their analysis. For

instance, P1 refereed to the topic term "new", and how they would overlooked it in the LDA model. However, seeing this word with different co-occurring terms and in different contexts, such as *new reports*, *new cases*, *new hospitalizations*, *new results*, *new symptoms*, *new tests*, helped better understand what the topic is about and what framing language are happening in the topic.

4.4.5.2 Clarity

4.4.5.2.1 Criterion Definition: In leveraging the models' results to identify framing processes, participants frequently noted their efforts to *clarify* and *confirm* a) meaning of topic terms, b) connections between topic terms' contexts, as well as c) their own inferences using each model's results. At times, viewing example phrases in which topic terms appeared was sufficient for participants to clarify meaning of topic words and their context, and confirm their analyses. However, in other cases, participants found it necessary to read the entire document or even multiple documents in full, depending on the model used, and the framing aspects they were attending to (e.g., identifying the issues vs. inferring the moral judgments).

4.4.5.2.2 Models efficacy in terms of clarity The different models varied in both a) the frequency of required clarifications and b) their effectiveness in providing supportive evidence to resolve confusion.

The LDA required the most frequent clarifications attempts. Specifically, participants mentioned they would need to go back to the example documents in which a topic term appeared, and in a lot of cases, read the whole document to confirm the meaning of topic terms and their relationships between topic terms. They also refereed to the documents to confirm whether the framing processes that they inferred using the topic terms was evidence in the immediate document in which topic term appeared. P2, for example, mentioned:

> I had to do a lot more reading and, um, like, it, it just kind of felt like the terms that were being pulled out, I guess they are helpful, but they didn't provide me nearly as much information as the other one [i.e., the LLTR model]. So **I had to do a lot more digging** and kinda reading with each of the terms. So

like, just having access and there are no other kind of related words. I'm like, okay, what access? Like that could mean so many things in so many different contexts. So then I actually had to read more.

Participants noted that verifying the consistency of inferred framing evidence across multiple example documents was essential to them, and they would not relying on a single occurrence of framing evidence within a document. For instance, in relation to the need for multiple instances to confirm framing inferences, P1 discussed the LDA model did not provide example documents to confirm their attempts to infer moral judgment.

like that's clearly a pretty like positive, um, appraisal, um, morally of healthcare workers. Um, but this honestly didn't stick out to me and it's hard to tell if this is indicative of, of sort of bigger things in these topics just because I think there's like one or two examples where it sticks out and it's, it's, it's much harder I think to assess like, does this cover this whole data set? Um, because it also felt a little bit random in terms of where it was showing up, like there was one here or there. Um, but it wasn't like all of the, um, examples under emergency, for example, were talking about sort of the positive response to an emergency. So I think that makes it a lot harder than with sort of the different topics. You can sort of make some assumption that okay, these are repeating multiple times, so these are probably part of this larger data set, not just sort of a fluke of this one example has this tone[4].

In addition to frequent needs for clarification, this model also made the process of clarification "more tedious" (P2, and P4) (e.g., "[with LDA model] the lack of information made things a lot more tedious [...] it took more time clarifying and took more time finding those confirmatory measures just to make sure I was able to assign certain things to like the correct component of framing. ", P4)

For the LDA-GR model, participants noted that they went through an additional clarification process compared to the LDA model, due to the inclusion of grammatical relationships as a unique component of this model. Participants noted to first engaged with

[4]Note that this quote was also used to explain the Context criterion.

the grammars in an attempt to connect topic terms using their grammatical relationship. P1 mentioned clarifying the connection between topic terms, compared to LDA, was relatively less challenging (e.g., "it's make me feel like, oh, there are less, um, uh, obstacles, less difficulties."). P3 participants found this component (i.e., grammatical relationships) to be effective in informing participants about the role of each topic term when reading its context. Specifically, P3 mentioned:

> Then I started to find where that particular word really is in that sentence, and sometimes kind of like different from what I would imagine. And that's interesting. So I, I sometimes also might not look at the term definition very closely. And for example, when I see order, like if I don't read the explanation carefully, I might be thinking about verbs, right? So things like here there's order, but there's a verb. I was probably thinking about that in my mind. And when I open up this window though, it highlights it. You put the sentence in bold, I will read, oh, this is emergency order, and there's more explanation of what that it really is. So that's helpful, especially when I come back and like look at this term a little more closely.

Although relatively useful for clarifying the role of each topic terms, given the effort required to leverage this grammatical relationships, participants reported to eventually give up on it (discussed in more detailed in section 4.4.5.3). Next, participants went over a similar process as for the LDA model to clarify meaning of words using the context, and to confirm if the framing they inferred is supported within and across the topic documents. Again, similar to the LDA model, participants noted to have to read the full text to get the meaning of words, and understand if they interpret the topic correctly. In this process, participants mentioned they still tried to accounted for the grammatical relationship in which topic term occurred while digging into the document, which made their reading process more intentional.

The LLTR model significantly reduced the frequency and complexity of clarifying term meanings and contexts, for two primary reasons. First, by presenting co-occurring terms for each topic term and the contexts in which they appeared, LLTR provided readily available

context, reducing the need for further clarification of contexts. Second, the capturing the diverse co-occurring terms for each topic terms, and consequently, diversity of these contexts for topic term pairs, offered a comprehensive overview of the corpus and the arguments being made across the documents. This accordance facilitated the identification of supporting evidence for framing language within and across documents, making the clarification process easier. Specifically, participants reported a reduced need to read entire documents, due to LLTR's co-occurring term feature, if at all For instance P2 mentioned:

> I see some patterns here of, of like these words that are showing up together. I know that continue working is showing up. I know that continue to provide is showing up, continue to support, continue to work. And so by seeing those patterns of, of these things are coming up together, that these words are coming up together, even just looking at this list, like work support, provide, take, working, keep serve efforts, ensure like all of these words are, uh, you know, linked to like a work ethic or hard work or like, um, consistency. And so it's definitely like this value that I'm seeing around, um, work ethic basically around hard work around, um, doing. And so I got that just from this model because I didn't even really need to look very hard at the examples. **I didn't need to read very much.** I could just tell from even this list of [i.e., co-occurring] words, um, of the kinds of values that are coming up in this text and or in these texts [referring to the excerpts in which topic terms and their co-occurring terms appear].

Similarly, P1 and P4 noted that the co-occurring terms clarified the relationship between terms, provided an overview of how the topic terms relate to one another, making the framing evidence much clearer and nearly eliminating the need to read entire documents for clarification.

4.4.5.3 Curve

4.4.5.3.1 Criterion Definition: In assessing the models' usability the models, participants considered the efforts required in working with each of the models, i.e., "learning

curve". This criterion includes two aspects, 1) **model learning curve**, which relates to understanding the different components of each model (e.g., top terms, grammatical relationships), and 2) **results learning curve**, which pertains to the effort needed to explore framing language through the model's results. In many cases, the models' learning curves were inversely related to the results learning curves. This section presents the results based on participants' evaluations of each model in terms of these two learning curves.

4.4.5.3.2 Models efficacy in terms of Curve The LDA model, is discussed to require the least model learning curve, mainly due to this model's reduced number of components (i.e., topic terms and example documents). However, this same reason, i.e., less supportive components, led to LDA having the greatest results learning curve in exploring framing evidence. Specifically, participants discussed the LDA model did not provide any supportive components to find the connection between topic terms, and to find the broader ideas being discussed beyond the example documents. As a results, this model made it more difficult to utilizing its results to effectively analyze and understand the nuances of framing within and across the documents. More specifically, participants found it challenging to determine the connection between the different contexts (example documents) in which each topic term occurs, to connect the contexts of different topic terms, and to examine whether the inferred framing is salient across documents. They reported to need to do more extensive reading of example documents, and spend more cognitive effort when exploring this model's results.

The LDA-GR model, exhibited a higher model learning curve compared to LDA, due to inclusion of grammatical relationships in which topic terms occurred. Overall, participants reported a moderate model learning curve in understanding how to interpret and utilize the grammatical relationships for each topic term. This model also had high (but not the highest) model learning curve. Specifically, the grammatical relationships provided more support to help read the documents more intentionally with respect to the role of each topic term, making this process relatively easier compared to the LDA model. However, this grammatical relationships still did not provide the support needed to effectively find framing evidence in the results. Put precisely, the primary challenge arose during the *results learning curve*, as participants found it more demanding to integrate topic terms,

their grammatical relationships, and associated documents to identify and analyze framing within the data. For instance, P5 relates the use of the LDA-GR model to solving a "puzzle", and mentions that "It takes more effort for me trying to like, you know, put all the things together, all the puzzle together".

Lastly, the LLTR model is argued to have the highest *model learning curve*, requiring the greatest effort from participants to understand its different components (i.e., topic terms, co-occurring terms, and example documents). Similar to the LDA-GR, the higher model learning curve is the result of the additional components offered in this model, i.e., the co-occurring terms. However, this additional components, i.e., the co-occurring terms captured for each topic, is ultimately rewarded, as the inclusion of co-occurring terms significantly facilitated the "results learning curve", enabling participants to more effectively utilize the model's output for framing analysis. For instance, P1 noted:

> **So there was maybe a little bit of a learning curve to figure out how to use it.** Um, I think the upside of all that complexity is that there's sort of a lot more nuance here. I think as I've been saying, like I do think these topics (in the LLTR model), those different, um, topics did sort out a lot more discreetly. **It was a much easier to sort of see them as different things. And all of the examples I think were incredibly useful** to sort of dig in and get a better sense of what, what these words were, were doing. Um, and I think I just, I this like the multiple bolded words in the example, I am sort of amazed at like how helpful that that is.

4.4.5.4 Confidence

4.4.5.4.1 Criterion Definition: Participants also discussed their experience in terms of **confidence** when working with each model. This criterion is examined under two main aspects, a) confidence in the results provided by each model, and b) confidence in their own analysis based on the model's outputs.

Participants confidence about the models' result mostly focuses on whether and how much the results provided by each model are representative of the corpus. More specifically,

participants emphasized the extent to which each model provided a more comprehensive view of the data in the corpus, as well as the diversity of the supporting documents, while maintaining topics' focus. Patricians' confidence in their own analysis is directly related to their trust in the model's results being representative of the corpus, as well as their perceived level of success in identifying framing processes using each model's results.

4.4.5.4.2 Models efficacy in terms of Confidence Participants expressed relatively the least confidence in the LDA model, moderate confidence in the LDA-GR, and the highest confidence in the LLTR model, both in terms of confidence about each model's results, and their own analysis using the model's results.

With the LDA model, participants emphasized how the model fall short in providing the full overview of the corpus, making them less certain about whether or not the provided context using this model is representative of the whole corpus. While the model provided multiple example documents for each topic term, these documents were reported to be very similar, thus the offered context that lacked diversity and comprehensiveness. P1, for example, mentioned while they saw one document with moral language in it, it "felt a little bit random ", thus they could not trust to make their analysis of moral judgment only based on one sample document. More specifically, P1 mentioned:

> I think [with LDA model] this was challenging 'cause I think there, there are some places like I'm seeing right now, like this one has some moral judgment. Um, I'm now seeing like heroes on the front lines, like that's clearly a pretty like positive, um, appraisal, um, morally of healthcare workers. but this honestly didn't stick out to me and it's hard to tell if this is indicative of, of sort of bigger things in these topics just because I think there's like one or two examples where it sticks out and it's, it's, it's much harder I think to assess like, does this cover this whole data set? Um, because **it also felt a little bit random in terms of where it was showing up, like there was one here or there.**

Participants confident in their analysis was directly related to their perceptions about whether and how much the model's results are representative of the whole corpus, and

their perceived success in inferring the framing processes. With respect to the first aspect, for example, P2 mentioned " I don't know what, I'm not framing because I'm seeing such a limited amount of the text, so I'm like wondering how these snippets were chosen". Regarding the second aspect, P4 mentioned "[with LDA model], I wasn't as thorough as I would have liked to be, or I didn't feel as confident in my responses than I would in the second model [i..e, the LLTR model]".

Participants discussed to be more confident with the LDA-GR compared to the LDA model. This assessment is directly related to the way LDA-GR provided the grammatical relationship in which topic terms appear, and the way this component provided relatively more support to connect topic terms with more certainty. For instance, P5 emphasized the way these grammatical relationship helps them to be "objective" when they were connecting the topic terms together to make sense of these terms and find framing evidence (P5) (ADD). Another participants also mentioned accounting for the grammatical relationship in which topic terms occurred helped them to correct their assumptions, making them more confident in their own analysis. More specifically, P5 mentioned that:

> they [i.e., grammatical relationships] were correcting my assumptions because I remember like before I click on it, I was thinking, oh, this must mean that things were getting worse, but it actually talk about the, it getting better or vice versa.

However, similar to the LDA model, the example documents provided by LDA-GR were not as diverse and participants felt these examples may not be representative of the whole corpus. As a result, this model fell short in providing a broader overview of the corpus, leading participants to feel less confident in whether the results were representative of the whole corpus.

Lastly, participants expressed the highest level of confidence in the results generated by the LLTR model, and consequently, in the framing analysis they conducted based on these results. This assessment is linked the breadth of information covered in the LLTR (P1, P2, and P4). Specifically, the different co-occurring terms captured for each topic terms, as well as readily capturing the way each pair of topic terms and its co-occurring terms

appear in the context of diverse set of documents offered a "holistic overview of the corpus" (P1). Having a clear overview of the corpus made participants feel more confident in the thoroughness of the results, and in turn their own analysis using the results. For instance, P1 noted that

> I think it's definitely, it's, **it's the volume but like the correlate to the volume is also the breadth.** It's, it's just like there's a lot, there's a lot more different stuff. **It's not just that there is more, there is a much bigger spread of material, um, that I think gives more confidence that you're getting a fuller picture of what this is.**

P2 compared their experience working with the LLTR model to their experience when doing inductive coding, stating that they would be more confident if the model supplement their analysis, due to two factors. First, it makes much less time to explore large volume of data. Second, the model can divide its attention equally to different part of the documents, since it works based on the idea of capturing the most probable words. More specificity, they mentioned:

> There are two big ones. One is **time** and like [the other one is] **attention**, right? [regarding the time aspect], it would have been a, a lot for reading all of that and then analyzing it would've taken a really, really long time. but then the second one [regarding the attention aspect] is after a while of reading, there are gonna be things that I'm gonna miss that. [...] the model, [however] is not gonna get tired of finding the probability, so it's a little more reliable than me reading through, and I probably wouldn't have read through, um, as systematically as, as like the model is gonna just find the probability. like there would be less consistency and I probably would've emphasized certain things just because I was reading more attentively at certain times compared to others.

All the participants unanimously mentioned they are interested to use computational models to "supplement" their framing analysis and provide "framing evidence", but they would not be comfortable with allowing these models to infer framing processes directly.

Criterion	LDA	LDA-GR	LLTR
Context	✓Sufficient Contexts to capture discussed issues ×Falls short in providing the broader overview of ideas discussed in corpus × Lacks diverse contexts	✓Sufficient contexts to capture discussed issues ×Falls short in providing the broader overview of ideas discussed in corpus × Lacks diverse contexts	✓Provided the most effective contexts by capturing co-occurring terms ✓ Context readily and easily available through offering co-occurring terms ✓ Offered contexts were diverse and comprehensive
Clarity	× Lacks clear connections between example documents × Required a lot of reading of the documents in full to clarify the meaning of words × Required a lot of reading of the documents in full to confirm the inferred framing	✓ Grammatical relationships helped clarify how the topic terms are used × Efforts required to account for grammatical relationships made this process less effective × Required a lot of reading to confirm inferred framing	✓ Made meaning of topic terms clear, due to providing their co-occurring terms ✓ Easy clarification process, due to providing the broader overview of the corpus
Confidence	× Lack of diverse contexts, and sparsity of context supporting each framing evidence made researchers less confident about the results being representative × Less confidence about representatives of results reduced participants' confidence in their analysis	× Offered more confidence about connecting the topic terms, due to providing grammatical relationships. × Offered less confidence about whether model results are representative of the broader corpus	✓ Offered confidence in model's results being representative of the broader overview of the corpus. ✓ Made participants more confident about their framing analysis
Curve	✓ The easiest model learning curve, due to the lowest numbers of components × The greatest results learning curve, due to the lack of supportive components to connect the results	× Increased model learning curve, due to the addition of grammatical relationships × High results learning curve, due to the lack of sufficient connections between example documents	× The steepest model learning curve, due to increased numbers of components ✓ The least results learning curve, due to ease of connecting the model's components, mostly made possible through capturing co-occurring terms

Table 4.1: Comparison of the LDA, LDA-GR, and LLTR models in terms of context, clarity, confidence, and curve. The LLTR model provided the most diverse and interconnected contexts, enhancing the clarity of framing evidence and resulting in the highest confidence in model results, thereby participant's highest confidence in their own framing analysis. However, LLTR requires the steepest learning curve.

The participants highlighted the nuances involved in framing processes, and how they would not rely on computational models to infer these nuances. In addition, participants all emphasized linguistic patterns are important, but it is important to see the context in which these linguistic patterns appear. Thus, any model that is designed to support researchers with these analysis should effectively provide the context in which these linguistic patterns appear, so that researchers can refer to these contexts to understand the broader ideas being discussed and to confirm whether or not there inferences are consist across different contexts. This feedback aligns with the approach adopted in this work, which designs and utilizes computational techniques to capture linguistic patterns indicative of framing with the goal to facilitate exploratory analysis, rather than designing these models to directly infer framing processes.

4.4.5.5 Summary of the models evaluations

This section summarizes effectiveness and usability of the three models as evaluated by participants across the four key assessment criteria, discussed in preceding section. Table 4.1 outlines an overview of this this summary, in terms of whether or not each model meets the criteria mentioned by the participants.

4.4.5.5.1 Assessments of LLTR: Participant assessments indicated that the LLTR model exhibited the highest efficacy in providing diverse and connected contexts. This assessment was attributed to the model's co-occurring term component, participants reported as highly effective in facilitating framing analysis. Participants reported that the LLTR model required the least attempts to clarify results interpretations, and the inferred framing processes. When clarification was needed, the process was significantly more straightforward compared to the LDA and LDA-GR models. This enhanced usability was attributed to the LLTR model's unique feature, i.e., the inclusion of co-occurring terms for each topic term within the context of example documents. By providing a comprehensive corpus overview through diverse contexts for each topic term, the LLTR model cultivated greater confidence in the representativeness of its results and enhanced participant confidence in their own framing analysis. However, the LLTR model exhibited the steepest model learning curve,

primarily due to its additional components (i.e., co-occurring terms). Nonetheless, once participants became familiar with these components, the results learning curve for examining framing evidence became greatly faster and smoother compared to the other two models.

4.4.5.5.2 Assessment of LDA: The LDA model is discussed to fall short in providing sufficient contexts required for framing analysis. Participants frequently reported the need to read the example documents in full to clarify meaning of terms and the connection between topic terms, and confirm whether the framing evidence they observe is consistent across documents. This necessity for clarifications, coupled with the perceived randomness of the context in which topic term occurred, resulted in participants' lower confidence in both the model's results and their own analyses using those results. In terms of learning curve, due to the model's reduced number of components (i.e., topic terms and the example documents), the LDA model has the least model learning curve. This reduced number of components, however, comes at the cost of the greatest results learning curve, as the participants did not have "supportive components" while they were examining framing evidence, and needed to read the documents in full more often.

4.4.5.5.3 Assessment of LDA-GR: Similar to LDA, the LDA-GR model is discussed to fall short in providing diverse contexts for topic term. Compared to LDA, however, grammatical relationships in LDA-GR helped participants be more intentional when reading example documents. That said, the provided grammatical relationships did not contributed to capturing the broader arguments being discussed across the documents. This component (i.e., grammatical relationship), made participants feel they are able to connect topic terms more "objectively" compared to LDA, making them relatively more confident in their framing analysis. However, given the cognitive efforts required to utilize grammatical relationships, participants reported to give up on this affordance of LDA-GR and relied more in reading the provided example documents. Compared to the LLTR model, this model falls short in providing the different contexts in which a topic word occurs, and therefore, does not provide the broader overview of the contexts. As a results, participants did not feel through enough with their analysis. Furthermore, the LDA-GR model exhibited the a

high results learning curve, which was due to efforts required to account for grammatical relationships, albeit often not effective, as well as the necessity felt to read the documents in full.

4.5 Contribution and Future Work Directions

This chapter makes a technical contribution by offering a computational model that facilitate the analysis of framing processes across the wide, heterogeneous information ecosystem. To summarize, to design these models, this chapter investigates linguistic attributes whose relevance to framing language is motivated by the concept of framing and the functions it performs, as well as prior work in identifying the prominent frames. This chapter then examines the effectiveness of these proposed attributes in identifying patterns that pertain to the language of framing, the work presented in this chapter both explores an existing unsupervised topic modeling approach, the LDA model, as used in prior studies (e.g., Blei et al., 2003; Walter and Ophir, 2019), and also expands this model, and design and develop two other models to incorporate linguistic attributes not integrated in prior models. This chapter evaluates the proposed models and examines how the linguistic patterns identified using these models might be effective in assisting researchers in identifying evidence of framing processes in language. The evolution study demonstrates that the Linked Latent Theta Role model (LLTR), by capturing both topic terms, and their interrelationships, and through offering the diverse contexts in which these terms appeared together, provides a more effective and more comprehensive understanding of framing processes within a large corpus of documents compared to the LDA and LDA-GR models.

The techniques designed and developed in this chapter contributes to a) misinformation research, by enabling to study the broader scope of misinformation and its impacts through the lens of framing (Discussed in Chapter 5), b) research on phenomena similar to misinformation, such as disinformation and rumor spreading to examine the manifestation of such phenomena and their impacts in framing processes, and more broadly to c) framing research by enabling to examine framing processes around different events using an ecological approach.

In the context of this dissertation on misinformation, the designed and developed computational techniques in this chapter, in particular the LLTR model, enhance the analysis of framing, and makes it possible for this dissertation to investigate the relationships between framing processes and the phenomenon of misinformation. In particular, the designed and developed LLTR model makes it possible to study and observe some of the manifestation of misinformation impacts as evidence in processes of framing (discussed and examined in great detail in the next chapter)

In addition, the computational techniques presented in this chapter makes a valuable contribution to research on similar phenomena, such as disinformation dynamics, the spread of rumors, and the role of deep stories in facilitating such rumors (e.g., Starbird et al., 2019; Prochaska et al., 2023). More specifically, the tools presented in this chapter enables to explore both the dynamic of such phenomena by looking at the content shared that are related to specific disinformation efforts, and/or rumor, as well as to examine the evidence of manifestation of such phenomena on people interactions and processes of meaning constructions. For instance, considering *disinformation* as a collaborative work that is distributed across media ecosystem (Starbird et al., 2019), Prochaska et al. (2023) studies the dynamics of these efforts by tracking *deep stories* (Polletta and Callahan, 2019). In their work, Prochaska et al. (2023) adopts a qualitative coding approach and suggest that disinformation might mobilize online audiences. Given that these processes are not always evident in single documents, and researchers may lack awareness of all relevant deep stories, qualitative coding for these stories can be overwhelming, if not impossible in some cases (e.g., when there is no prior knowledge of potential deep stories). The LLTR model developed in this chapter can be employed to examine the framing processes involved and may offer a valuable approach to examining disinformation dynamics, potentially addressing the challenges associated with the qualitative coding approach

Furthermore, the techniques presented in this chapter more broadly contribute to framing research with different focuses. Specifically, by enabling researchers to analyze framing processes related to different events across the wide, diverse information ecology, the techniques designed and developed in this chapter enables framing researchers to examine framing processes with a more ecologically situated approach on framing evidence . For

example, using these techniques research can expand knowledge on framings around how people interpret different events (e.g., residential elections, ongoing global conflicts, or policies concerning minority groups) using a more ecological approach. Put concisely, not only the techniques provided in this chapter addresses the scalability concerns of manual coding, a techniques that is commonly used to analyze framing, these techniques also enable to expand our horizon when examining framing processes beyond the framing evidences within a single document. Put differently, using these techniques, enables to synthesis on evidence across documents to infer framing processes that might be distributed across multiple documents.

While the LLTR model, designed and developed in this chapter, is shown to be effective in facilitating exploratory analysis of framing, it represents an initial step in applying computational techniques to examine framing from a dynamic, processual perspective. Therefore, this model offers opportunities for future research to build upon. First, future work can improve the usability of this model, particularly given the model's steep learning curve. Conducting human-subject studies with researchers as participants could provide valuable insights for enhancing user experience with this model. Second, future work should explore the integration of additional linguistic features, such as catchphrases and metaphors, to enrich the model effectiveness in providing evidence of framing processes. Future work can also examine implementing variable weighting for document sections (e.g., titles, introductions, bodies, conclusions) based on their potential framing influence, as suggested by researchers who evaluated and discussed the efficacy of the tested models.

Chapter 5

Exploring Misinformation as a Sociological Phenomenon: The Entanglement of False and Misleading Content with Framing Processes

5.1 Introduction and Motivations

In line with this dissertation's view of misinformation as a broad sociological phenomenon that transcends any individual piece of false or misleading content, this chapter examines the relationships between the processes of framing and the phenomenon of misinformation. Specifically, it examines whether and how there are differences in the way communities with different prevalence of false and misleading content perceive and interpret world events, including the way they frame different events, and the way they respond to framing presented by other sources, either implicitly or explicitly.

Building on the findings from Chapter 3, which demonstrates the broad impacts of **the prevalence of false and misleading content** and **community response** at the com-

munity level (e.g., influencing individuals' perceptions about a community and its norms), this chapter explores how the interplay between these elements could manifest in the processes of meaning constructions about an event (i.e., framing). Chapter 3 primarily focuses on community responses in terms of expression of agreement and disagreement with shared content. However, community responses are indeed more nuanced than simple expressions of agreement or disagreement. To capture these complexities and further explore the interaction between community responses and false or misleading content, this chapter adopts the concept of framing, as motivated in the previous chapter. In particular, it investigates how framing is manifested in both individual responses to news media articles and in the collective discourse within online communities regarding these articles.

To account for **the prevalence of false and misleading content** and to examine its interplay with **community response**, this chapter examines framing processes in two communities with varying levels of false and misleading content prevalence. It explores the way these two communities respond to the framing of the same event in the mainstream news media articles, focusing on the differences in frame manifestation between these two communities.

Indeed, online communities might respond to framing depicted in the news media in different ways, from reinforcing the framing in the original content to revising it in some way, to completely rejecting the provided framing and offering their own perspectives. This chapter explores whether there are meaningful differences in how communities engage with these ways of responding in online communities wherein false and misleading content is less or more prominent [1]. When referring to community responses, this work both accounts for the content individual community members share about mainstream news media articles (i.e., posts), as well as community members' discussions around the content shared (i.e., comments). Put precisely, this chapter asks:

RQ: What are the differences in how the processes of framing unfold, as observed in

[1] Given the numerous factors that come into play when communities interpret events and respond to news content from other sources, the work presented in this chapter is not intended to test causal impacts of the prevalence of false and misleading content on the way communities frame an event. Instead, this study takes an observational approach and explores whether there are any meaningful differences in the way an event is framed in response to mainstream news media, in online communities wherein false and misleading content is less or more prominent.

community members' interactions/discussions in response to events covered in (mainstream) news media when false and misleading content is more or less prevalent within a community?

To investigate the processes involved in framing, this chapter leverages the computational model developed in Chapter 4, specifically the Linked Theta Role Model (LLTR), that is identified to be the most effective in assisting researchers with analysis framing processes. As a testbed to explore processes involved in framing and investigate the research question motivated above, consistent with the focus of the data in Chapter 4, this chapter sticks with the COVID-19 pandemic, for the reasons described in the preceding chapter.

5.2 Methods and Experiments

This section outlines the methods developed and utilized to investigate the research question presented above, including the study designs, data collection processes, and analysis procedures.

5.2.1 Study Design: Examining the Interplay between Misinformation and Framing Processes, as Evidenced in Community Responses to Mainstream News Media

As motivated in Section 5.1, to examine the interplay between false and misleading content and community response through the lens of framing, this chapter examines the differences in framing manifestations evidenced in communities response to news media framing of the same event (i.e., RQ stated in section 5.1). To do so, it focuses on the *content* (i.e., posts) that community members share about news covered in mainstream news media about an event, and the *discussions* that occur around those posts (i.e., comments).

With this goal, this chapter employs the approach described in Chapter 4 that is the most effective in analyzing framing processes, i.e., Linked Latent Theta Role (LLTR). Specifically, it investigates the framing of the COVID-19 pandemic in two communities wherein false and misleading content is less and more prevalent (e.g., r/science, and r/conspiracy), as well as the framing of the same event in the news media articles, and compare and contrast these processes. It examines how each of these communities might shift the framing depicted

in the news media articles about the COVID-19 pandemic, in a way that might reinforce, revise, or entirely reject the framing in the original content.

To compare and contrast these processes, this chapter mainly conducts a qualitative, exploratory analysis of the inferred framing processes across each of these communities and the news media articles. The qualitative portion is motivated with the nature of framing, that is inherently exploratory and requires engagement with the actual documents and nuances in language. Next, it explores quantitative investigations to to bring specific evidence to support the aforementioned qualitative analysis.

Thus, as the first step, this work examines the evidence of framing processes in each of the two aforementioned communities, as well as in the news media articles that are being shared. Specifically, it examines the functions which framing performs in each of these communities, such as determining what counts as problems, who are assigned the agencies, what remedies are suggested, and potential moral judgments that are discussed regarding the situation. Next, this work investigates the differences between the framings that are evidenced in each of these communities with the framings manifested in the news media.

5.2.2 Creating the Corpus

The corpus includes documents from the following three sources: a) an online community wherein false and misleading content is more prominent, b) an online community wherein false and misleading content is less prominent, and c) prominent news media in the U.S. To be specific, for an example of online communities wherein false and misleading content is relatively more prominent, this work focuses on a community on the Reddit platform, the r/conspiracy subreddit. This community has been referred to as the most prominent conspiracy theories community in prior work (Phadke et al., 2022), and is well-known in research around misinformation on social media (e.g., Phadke et al., 2022; Klein et al., 2019). For an example of online communities wherein false and misleading content is relatively less prominent, this work focuses on the r/science community on Reddit, which is known as the largest scientific community on the Reddit platform. This community is referred to as the polar opposite of the r/conspiracy community in terms of prevalence of false and misleading content (Phadke et al., 2022). To examine the framing of the COVID-19 pandemic in news

media, this work focuses on the mainstream news media outlets that are linked in these sub communities. Informed by prior work, mainstream news media is considered any news outlet that has a high reach, and thus, can potentially influence many people opinion and thoughts of current events (Chomsky, 1997). These outlets hire trained journalists and editors to ensure quality of their content, and present the news across channels (e.g., AP, CNN, WSJ, The New York Times, MSNBC, ABC News) (Maryville University, [n. d.]; new, 2022; Basch et al., 2020).

The data for this study is collected from a public repository, named Academic Torrent. This repository is collected and maintained by researchers. To collect data of interest from this resource, first the posts that contains at least one of the COVID-19 related keywords (e.g., "COVID19", "coronavirus", "ncov2019", "2019ncov", "nCoV", "CoV2019", "2019nCoV", "COVID19") (Wicke and Bolognesi, 2020) are extracted. This step resulted in a dataset containing 552 posts from the r/science community and 902 posts from the r/conspiracy community, spanning the years 2019 to 2023. Next, given that this study focuses on the evolution of framing in response to framing of the COVID-19 pandemic in the news media, the dataset was filtered to include only posts that link to articles from mainstream media outlets. This inclusion criterion resulted in 132 posts from the r/science community and 218 posts from the r/conspiracy community. Following this step, all comments on these posts were retrieved to facilitate an exploration of the discussions surrounding the shared news. The posts received a varying number of comments, ranging from 0 to 290, with a mean of 20 comments (SD = 39.54).

Most of the posts were short, and much of the discussions occurred in the comment sections following the posts. To address the issue of short documents, the pre-processing step involved concatenating each post with all the associated comments into a single document. However, since posts received varying numbers and lengths of comments, the resulting documents varied significantly in length. To create more uniform-length documents, the documents were truncated to approximately 300 tokens. This length was chosen arbitrarily based on exploratory analysis of the average length of discussions around shared news media. As a result, the final corpus consists of 1517 documents, each with an average length of 300 tokens.

5.2.3 Analyzing Framing by Synthesizing Framing Evidence Across Topics

Leveraging the Linked Latent Theta Role Model (LLTR), which is identified to be the most effective model in assisting researchers to identify evidence of framing in Chapter 4, this chapter examines the manifestation of framing processes across LLTR results (i.e., topics).

Due to the inherent complexity of framing analysis, a qualitative investigation is conducted to enable engaging with the linguistic patterns captured in LLTR topic results, and infer the evidence of framing processes within these results. This section details the multi-step procedure used to leverage LLTR topics for identifying framing processes (Section 5.2.3.1). Subsequently, it details the approach taken to identify framings *within* specific sources – namely, r/science, r/conspiracy, and mainstream news media (Section 5.2.4). Finally, it describes the methods employed for the *cross-source comparison* of framing processes (Section 5.2.5).

5.2.3.1 From Topics to Framings

This section outlines the methods taken to analyze framing, which include a) extracting topics using the LLTR model, b) analyzing each topic captured by LLTR to identify key points at the topic level, and c) synthesizing across the captured topics to infer framing evidence and framing processes across the corpus.

5.2.3.1.1 Extracting Topics Using LLTR As discussed, given the efficacy of LLTR model in facilitating framing analysis, this model is leveraged for framing analysis in this chapter.

For using this model, the first step was to find the number of topics and number of the linked latent theta roles. The optimal number of topics for the corpus detailed in Section 5.2.2 was established through coherence metric analysis. This analysis revealed that a topic count of four (K=4) resulted in the most cohesive topic representation for this corpus. To determine the optimal number of linked latent theta roles (T), this dissertation followed the approach outlined in Bamman et al. (2014), evaluating T values of five, ten, fifteen,

and twenty. Selection of T was based on the cohesion of co-occurring words (i.e., a key component of the LLTR model), and resulted in the choice of five linked theta roles (T=5). That said, future research should explore the feasibility of employing coherence metrics to inform the selection of T, similar to their application in determining K.

5.2.3.1.2 Analyze each individual topic First, the theme of each topic is identified to provide a preliminary understanding of the arguments presented. This step is done by reviewing topic terms, their co-occurring terms, and reviewing example documents in which topic terms co-occurred together. Next, a detailed review of each topic was conducted to take notes on the key points observed within each topic. These notes are oriented using the various functions by which framings perform (i.e., what problems are highlighted in each corpus, who are assigned the agencies of those problems to, what remedies are discussed with each corpus, as well as any potential moral judgments around the event and its surrounding issues). While these functions are leveraged to orient the analysis in this step, any other evidence that seemed important to process by which people understand the pandemic is also taken notes of in this step.

To clarify, this step was not aimed at identify the aforementioned functions to map them to a framing package. Indeed, as discussed in Chapter 4, this study avoids a one-to-one correspondence between topics and frames. Thus, the notes generated during this step were not used for direct topic-to-frame mapping. Instead, these notes were taken to facilitate synthesizing across topics, which is done in the next step, for identifying framing processes evident both within and across topics (Described in the following section).

5.2.3.1.3 Synthesis across topics to infer framings The notes taken in the previous section were leveraged to identify framing processes that might appear either within or across the topics. This step is inspired by the thematic analysis by Braun and Clarke (2006) to enable identifying latent evidence of framing processes, and the analysis in this step are oriented by the functions by which framing performs (i.e., what problems are highlighted in each corpus, who are assigned the agencies of those problems to, what remedies are discussed with each corpus, as well as any potential moral judgments around the event and

its surrounding issues) (Gamson and Modigliani, 1989). Specifically, similar to the previous step, in this step, the four main functions by which framing performs (i.e., what problems are highlighted in each corpus, who are assigned the agencies of those problems to, what remedies are discussed with each corpus, as well as any potential moral judgments around the event and its surrounding issues) are first answered using the notes identified in the preceding section. To emphasizes, these notes served to orienting the analysis, and were not treated as the primary sources of evidence. Indeed, the analysis involved a further review of the topics to extract framing evidence informed by these initial observations gained using these notes.

Next, the extended notes gained from this step were organized to capture the framing packages by which the COVID-19 pandemic (i.e., the event under discussion), and its surrounding aspects were interpreted and understood within each corpora (i.e., r/science, r/conspiracy, and mainstream news media). Concrete example results were reviewed and included in describing these identified framings. This step creates the final framing packages that are described in Section 5.3. The following section describes the approach taken to divide the topic results to pertain to a single corpora, making it possible to infer framing processes within each of the aforementioned sources.

Note that this analysis do not enforce any fixed numbers of framing packages. Rather, the identified framings are naturally emerged using the author's analysis, which are also reviewed, refined, and confirmed in discussion with her advisor. Thus, different number of framings might be identified within different sources (i.e., r/science, r/conspiracy, and news media).

5.2.4 Identifying Framings Within Individual Sources

Following the procedure outlined in the previous section (i.e., Section 5.2.3.1), framing patterns were identified within each source (r/science, r/conspiracy, and news media). While topics were generated from the combined corpus, to facilitate the analysis, each topic example was labeled by its source and visualized separately using the interface described in Section 4.4.4 to facilitate this analysis. As previously stated, this analysis did not impose a predetermined number of framings for each source. Instead, framings were allowed to

emerge organically based on the framing evidence emerged within the results, and from the author's interpretations of the results. Consequently, the number of identified framings varied across the analyzed sources.

It is important to acknowledge that due to the inherent subjectivity of framing analysis, alternative framings, and possibly different numbers of framings might arise from independent interpretations and analysis of the same corpus.

5.2.5 Cross-Source Comparison of Framing Processes

As motivated in Section 5.1, this chapter examines whether and how the responses to the framing processes of the COVID-19 pandemic in mainstream news media articles vary across the two aforementioned communities with varying degree of false and misleading content. To do so, it focuses on the content that each of these communities share in responses to the news media articles that they share with their community. Specifically, it compares the framing processes that are inferred in the previous step in each of these communities.

In these analyses, this work examines how each of these communities responds to framing in mainstream news media through practices such as *reinforcing*, *revising*, or completely *rejecting* the framing of this event in news media. By **reinforcing** a particular framing, this work refers to understanding the event in a similar way as it was framed in the news media. This form of framing evolution does not imply simply repeating the original framing verbatim, but rather framing the event in a way that maintains the same key issues, the same causes involved, and the same line of thoughts to the potential remedies. For instance, if news media frame the pandemic as a health crisis, and a community that comes to discuss such news also understands the pandemic in the same way, acknowledging the same causes and key entities involved, this response would be an example of reinforcing the original framing of news media. In the case of **revising** a framing, while the community that comes to discuss the news does not disagree with the news media framing of the same event, the community does not accept the original framing in its entirety. Rather, they might modify certain aspects of the original framing, and introduce new aspects of the issues not originally covered in the original framing. However, there is significant overlap on different aspects of framing, such as the major issues, and the causes to those issues, and the potential

remedies that should be considered. For instance, when a community come engage with news that frames the pandemic primarily as a health crisis, it might revise this framing by acknowledging not only the health implications but also the economic consequences, thereby broadening the scope of the issue. Finally, by **rejecting** a framing, this work refers to when a community disagrees with the way the event is portrayed in the news media, offering an alternative framing, which often directly contradicts the original framing. For example, refusing news media framing of the pandemic as a health crisis, and describing the event as a normal situation would be an example of rejecting news media framing of the pandemic.

There is indeed prior work on framing evolution (e.g., Snow et al., 1986; Benford and Snow, 2000), much of which has been conducted in the context of social movements. For example, Snow et al. (1986) identifies four categories of frame alignment: frame bridging, frame amplification, frame extension, and frame transformation. The three types of framing evolution discussed in this study (i.e., reinforcing, revising, rejecting) are indeed influenced by the aforementioned frame alignment processes. For instance, the concept of reinforcement shares similarities with frame amplification as described by Snow et al. (1986), which refers to the clarification and invigoration of an interpretive frame to encourage greater participation in an social movement. Similarly, in the context of frame reinforcement in this work, individuals may discuss an event within the original frame, making it clearer to themselves and their communities. However, as noted earlier, frame amplification and the other frame alignment processes are deeply embedded in the context of social movements. As a result, motivations for frame alignments differ from the motivations underlying frame evolution in this study.

5.3 Results

This section begins with reviewing the topics that emerged based on fitting the theta role model developed in paper 4 and on the corpus discussed in Section 5.2. Next, it describes the framing identified in each of corpora, following the approach described in 5.2.3. It starts by analyzing framings in evidenced in the news media sub-corpora, which serves as

the baseline for comparison in this paper. It then examines the framings observed in the two subreddits of interest, *r/science* and *r/conspiracy*, and explores how each community responds to the framings of the COVID-19 pandemic event in news media articles.

5.3.1 Topics Captured using Linked Latent Theta Role

The linked theta role trained on the discussed corpus captured four latent topics. This work analyzes each topic as follows. A topic in the linked theta role is defined by a) its probability distribution over words, b) the probability distribution for each word over grammatical relationships, and c) the other topic words that co-occur with the topic terms in a grammatical relationship. This section discusses a topic using the topic top words, alongside with the top co-occuring words that appear with the top word in the corpus. In addition, to better understand the topics, for each topic term and its co-occurring terms, example documents in which they occurred is also captured, provided that the document is a representative document for the topic (i.e., the document should have a high topic probability). Based on the aforementioned components, a high-level label is assigned to each topic to facilitate the discussion of each topic.

5.3.1.1 The spread of the COVID-19 virus

This topic, along with its associated terms and co-occurring terms, relates to the spread of the COVID-19 virus. Table 5.1 provides a portion of this topic.

5.3.1.2 The origin of the COVID-19 virus

This topic, including the topic words, as well as their co-occurring terms for the topic words pertain to the virus outbreak in Wuhan, Chian and the origin of COVID-19 virus and different. Table 5.2 illustrates a portion of this topic.

5.3.1.3 Public health and officials response to the pandemic

This topic, including the topic words, as well as their co-occurring terms for the topic words, and the context in which the topic terms appear relates to public health and the officials responses to the pandemic. Table 5.3 illustrates a portion of this topic.

Word	co-occurring term	Example
virus	spread	china's national health commission confirmed the virus can spread person-to-person, with patients in major cities like beijing and shanghai ...
	causes	experts warn coronavirus may cause 'wave' of neurological conditions including parkinson's disease ...
	spreading	but this isn't necessarily a sign that the virus itself is spreading throughout the body ...
	infected	more study may be able to reveal whether the virus first infected a small number of people ...
	uses	given that most brain cells lack the ace2 receptor the virus normally uses to break into cells ...
	emerged	in the four months since the virus emerged in the central chinese city of wuhan ...
	mutating	the new mutation makes the virus more likely to infect people ...
	found	the article says the deer caught it from humans, it simply means the virus has found a suitable host population ...
	caused	the virus has caused alarm because it is from the same family of viruses as sars ...
	...	...
people	infected	overall this will lead to more people dying in the event of overpopulation simply because more people will be infected ...
	vaccinated	6 million deaths and shrug as a few deaths meanwhile more people have been vaccinated ...
	died	that's in contrast to 1918, when many young people died ...
	tested	according to jones' portal, 81,269 people have tested positive for COVID-19 in florida since the beginning of march ...
	infect	source: "people infected with COVID-19 can still infect others after they stop feeling sick, so these measures should continue ...
	younger	for people younger than like 30 and this is why so many people want to wait longer before taking the vaccine ...
...		

Table 5.1: Topic 1, the spread of the COVID-19 virus. Note: this table only present part of this topic, to give an overview of the results, while ensuring concision.

Word	co-occurring term	Example
china	wuhan	the deadly disease in the book is named after the place it originated - wuhan in china,
	lab	in a letter published thursday in the journal science, they argue that there is not yet enough evidence to rule out the possibility that the sars-cov-2 virus escaped from a lab in china
	markets	a virus appeared in wildlife markets in southern china, and it was unlike any the world had seen ...
	scientists	using samples of the virus isolated from patients, scientists in china have determined the genetic code of the virus...
	trade	Photograph by Edwin Remsberg, VWPics/AP Rebecca Wong, assistant professor of sociology and behavioral sciences at the City University of Hong Kong, argues in her 2019 book about the illegal wildlife trade in China that consuming wildlife "is a common phenomenon in mainland China...
	...	...
wuhan	market	The spread of a deadly strain of coronavirus, sourced to a wildlife market in Wuhan and now a global epidemic, has thrust China's live wild animal trade into the spotlight...
	lab	daszak appeared to express gratitude to fauci for downplaying the theory that the COVID-19 was created in a lab in wuhan,...
	virus	8 report of the first case of pneumonia from an unknown virus in wuhan...
	outbreak	advertisement eric toner, a scientist at the johns hopkins center for health security, wasn't shocked when news of a mysterious coronavirus outbreak in wuhan, china, surfaced in early january...
	...	...
...		

Table 5.2: Topic 2, the origin of the COVID-19 virus. Note: this table only presents part of this topic, to give an overview of the results, while ensuring concision.

Word	co-occurring term	Example
public	england	the department of health and social care has mandated that all tests must be carried out by the nhs and public health england ...
	health	chan school of public health, said the retweet was damaging ...
	officials	americans have participated in elections during challenging times in the past, and based on the best information we have from public health officials, we are confident that voters in our states can safely and securely cast their ballots in this election...
	experts	as coronavirus cases rise nationwide, public health experts urge caution...
	service	the washington post is providing this news free to all readers as a public service...
	spaces	masks, which were already compulsory on public transport, in enclosed public spaces, and outdoors in paris in certain high-congestion areas around tourist sites...
	emergency	fema spokesperson lizzie litzow said the agency is currently focused on supporting the department of health and human services (hhs), which separately declared a "public health emergency" on jan...
	...	...
health	officials	when city health officials arrived at the residence to notify her, they realized she had lied about her identity ...
	department	the woman who said she was fired from the florida health department for refusing to alter coronavirus statistics is now publishing data on her own...
	experts	US underprepared for coronavirus due to Trump cuts, say health experts — Steps put in place after Ebola outbreak have been scrapped. ...
	secretary	fourteen people in britain tested for coronavirus, as health secretary says uk is prepared for virus ...
	workers	he defense production act, enacted in 1950, allows the president to force american businesses to produce materials in the national defense, such as ventilators and medical supplies for health care workers ...
	...	...
...		

Table 5.3: Topic 3, the officials reposes to the pandemic. Note: this table only presents part of this topic, to give an overview of the results, while ensuring concision.

5.3.1.4 Opinions on the COVID-19 Vaccine:

When reviewing the topic's top terms at first glance, this topic does not seem to capture any meaningful underlying theme, which is considered as junk topic [2] in prior work (Steyvers and Griffiths, 2007). However, reviewing the other component of this topic, including the co-occurring terms, and the example documents suggests that this topic mostly captures people opinions about vaccines, and whether or not they and others should take the COVID-19 vaccine. Table 5.4 depicts a portion of this topic.

5.3.2 Framings of the Pandemic in Mainstream News Media

The analysis suggests there are four main framings of the pandemic witnessed in the news media corpora. These framings do not necessarily correspond directly to the four topics identified in the model and may be based on one or multiple topics. The first framing,

[2] A junk topic is an "uninterpretable topic that picks out idiosyncratic word combinations" (Steyvers and Griffiths, 2007; AlSumait et al., 2009)

Word	co-occurring term	Example
people	died	He couldn't even handle that "Well it's not like a hundred thousand people died !" *Leans in to hear whisper from junior adviser. ...
	vaccinated	2.6 million deaths and shrug as a few deaths Meanwhile more people have been vaccinated than people have been infected and you have so few deaths that you can make a newspaper article for each one and that is scary? Smooth brains all around It looks like you shared an AMP link...
	die	Yes it sucks that people die from the vaccine but it's soooooooo much safer than getting COVID. ...
	infected	Nurse treating coronavirus sufferers in China claims 90,000 people have already been infected / I know it's dailymail , but it's an enormous post with full details and stuff , idk check it out (they agree with the conspiracy) ...
	want	Governments are gonna do what they want, regardless of what the people want, that much is clear...
	going	If people aren't going to listen or if people are just going to brush me off as "not knowing more than the average layman", what is the point? I know deep down I want to save lives, get my Masters in both nursing and Microbiology, and go on to study virology...
	...	...
get	vaccine	You're naive if you think you're simply not going to get a vaccine....
	people	If the vaccines pan out and we can somehow get most people vaccinated within 14 months we might be able to keep the casualties below a couple hundred thousand...
	sick	Plenty of people get a vaccinnee, get sick, and blame the vaccinnee because they got it recently...
	worse	And it will get much worse as long people are not united and they will not unite...
	get	You are right on that one, inhere they try to get all available doctors and nurses on the epidemic to get through the worst phase...
	...	...
...		

Table 5.4: Topic 4, focused on opinion around COVID-19 vaccines. This topic emerged mostly based on discussions around the COVID-19 news, and less in the original news. Note: this table only presents part of this topic, to give an overview of the results, while ensuring concision.

new territory framing, characterizes the pandemic as an unprecedented and entirely novel event. The second, **tragedy** framing, emphasizes the pandemic as both a health and economic crisis. The third, **urgency** framing, highlights the pandemic as an emergency requiring immediate responses. Finally, the **irresponsible** framing links the pandemic to irresponsible actions and reactions by various entities. These labels are used to facilitate the discussion, but as noted in Chapter 4, framing is more nuanced than can be captured by a single term. Each of these framings is explored in greater detail in the following sections.

5.3.2.1 New Territory

The news media articles frame the COVID-19 pandemic is as a "new territory", distinct from any previous event or experience encountered by society. The virus itself is described as "novel," with "much remaining unknown about the novel coronavirus". News media discusses the crisis's novelty from various perspectives. For example, discussions range from proposals for "extraordinary plans" to "place military commanders in control around America" to ensure the continuity of government (Topic 3), to concerns about "widespread domestic violence as a result of food shortages" (Topic 3), to calls for a temporary "global government" and a "coordinated global response" to address both the medical and economic crises triggered by the pandemic (Topic 3). Other discussions highlight how "family gatherings at Christmas holidays would pose substantial risks" and should be postponed (Topic 1). These discussions underscore factors that are unique to the pandemic, distinguished from previous events.

In framing the pandemic as a new territory, a variety of stakeholders are involved, from government officials and military commanders, to healthcare officials, to each individual across the globe. This framing implies that, to effectively address the various aspects of this unprecedented situation, the essential and the right line of response would be for different organizations and individuals to collaborate in combating the pandemic and do their parts. For instance, the news highlights how "go-it-alone approach" does not work to tackle this crisis, and "there has to be a coordinated global response" (Topic 3).

Through this framing, news media not only consider the need for collaborative efforts in combating the pandemic as the remedy to this tragedy, but also highlights the moral aspect

of such collaborative efforts. For instance, the news highlights how the government officials and public health officials, and countywide leaders continue to work while "keeping the public's best interest in mind". They highlights the importance of "working collaboratively as a community as the best approach to combating this virus." (Topic 2)

5.3.2.2 Tragedy

News media frames the pandemic as a tragedy, emphasizing the fatal aspect of this multifaceted crisis. For instance, the virus is discussed as "deadly" (Topic 1), which is "kills millions of people" (Topic 3). "Excessive human toll", "daily fatality tolls", "ongoing death tolls" are repeatedly emphasized (Topic 1), and number of "urns" being handled during this time window is reported and compared to normal time frame. In addition to the deaths numbers, infected cases are frequently reported, emphasizing the magnitude of the virus spread. Similarly, the death of many frontline health worker due to insufficient PPE (personal protective equipment) is highlighted (Topic 3).

The tragedy framing highlights the various ways in which people's lives have been disrupted, from personal well-being to daily routines. Example includes difficulties to shop for groceries (Topi 3), closer of schools for kids, people losing their jobs and depleting their savings, missing to go to funeral of their lived ones (Topic 1), among others.

Different entities, from the public health officials to the country leaders are involved to help address the global tragedy of the COVID-19 pandemic. In this framing, it is frequently emphasized how people lives should be the main focus and various organizations should work together to help tackle this tragedy. For instance, under this framing, news media highlights how during this tragedy "everybody should get the medical treatment they need regardless of their income" (Topic 3). [3].

[3]While there is not necessarily a one-to-one correspondence between topics and framings, this framing is most prominent in Topic 3, which pertains to official responses to the pandemic. Specifically, when reporting on government responses to the pandemic captured in this topic, officials emphasize the tragic aspects of the situation and justify their actions as efforts to mitigate this tragedy

5.3.2.3 Urgency

News media articles frame the pandemic as "urgent". For instance, the Department of Health and Human Services (HHS) declared the pandemic a public "health emergency" (Topic 1), and the World Health Organization declared it a "global health emergency" (Topic 1). Such emergencies demand rapid government action, utilizing all available resources (Topic 1). Additionally, President Trump declared the coronavirus a national emergency (Topic 1).

To address this urgent situation, calls for "immediate response" and "all hands on deck" were made (Topic 1). Specifically, this urgency needs full personnel commitment and the mobilization of all available resources to address the new challenges created by the new territory, the pandemic. This emergency requires the government as well as other organization involved, such as healthcare providers, to act quickly and with all tools available to them (Topic 2) In this context, President Trump allocated $50 billion in emergency funds (Topic 3). Align the same effort, "the Food and Drug Administration granted *emergency authorization* (emphasis added) to Moderna's coronavirus vaccine" (Topic 1). While this expedited approval process is rare, it was deemed necessary due to the extraordinary nature of the pandemic.

The news also emphasizes the extraordinary and urgent circumstances, that makes it a "must" for the senates and the president to approve financial supports needed by "state and local agencies on the front-lines" (e.g., healthcare workers) to enable them deal with this virus and its scope (Topic 3). Regarding urgency to provide medical supplies, the news draws an analogy between the outbreak and "war time," asserting that "we must mobilize as if it were a time of war, especially in relation to hospital beds."

5.3.2.4 Irresponsible

The news media articles also frame the pandemic and its different dimensions as the consequence of irresponsible actions, and reactions, taken by different entities. The initial outbreak of the virus is portrayed as an irresponsible response by the Chinese government, allowing the virus to reach across the globe and resulted in excessive death tolls (Topic 2).

Additionally, the Chinese government is held accountable for "mishandled the crisis", and allegedly covering up both " the coronavirus's spread" and "the outbreak's severity" (Topic 1). This irresponsible actions and reactions by the Chinese government are argued to have facilitated the further spread of the virus and exacerbated its consequences (Topic 1 and Topic 3). In addition, even after the pandemic "killed over three million worldwide in a year", "China still refuses to share critical information", which is argued to be clearly an irresponsible act by the Chinese government (Topic 3).

The irresponsible framing isn't limited to the Chinese government's response to the outbreak. For instance, the response from various entities within the U.S. government, across different political parties, also shifted attention away from the pandemic and downplayed its scale. The news media, for example, holds President Trump accountable for referring to the virus as "their new hoax" and for "misleading the American public about the threat posed by the COVID-19 virus" (Topic 3), framing this as an act of irresponsibility on his part. The news considers the president Trump response to the virus as "incompetent, political and reckless" (Topic 1), and argues he "downplay[s] the outbreak" to show "he has everything under control" (Topic 1). For instance, the news critiques how the president "downplayed the risks of COVID-19, comparing it to the common flu, which is not as dangerous as the novel coronavirus, and suggested that "one day it's like a miracle, it will disappear.". These "messaging and sending unclear signals to the general public during a national crisis" are considered irresponsible act, contributing to the pandemic magnitude (Topic 1).

In addition to Trump, the news also criticizes "his allies in the media" for being "busy trying to downplay the virus" and the danger associated with it (Topic 4). Similarly, democrats focus on impeaching President Trump is discussed as an irresponsible act, drawing attention away from the pandemic and allowing it to hit the United States (topic 4).

The framing of irresponsibility is further evident in discussions regarding the failure of the government and various organizations to adequately prepare for the new crisis. For example, the deaths of frontline healthcare workers, attributed to a lack of sufficient protective equipment, are framed as a direct consequence of the government's failure to ensure these individuals were properly protected, illustrating a broader neglect of responsibility and a lack of care. Similarly, the news also emphasizes the federal government's unpreparedness

to manage operations in the context of a new, unprecedented crisis. This unpreparedness is particularly emphasized in light of the lessons learned from the 9/11 incident, which prompted the government to pass new laws and establish procedures for handling emergencies and remote operations. Despite this precedent, the federal government's failure to prepare for a pandemic is framed as a clear indication of irresponsibility (Topic 3).

5.3.3 Framings of the pandemic in the r/science subreddit

This community present three primary framings of the pandemic. These framings include the **urgency** of the crisis, the association of this pandemic with **irresponsible** actions and reactions by various actors, and the characterization of the pandemic as an **information crisis**.

5.3.4 Tragedy

This community reinforces the tragedy framing in the news media (discussed in section 5.3.2.2). Specifically, similar to how the news media highlights the virus as being a deadly virus, and how it is killing millions of people, the "deaths numbers", and "infected cases" get shared occasionally in this community (Topic 1 and Topic 3), and the community also highlights "million cases" and "million killed" by the virus (Topic 2).

However, the tragedy framing is more implicit under the urgency framing in this community. That is, instead of elaborating the tragedy aspects of the pandemic, this community mostly focuses on the urgent need to find remedies for this tragic crisis and to learn about it for future such crisis, discussed in the following section. This shift in focus may stem from the community already recognizing the tragedy, and thus prioritizing discussions around actions that require immediate attention to mitigate the pandemic (Discussed further in the following section on *urgency* framing).

5.3.4.1 Urgency

The r/science community reinforces the way news media frames the pandemic as an urgent matter. This sense of urgency extends from the immediate need for the scientific community to identify the origin of the virus and study its behavior, for the governments to implement

effective measures to mitigate the spread of the virus, and for the scientist community to urge cures (i.e., vaccine) for the virus, and finally but equally importantly to the individual responsibility of taking the vaccine to help mitigate transmission.

The community also shares and discusses that scientists all around the world are"urging their colleagues to dig deeper into the origins of the coronavirus responsible for the global pandemic" (Topic 2). Understanding the origin of the virtus, this community argues, is not only urgent in order to help find remedies for the virus and inform us about how to mitigate the risk for the current pandemic, but also better prepare the world in case of encountering another future such pandemic (Topic 2).

This community highlights "million cases" and "million killed" by the virus (Topic 2), urging the need to find cures for this pandemic. They emphasize the urgent need for the scientific community to all get involved in pushing the development of the vaccine for this novel COVID-19 virus. For instance, while the mRNA vaccine tech is something that researchers have been working on for a while, "immediacy of the need for a coronavirus vaccine means it's "all hands on deck", ", pushed the development of an mRNA-based vaccine (topic 1).

The r/science community also calls for urgent response from the government officials to both implement measures, either mandate or otherwise [4], to mitigate the spread for current virus, and to prevent the positional future such outbreaks. With respect to the current outbreak, this community view it essential to push "mask mandate", "social distancing", and "vaccination" (Topic 1 and 4). Regarding the future such virus, given the high possibility of natural origin of the COVID-19 virus, this community emphasize the need for the rule makers to push to"shut down the wildlife market" (Topic 2).

In addition to the urgent response from the scientific community and officials, this community emphasizes the critical need for individuals to understand and adhere to guidelines designed to mitigate the spread of the pandemic. For example, some members argue that it is essential for people to "get vaccinated" (Topics 1 and 4). The community asserts that

[4]This community does not take a definitive stance on whether measures such as masking and vaccination should be mandated. While they generally view compliance with these measures as an individual responsibility, they recognize the reality of widespread non-compliance and therefore seem to implicitly support mandates as a means to mitigate the spread of the virus.

until a sufficient proportion of the population is vaccinated, it remains urgent for those who believe in the efficacy of vaccines to get vaccinated in order to help curb the spread of the virus. While there is uncertainty about the effectiveness of vaccine mandates, many in the community express a desire for such mandates to be implemented as a means of ensuring immediate vaccination and helping to mitigate the spread of the virus.

5.3.4.2 Irresponsible

The r/science community frames the pandemic as the result of irresponsible actions by various entities, reinforcing the irresponsible framing manifested in news media (Discussed in 5.3.2.4). For example, the outbreak is often linked to the Chinese government's failure to promptly inform the global community about the spread of the virus. Statements such as, 'China lost vital weeks in notifying the world about the pandemic,' reflect this view. In addition to concealing the virus outbreak, this community argues "Chinese government's initial response was poor", and believe the Chinese government acted too late, again, a sign of irresponsible management (e.e., "They lost several vital weeks telling people to shut up already") (Topic 2).

Similarly, the response of the U.S. government, particularly during the early stages of the pandemic, is characterized as irresponsible, contributing to the further spread of the virus and exacerbating its impacts. The repeated "downplaying of the pandemic's severity" and "diminishing the death toll" by U.S. officials is frequently cited as a key example of this negligence.

Beyond government actions, the failure of individuals to adhere to public health guidelines is also framed as irresponsible behaviors that exacerbated the spread and consequences of the virus. For instance, "the selfish choice to wait or to abstain from the vaccine" is characterized as an irresponsible act with consequences for the broader society (Topic 1). Similarly, resistance to vaccine mandates, often framed as an infringement on individuals' freedom, has been viewed by some as an irresponsible act.

In this framing of the pandemic, both government officials and citizens are urged to recognize their responsibility in addressing the pandemic and mitigating its effects. The government officials are urged to put all their resources to mitigate the spread, and indi-

viduals are urge to do their parts by following the guidance shared to mitigate the spread (e.g., “social distancing”, and “masking”).

5.3.4.3 Information Crisis

The pandemic is also framed as information crisis. The overwhelming volume of both “information and misinforma tion” (Topic 4) surrounding the COVID-19 pandemic has created significant challenges for the public in discerning the facts while they are making sense of the pandemic. For example, this community argues that ”There is so much conflicting info out there that people are going to ignore everything”, and ”This much of content make the doors open for conspiracy theories until people stop taking it seriously”.

This community also criticizes publishers, arguing that “everyone wants to publish on COVID-19” to make profit, resulting in “hastily published articles”, worked around the virus. For instance, scientists are critique about urging to publish around this matter, and see the pandemic as ”a way to publish [more]”, thus ”loose with these studies” (Topic 4). The community makes a call for the need to cautious on what gets published to ensure that the information being shared is accurate(Topic 2).

Some in the community argue that the information crisis cannot be addressed solely through regulation of media coverage, as it is such a complex problem and can be easily corrupted by “those in power”. Instead, they view it as a “cultural problem than something that legislation could tackle” (topic 4), in which every individual plays a role. The superficiality of many online discussions exacerbates this issue, as people often rely on headlines or brief snippets rather than engaging with the full complexity of topics, particularly during the pandemic. This leads to widespread misunderstanding, where exaggerated claims about unproven treatments or “miracle cures” gain traction and spread misinformation. Critics within the community condemn the “just read the title” culture, deeming it irresponsible (topic 4). They believe that individuals sharing content should provide “vital information” rather than superficial excerpts or fragments of articles. As a result, the responsibility for sharing reliable information extends beyond media outlets to everyone who posts or shares content online.

This framing is a revision of the “tragedy framing”, covered in news media (Section

5.3.2.2). While this community reinforces the tragedy framing depicted in news media, considering it as a health crisis and highlighting health implications associated with this crisis, it also emphasizes the pandemic as an information crisis. In this view, individuals and the community as a whole face significant challenges in reliably understanding the nature of the pandemic due to the widespread misinformation. The community asserts that combating the spread of misinformation requires a heightened sense of responsibility from media outlets, scientists, as well as individuals sharing content. This responsibility entails not only ensuring the accuracy of the information but also considering the broader social and personal impact of what is shared. Therefore, this community calls for a collective effort to prioritize transparency and accuracy, recognizing these as essential to mitigating the risks posed by misinformation during the critical period of the COVID-19 pandemic.

5.3.4.4 Framings of the pandemic in the Conspiracy subreddit

Results suggests three main framings of the pandemic. These framings include **intentionality**, **totalitarian**, and **denial** of the pandemic. The following sections describe each of these framings, and the way each might revise or reject certain framings of the same event in the news media articles that are shared in this community.

5.3.4.4.1 Intentionality This framing makes sense of the pandemic by calling it a *plandameic*, an event that was intentionally planned with certain interests and hidden agenda. There are different manifestation of intentionally framing that discusses the pandemic as an "engineered crisis" (Topic 4). Many people in this community claim that "Wuhan labs is 100% Chinese Communist Party," and they believe that the pandemic was designed as a bio-weapon "dedicated to vanquishing [China's] global rival." Many others argue that blaming china is an oversimplification of who is involved in this plan, and other nations, and mostly the US government, are involved to (e.g., "They [the US officials] want you to want war with China", topic 4, "Gene editing bioweapon COVID-19 to sterilize Chinese, Iranians, etc.", topic 2).

In addition to political motives for engineering the pandemic, this community also highlights business and financial interests (Topics 3 and 4) as justifications for the intentionality

framing. For example, some argue the virus, and the distributed tests kits following this event, are designed as a tool kit for the Gates's foundation "digital certificates" to track everyone and their information (Topic 3), and/or to profit from vaccines for this engineered virus. This manifestation of the intentionality framing is tied to existing conspiracy theories about the bill gates foundations (See (Goodman and Carmichael, 2020) for more details on this conspiracy theory).

While there are variations in the causes discussed for the intentionality framing of the pandemic as well as the actors involved, in all instances, the pandemic is conceptualized not as a natural event, but as a deliberate, engineered occurrence, orchestrated by those in positions of power. In this framing, a lack of concern by the governments and the elites for people's well-being and lives is implicitly conveyed. For instance, the portrayal of the pandemic as a deliberately collaborative effort to provoke conflict between nations and incite " world war" (Topic 2) reflects the perception within this community that those in power are indifferent to human welfare and the value of individuals' lives.

The intentionality framing provides a clear example of how the r/conspiracy community revise the framings of the pandemic in news media. Put precisely, this framing is a response to both *new territory* and *irresponsible* framings of the pandemic in news media articles. Specifically, by considering the virus as "engineered", this community pose a revision of how the pandemic is a health crisis that put the world in an entirely new area. In addition, with considering the pandemic as "planed", this community revise viewing the act of those in power in response to the pandemic as irresponsible, but rather as designer of this "engineered" pandemic.

5.3.4.4.2 Totalitarian One of the key reasons often cited as a motivation for the intentionality framing is the perceived expansion of government power and authority, a concept commonly referred to as the totalitarian framing. In this framing, the community believe that the government is using the pandemic as a tool to excessively extend its authority.

For instance, there are arguments that the pandemic is leveraged by the government to inform people the importance of allowing the government to control people and use oppressive methods (e.g., "authoritarian draconian lockdown measures") to save people lives (topic

3). Phrases such as "Totalitarian tiptoe", "corrupt elites", "rapidly boiling frogs"(topic 2), "corrupt global elite power structure" (Topic 4) , and "anti liberties and [anti] freedoms" (Topic 4) illustrate how this framing is linked to lack of trust in the government, authorities, and the elites connected to them. Note that there is some overlap between the intentionality and totalitarian framings. In these overleaping arguments, this community argue that the pandemic was deliberately engineered as a means to consolidate government power.

Not all instances of totalitarian framing, however, are based on the belief that the pandemic was deliberately engineered. There are certainly members in this community that do not view the pandemic as a purposeful or planned event, yet they still perceive *the government's response* to the pandemic as an overreach of "power and control", topic 0. For example, while many in this group acknowledge the reality of the virus, and accept that its origins remain unconfirmed, they interpret measures such as mask mandates and vaccination requirements, especially since "mandated", as signs of increasing government control and a gradual erosion of civil liberties (Topics 1 and 4).

Framing the pandemic as totalitarian and act of authority represents a revision of how the news media depicts the pandemic as a novel and unprecedented situation (i.e., new territory, Section 5.3.2.1). In this perspective, this community acknowledges the pandemic as a significant development, however, not because of the virus itself or its associated health consequences, but because of the ways in which governments are consolidating power and exerting greater control over the population, using the pandemic as a justification.

Both of these framing discussed above (i.e., intentionality framing and totalitarian framing) have evidence of skepticism and distrust in both science and authorities. Skepticism and distrust are very common among conspiracy theories, and studied in prior work as well (Phadke et al., 2021; Anderson and Rainie, 2020; Starbird et al., 2016). The results here show that these skepticism and distrust have taken to a great degree in the context of the COVID-19 pandemic, manifesting as complete denial of the pandemic (discussed in the following section).

5.3.4.4.3 Denial Framing This community also frames the pandemic as a hoax, either denying its existence altogether or downplaying its severity. Within this framing, many

people assert that "the so-called COVID-19 virus" is a fabrication, driven by "those who seel to profit, financially or politically" (Topic 4). Public officials, health authorities, and scientists are often portrayed as key actors in this narrative, with motives ranging from personal profit to exercises of power. For instance, many claim they know no one who is infected by the virus. Hence, a hoax. This rationality acknowledges prior work on how people try to understand events in light of what touches their lives (Gamson and Modigliani, 1989). However, in the case of the pandemic, were the virus hits different countries and even different counties disproportionally, people personal experience with the virus might be limited.

Others within the community acknowledge the virus's existence but argue that its lethality has been grossly exaggerated by the media and political figures, again with certain hidden goals. This group compare, and see the symptoms as equivalent to those of flu (topic 2). In this view, governments and health authorities are accused of inflating the crisis to further undisclosed objectives. For example, some claim that the reported "exaggerated COVID deaths" are inflated to present the pandemic as more severe than it is to implement "draconian methods" (Topic 1). Additionally, some question the timing of certain high-profile individuals having the virus, suggesting that these events are orchestrated to "fuel the panic" and "justify government actions" (Topics 1 and 4). In these narratives, individuals are often depicted as victims of elite manipulation, with the actions of those in power viewed as immoral and primarily serving their own interests rather than the public's well-being.

While both the intentionality framing and the totalitarian framing reflect a broader distrust of government officials and health authorities, the denial framing underscores the extent to which this distrust invoke individuals and communities to disregard the readily available evidence of the virus and instead adopt their own narratives, picturing the pandemic as a complete hoax.

Framing the pandemic as a hoax rejects both the tragedy and urgency framings of the pandemic in news media. More specifically, this community does not view the pandemic as a tragedy, and instead offers its own framing of the pandemic as a fabricated narrative (i.e., a hoax) that either never occurred or is vastly exaggerated compared to the representations

in mainstream media. Similarly, it rejects the pandemic as a matter that requires urgent response from the entities such as official government and healthcare authorities, mostly because this community believes the pandemic is just a fabricated narrative, which either never happened, or is as minor as flu virus and its dimensions are being exaggerated. This framing is also closely related to totalitarian framing, which views governmental responses to the pandemic as expressions of authoritarian control, rather than genuine concern for public health and societal well-being.

5.3.5 Differences in Framings Evidenced in Frequent Topic Terms and their Co-occurrence Terms Across Sub-Corpora: Illustrative Examples

An important component in the linked theta role model is the way it captures the co-occurring terms for each topic words. Co-occurring terms can signal how words that have high probability in a topic (i.e., topic terms) might co-occur with different terms across different sub-corpora. Thus, quantitative analysis of differences in co-occurring term at the topic level might indicate differences in how each topic is discussed across the sub-corpora. However, given that there is no one to one mapping between topics and framing, these analysis less readily highlights the differences in framing across the different sub-corpora.

Therefore, this section instead illustrate some examples of how specific topic terms, identified in the exploratory framing analysis in preceding section, are associated with different sets of co-occurring terms across sub-corpora. Specifically, it provides examples that illustrate the way a) certain topics terms might be emphasized less and more across different sub-corpora, and b) how these topic terms co-occur with different sets of topic terms across the sub-corpora. It is important to note that the analysis presented with these examples is not meant to identify framing differences across the sub-corpora for two main reasons. First, framing evidences are often interwoven across documents and cannot be simply by focus on individual examples. Second, as discussed in Chapter 4, framing analysis requires the examining the contexts in which such evidence appear, therefore, cannot be simply captured in single topic terms and their co-occurring words. Therefore, these examples serve to

supplement the qualitative analysis by providing concrete instances of how word choice can reflect the different ways framing manifests in public discussions of the same event across these different communities.

"Outbreak" is one of the major terms that is frequency used in news media when framing the pandemic as new territory, and tragedy. In this sub-corpora (i.e., news media), this term co-occur with terms including Wuhan, virus, eloba, began, started, global, declared, mysterious, highlighting the novelty of the pandemic (i.e., new territory framing), and its and its magnitude (i.e., tragedy framing). This term, however, is less used in the r/conspiracy community, and only co-occur with terms "wuhan", "new", and "epicenter", that is in line with how this community denies the pandemic and/or ignores its scope. Similarly, the news media emphasizes the term "state" with a variety of other topic terms (e.g., media, department, portal, cctv, xinhua, governments, actors, outles, election, vermont, emergency, Washington, government, local), highlighting the different states and entities dealing with the pandemic and its impacts (i.e., tragedy framing). However, the r/conspiracy community seem to ignore all these aspects of the news as this term only appear with a small subset of these co-occurring terms (e.g., media, department, government, portal), which is again in line with how this community denies the pandemic and its magnitude. News media consider the natural origin In framing the pandemic as a new territory, news media discusses the natural origin of the pandemic, where the word "market" appears, with a variety of other topic terms (e.g., animal, bats, wet, seafood, live, closed, originated, closed, signs). However, this term rarely appear in r/conspiracy community, which can be related to how this community indeed ignores the evidence around natural origin of the pandemic, considering the pandemic as a *plandemic* (i.e., intentionality framing).

There are indeed cases were the topic terms used in news media articles appear frequency in r/conspiracy discussions as well. However, these terms co-occur with different set of terms. For instance, the topic term "government" co-occur with a variety of terms in news media, including response, officials, continuity, succession, ornstein, spending, support, adviser, commission, report, chinese, federal, global, british, support, adviser, cripples, continuity, form, continuation, support, india, among others.

Some of these co-occuring words, such as censor, global, and rips are also used in

r/conspiracy community. However, this topic term also appear with new set of topic terms, such as "settled", "ordered", and "policy". The choice of terms in r/conspiracy community might be related to how this community discusses the conspiracy theories around how the government is censoring some, if not all, aspects of the pandemic to push their hidden agenda and gain more control and authorities over people. Similarly, the the topic terms "news" appears with news media with other topic terms such as conference, xinhua, reported, pictures, etc. However this term only co-occur with terms outlet and fake in r/conspiracy community. This selection of terms suggests how this community ignores all the news surrounding news reporting cases, death etc. and only talk about how news might be fake.

In the r/science community, however, topic terms co-occur with terms that are more similar to news media, but deviated from the choice of terms in r/conspiracy community. This is inline with how this community often reinforces, and in some case slightly revises the framing of the pandemic as manifested in the news media. For example, similar to how the topic term pandemic appears with terms such as "global", "epidemic", to highlight the scale of the pandemic and its impacts (i.e., tragedy framing), the r/science also uses similar words. In addition, similar to the news media framing, this community also discusses the pandemic is an irresponible act by different entities. This framing is evidenced in how topic term "pandemic" co-occurs with "irresponsible" in both of these sub-corpora.

Again, it is important to note these examples alone do not provide a comprehensive understanding of how these communities frame the pandemic. Rather, they serve as illustrations of how framing processes can be observed through the patterns of co-occurring terms.

5.4 Discussion

This chapter demonstrates the interconnections between misinformation as a societal phenomenon and the way people come to understand and make sense of events around them (i.e., framing). This section, building on the results described in the preceding section, discusses the way news media framing of the COVID-19 pandemic unfold in the two communi-

ties with different prevalence of false and misleading content, i.e., r/science and r/conspiracy, highlighting the differences in framing evolutions in these communities. By synthesizing across the presented results, this section shows that r/science more often reinforces, and in some case revises framings from news media. However, r/conspiracy, a polar opposite of r/science community in terms of prevalence of false and misleading content (Phadke et al., 2022), more often significantly revises or completely rejects news media framings (Section 5.4.1).

Next, the section outlines this chapter's implications for researchers within the misinformation domain in terms of the approach this work present to study misinformation beyond pieces of content, and for public health communicators in terms of how the content they create might be re-framed within different communities (Section **??**).

5.4.1 Framing Evolutions in Responses to News Media Framing of the COVID-19 Pandemic: Reinforcing, Revising, and Rejecting

The analysis of responses to the news around COVID-19 pandemic in the r/conspiracy and r/science demonstrate how framings of this event evolve in these communities in ways that **reinforce**, **revise**, or completely **reject** framings of this event to construct alternative interpretations of the same event. To reiterate, **the r/science community more often reinforces**, or slightly revises framing of the pandemic as discussed in news media articles, while **the r/conspiracy community more often rejects** or significantly revises those framing (See Figure 5.1). Drawing on existing literature in the domain of misinformation, this section explores potential reasons that may have influenced framing evolutions within these communities, as well as the differences in the processes underlying framing evolution across these communities [5].

As shown in this figure, the r/science community reinforces framing the pandemic as a tragedy discussed in the news media (i.e., tragedy framing), as well as the urgent need to implement measures to help mitigate the tragedy (i.e., urgency framing). In reinforcing both these framings, this community not only highlights the evidence from the news articles, but

[5] As noted in the introduction of this chapter, given the observational nature of the conducted study, the work presented in this chapter avoids from making any casual claims about the factors involved and their specific roles.

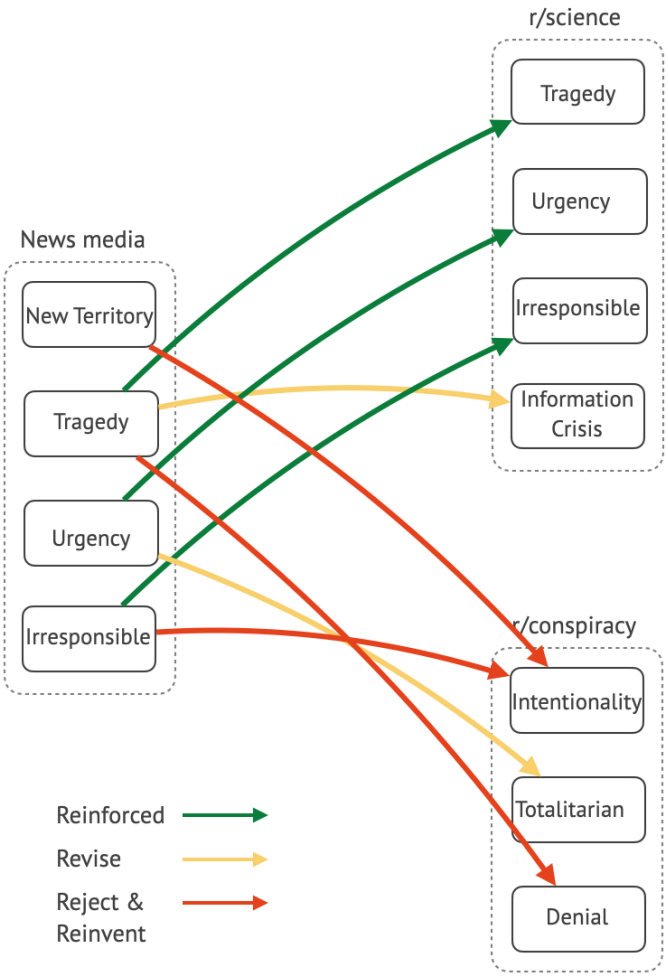


Figure 5.1: r/science community mainly reinforces the framings from news media, and in one case revise the original framing. However, the r/conspiracy community more often rejects the news media framing and offer their own framing of the pandemic.

also brings evidence from other relevant resources (e.g., scientific journals). This pattern is inline with prior work about how it is normative for this community to seek and share scientific information when making sense of world's events (Kaiser et al., 2023; Jones et al., 2019; Hara et al., 2019), such as the COVID-19 pandemic.

The r/science community not only reinforces the "tragedy" framing but also expands upon it, conceptualizing the pandemic not merely as a health and/or economic crisis, but also as an *information crisis*. The information crisis framing encompasses both a) the overwhelming deluge of content (including accurate and misleading content) about the pandemic, and b) concerns about the reliability of that information. Framing the pandemic as an information crisis in r/science suggests that this community, while acknowledging that

people need to understand the pandemic, emphasizes the importance of reliable information, trustworthy publications, and individual accountability in ensuring the reliability of content before being sharing it with others.

Within this framing, this community asks members (although comparatively less in numbers, as suggested by the results) who share opinions based unreliable or incomplete information (e.g., solely an article's title), to be account for the potential consequences of the spread of unreliable content, particularly in terms of how such content may overwhelm others as they try to understand the emerging event of COVID-19 pandemic, and may make people more vulnerable to misleading understandings about the pandemic, or even to fall for conspiracy theories. These discussions suggests that while the community enforces rules to maintain these norms, community's rules do not seem to be sufficient to achieve this goal, and community members themselves socially engage to maintain these norms. This speculation is indeed in line with the work presented in Chapter 3 that suggests rules alone may not influence people perceptions of a community norms as much as community responses do (Aghajari et al., 2023c).

The **r/conspiracy** community responses to news media articles about the pandemic shows a fundamentally different evolution of framings. As stated earlier, this community more often rejects the framing of the pandemic as depicted in news media articles, and shifts the interpretation of the same event that are depicted in news media, in drastically divergent directions (See Figure 5.1). For example, r/conspiracy rejects the new territory framing of the pandemic (5.3.2.1), considering the pandemic not as a novel, sudden natural crisis, but as an engineered, coordinated efforts enforced by those in power (i.e., intentionality framing, Section 5.3.4.4.1). This community also significantly revises framing the pandemic as an urgent matter, which news discusses to require coordinated efforts and urgent responses from government, healthcare officials, and citizens to mitigate the virus's spread and protect public health (Section 5.3.2.3). Instead, it considers the pandemic as an expansion of government power and authority (i.e., totalitarian framing) (Section 5.3.4.4.2), which, this community argues, is pushed through practices such as masking mandates, and lockdown.

Neither of these framings of the pandemic and the way they have evolved are drawn only from individual pieces of content. Instead, the framing evidence in results suggests that

the community's mistrust in government, health authorities, and the elite, as well as prior conspiracy theories, have given rise to these shifted framings of the pandemic. For example, in rejecting the new territory framing of the pandemic, depicted in news media, and coming to understand the pandemic as an intentionally planned event, the r/conspiracy community does not refer to individual pieces of evidence about the pandemic itself. Instead, they draw upon pre-existing conspiracy theories (e.g., those concerning the Gates Foundation and its purported hidden agenda) and emphasize the perceived untrustworthiness of government agencies, healthcare authorities, and elites. This community believes these entities disregard public well-being and have orchestrated the pandemic to advance their hidden motives. Put differently, this community's understanding of the pandemic is informed by their "distrust" of involved entities (i.e., government agencies and health authorities) and "existing conspiracy theories", where they used these factors as "evidence" when they come to understand the pandemic as an intentionally planned event, or deny its existence.

Both these patterns of drawing on existing conspiracy theories and highlighting distrust in authorities to validate new ones are indeed acknowledged in prior work as well. In terms of the influence of prior conspiracy theories, for instance, Gagliardi (2023) similarly argues that, in many conspiracy theories, individuals often do not rely on factual evidence to support new claims. Instead, they reference prior conspiracy narratives as a form of validation for their distrust of authorities, using these narratives to rationalize the adoption of new theories. Therefore, even with an infinite amount of correct information, Rabin and Schrag (1999) argue, people can still get convinced of their misconceptions, as many people tend to read the evidence in a way that confirms their prior beliefs.

Similarly, the role of distrust in authorities a strong driver for the formation of framings that are deviated from the reality in r/conspiracy (e.g., intentionality, and denial framings) aligns with prior work about the connections between distrust in authorities and conspiratorial beliefs (Phadke et al., 2021; Gagliardi, 2023; Anderson and Rainie, 2020). Further, this study suggests distrust in authorities contributes to the evolution of framings that deviate from reality, and already existing conspiracy theories (e.g., theories concerning the Gates Foundation) serve both as motives for these shifted framings, and as validations of these framings. Put differently, these results suggests a potential feedback loop between distrust

in authorities, and conspiratorial beliefs.

This work expands our understanding of the interplay between distrust in authorities and already existing conspiracy theories, particularly in terms of their combined influence on the re-framing of events in ways that deviate from reality. While the nuanced analysis of the r/conspiracy community's interpretation of the pandemic provides valuable insights, the observational nature of this study precludes definitive causal claims. Several hypotheses emerge: Does framing divergence stem directly from distrust in authorities? Is it driven by already existing conspiratorial beliefs? Or does it arise from the interaction between these factors? These hypotheses, informed based on the analysis conducted in this chapter, warrant further investigation, potentially through experimental research, to elucidate the nature and dynamics of the relationships involved.

These divergent framings within a community can significantly influence people's understanding of reality, extending beyond misleading them about individual pieces of content items. This influence, in turn, can negatively impact their responses to surrounding events. For instance, rejecting the framing of the pandemic as a tragedy and instead reinventing the denial framing (i.e., the pandemic is a hoax or a tool used by those in power to control society) can impact how many people perceive the government responses to the pandemic, such as lockdown and enforcing mask mandates, thereby negatively influencing their own responses to these implemented measures. For example, in the case of COVID-19 vaccines, misconceptions that denial framing and/or totalitarian framing contribute to might increase vaccine hesitancy within the community (Ullah et al., 2021; Kreps et al., 2021), and beyond (Ghosh et al., 2024; Abdallah and Lee, 2021; Taylor et al., 2016).

To reiterate, understanding framing evolution as a broader impact of misinformation, which can shape and in times create misperception of reality and influence society beyond misleading individuals about pieces of content, would not be possible to gain, and necessitates an ecological approach to misinformation research.

5.4.2 Implications

The chapter makes an important implication for community moderators to expand their moderation practices beyond individual content moderation. Specifically, community mod-

erators can utilize the approach taken in this chapter to proactively identify and mitigate potential misleading framings within their communities, and address the broader impacts of misinformation on their communities come to understand events around them.

Indeed, the analysis conducted in this chapter suggests that, much like the misleading framings of the COVID-19 pandemic and the misperception they contribute to are not the results of any individual pieces of content, they are unlikely to be simply addressed by only accounting for accuracy of news content, or providing corrective content with an individualistic perspective (in line with the argument made by (Aghajari et al., 2023b) about broader scope of misinformation). To address the evolution of framings that evolved using misleading evidence, instead, community moderators can utilize the approach taken in this chapter and examine the processes by which their communities come to frame world's events (e.g., presidential elections, ongoing global conflicts, or policies concerning minority groups), and attempt to mitigate the evolution and spread of misleading framings.

For example, in addressing framings that are evolved based on misleading evidence, which might not be necessarily factually incorrect when examined individually but utilized to present misleading framings, instead of providing alternative facts (an individualized intervention reviewed in Chapter 2), community moderators might consider offering alternative *interpretive packages* (Gamson and Modigliani, 1989). Such efforts might not only prevent misleading framings from becoming dominate framings within their communities, it can also mitigate the impacts of these framings on vulnerable individuals who are seeking to make sense of events around them. Given the important role of community moderators on and how their communities run and their influence on community members (Seering et al., 2017; Cullen and Kairam, 2022), framings that these stakeholders provide is likely to be well-received by their communities.

In this effort, community moderators can account for the role of community members and their responses in influencing other members' view and design around their influence (Aghajari et al., 2023a; Lo, 2018). For example, community moderators can target community members whose framings of the event is based on actual evidence that pertain to the event under discussion, and the ones who attempt to contribute to addressing misperception based on their prior interactions within the community. More specifically, they can

encourage such members to contribute to interpretive packages by which their communities understand events around them, and to promote leading narratives about the ongoing events (e.g., the tragedy framing and urgent framings). Encouraging community members to join this effort can potentially influence dominant framings of the ongoing event and affect the framings other members encounter within their community, thereby helping to mitigate shifts in framing that are impacted by misinformation [6].

5.5 Contributions and Future Work Directions

This chapter makes two key contributions in advancing research on misinformation. **First**, by employing the concept of framing and examining the functions it performs, this chapter empirically demonstrates that misinformation plays a role in the way online communities come to interpret and understand events around them (i.e., framing). Specifically, the findings from this chapter expands knowledge around the role of misinformation in the overall shifts in the processes of meaning constructions and people's understanding about the world's events, a broad impacts that exceed beyond misleading individual about individual pieces of content. Put concisely, this chapter shows in a community wherein false and misleading content is prevalent, framing evolutions occur in ways that more often reject the way the framing of the same event in mainstream news media, reinventing alternative framings, often contradicting the mainstream news media framing of the same event.

To reiterate, none of these alternative framing arise from any individual piece of content, but are instead motivated and shaped by the broader scope of misinformation (e.g., the way people perceive the different entities involved, and the way they selectively gather their evidence). Therefore, these impacts of misinformation would not be understood with an individualistic focus on misinformation, and can only be addressed when conceiving of misinformation as a broader, societal phenomenon that impacts online communities beyond misleading them about pieces of content. This understanding of the broader impacts of misinformation can inform the development of interventions aimed at mitigating the

[6] It is important to note that such interventions are less likely to work in communities that are invested in conspiracy beliefs Aghajari et al. (2023b). However, these interventions can potentially help mitigate framing evolutions that are creating misperceptions about the reality in communities that are vulnerable to misinformation, but value the truth, seek to understand the events around them.

community-level impacts of misinformation on the processes involved in meaning constructions (discussed in the implication section).

Second, this chapter contributes to approaches of studying misinformation with an ecological perspective. Specifically, it concretely designs, and develops a novel approach for studying misinformation as a societal phenomenon, extending beyond an individualistic focus on misinformation. The presented approach directly responds to the call made in the systematic literature review in Chapter 2, which highlights the need for approaches that address the broader scope of misinformation, beyond isolated pieces of content. Put concisely, this approach embraces the concept of framing from sociological research (Gamson, 1989; Scheufele, 1999; Benford and Snow, 2000; Druckman, 2001) (motivated in chapter 4), utilizes the computational technique designed and developed in chapter 4 and analyzes framing as processes by which people come to understand an event. Using the COVID-19 pandemic as a case study, this chapter examines the way misinformation contributes to understanding around this major global event, and impacts people's understanding about it, beyond misleading them about individual pieces of content. Put differently, this study enables us to understand and account for the societal-level impact of misinformation, which would be impossible to attend to via an individualistic approach to this phenomenon.

Indeed, the methods explored in this work can be applied not only to research on misinformation and its broader impacts, but can also be adapted to analyze related phenomena, such as disinformation dynamics, the spread of rumors, and the role of deep stories in facilitating such rumors (e.g., Starbird et al., 2019; Prochaska et al., 2023). For example, Starbird et al. (2019) argue that *disinformation* is indeed a collaborative work distributed across media ecosystem (Starbird et al., 2019), which contributes to the construction of a misleading version of reality Prochaska et al. (2023); Starbird et al. (2019). To examine the dynamic through which disinformation functions, Prochaska et al. (2023) captures *deep stories* (Polletta and Callahan, 2019) using a qualitative coding approach, and demonstrates how disinformation might mobilize online audiences. Given how these processes are not necessarily evidence in single documents, and the fact that researchers might not be aware of all deep stories at play, qualitative coding for these stories can be overwhelming, if not impossible in some cases (e.g., in contexts that there is no prior knowledge about

potential deep stories at play). This work suggests the methods employed in this work, including leveraging framing and the developed computational model, can provide a means to examine disinformation dynamics, and potentially address the aforementioned challenges.

This study has one primary limitation, which stems from the challenges associated with accessing and analyzing data from online communities. Specifically, given the challenges involved in accessing online communities' data and recent changes in scraping such data (Luscombe et al., 2022; Dogucu and Çetinkaya-Rundel, 2021; Ciani Sciolla, 2023), this chapter does not tease apart how responses to framing in news media might vary in the way individuals post about an event, and the way they come to discuss it as a community. This research, however, encourages future work to explore such potential distinctions, if and when such data can be accessible. Investigating these distinctions could potentially offer valuable insights into where social media platforms and online community moderators should focus their efforts to address the evolution of framing that contributes to misperception of major events, such as the COVID-19 pandemic.

Chapter 6

Conclusion and Overall Contributions

Misinformation plays a significant role on people's lives, affecting the way they come to understand the world's events, their interpersonal interactions, perceptions of social norms and acceptable behaviors within a community, and ultimately, the trajectory of societal evolution. This dissertation demonstrates that misinformation and its impacts encompass factors beyond individual pieces of content (Chapter 2). In particular, misinformation involves community oriented mechanisms, such as the role of social norms, the processes by which communities come to collectively make sense of events, and majority illusion effects, among others. However, interventions aimed to address this phenomenon merely focus on addressing misinformation as individual pieces of false and misleading content. Focusing on perceived social norms as a key community-oriented mechanism involved in misinformation, this dissertation demonstrates that these factors both contribute to and are influenced by the broader impacts of misinformation (Chapter 3). Based on knowledge gained in these two aforementioned studies, this dissertation argues that to better understand and address the broader scope of misinformation requires to study misinformation as a societal phenomenon, which transcends any isolated, individual pieces of content. Put precisely, it posits that there is a need to **adopt an ecological perspective to the phenomenon of misinformation** that enables considering the various elements within the information

ecosystem that influence people's understanding of the world's events, and study misinformation as the interaction among false and misleading content, community response, and the processes of meaning constructions. With this goal, embracing the concept of framing from sociological research (Gamson, 1989; Scheufele, 1999; Kirdemir et al., 2021; Benford and Snow, 2000; Druckman, 2001), this dissertation presents an approach that enables examining the interplay between false and misleading content, community responses, and the way communities come to understand and respond to events around them, and study misinformation with an ecological approach (Chapters 4, and 5).

This final chapter summarizes the main implications from this dissertation for researchers, community moderators, and social media designers. Specifically, it elaborates on how the approach taken in this dissertation, i.e., conceiving of misinformation as a broad phenomenon that transcends individual pieces of false or misleading content, provides opportunities for researchers to expand knowledge around this societal phenomenon, for social media designers, to develop interventions that address the broader scope and impacts of misinformation, and for community moderators, to monitor framing processes and employ moderation practices that transcend individual content moderation. Put concisely, this chapter highlights how the ecological perspective to misinformation, as taken and advocated for in this dissertation, can enable these stakeholders to better understand the broader scope of misinformation and to incorporate this understanding into the design of interventions to address this societal phenomenon and its impacts.

6.1 Researchers: Conceiving of Misinformation as a Societal Phenomenon

Prior work maintains a primarily individualistic focus on misinformation, conceptualizing it as pieces of false and misleading content that performs in isolation (e.g., Lazer et al., 2017, 2018; Wardle et al., 2018). With this view of misinformation, numerous approaches are designed to address this presumably isolated phenomenon (i.e., fact-checkers, signaling credibility of the content, signaling credibility of the source of content, and providing more perspectives on the content) (reviewed in (Aghajari et al., 2023b)).

This dissertation, however, argues that researchers need to conceive of misinformation as a broad phenomenon that transcends individual pieces of false or misleading content. In this view, misinformation is not an isolated phenomenon. Rather, it is situated in a complex information ecosystem where multiple factors interact and influence individuals' understandings of the world's events, thereby impacting their responses to those events. The impacts of misinformation similarly extend beyond misleading individuals about individual pieces of content. For example, conducting an experimental study in Chapter 3, this dissertation empirically demonstrates how misinformation impacts perceptions about norms in online communities, and can lead to broader consequences (e.g., shifts in perceived expectations and acceptable behaviors regarding sharing misleading theories). Similarly, conducting an observational study in Chapter 5, this dissertation demonstrates the broader impacts of misinformation on the way people come to understand and make sene of events around them (i.e., framing) (Gamson and Modigliani, 1989; Gamson, 1989). This study demonstrates that these broader impacts do not stem from any isolated piece of content. Rather, they pertain to social factors within the information ecosystem, such as community members' perceptions, biases, and trust in the entities associated with the events they are interpreting and making sense of.

Therefore, grounded in the insights from the aforementioned studies, this dissertation encourages researchers to similarly incorporate ecological approaches to misinformation, and study and address misinformation with practices that extend beyond isolated, individual pieces of false and misleading content. The remainder of the chapter will offer some guidance on how researchers can move towards this direction, and examine misinformation as phenomenon broader than individual pieces of content.

Admittedly, viewing misinformation as a societal phenomenon using an ecological approach involves considering the complexities of the information ecosystem and the various factors that play a role in this ecosystem. However, we do not need to start this effort from scratch. There is indeed a wealth of valuable insights from the psychological and social literature on the factors influencing how people process information, which may similarly play a role in the context of online misinformation and its societal impacts. The literature review presented in Chapter 2 outlines a variety of these factors, and particularly highlights the

important, yet under-explored community oriented mechanisms involved in misinformation (Aghajari et al., 2023b). Despite the role of these factors, there remains a significant gap in understanding how to effectively incorporate them into interventions designed to address the broader scope and impacts of misinformation. Researchers should therefore benefit from this extensive body of work to expand knowledge around the mechanisms by which these factors play out in the context of online misinformation.

To do so, this dissertations acknowledges the benefits and relevance of both *experimental*, and *observational* studies, and argues that these two approaches, to be effective, should inform one another. Specifically, informed by observations about potential mechanisms in information ecosystem that can be involved in the scope of misinformation, researchers can design and conduct controlled *experimental studies* to help expand knowledge around the mechanisms by which these factors function in the context of online misinformation. In this line, the study conducted in Chapter 3 designs and conducts experimental studies and offers insights into how norms are influenced by misinformation and identifies community responses as critical factors that can mitigate the broader, community-level impacts of misinformation. Indeed, social norms are not the only community oriented mechanism involved in misinformation. Numerous other community-oriented mechanisms may contribute to the spread and impacts of misinformation, which warrant future research attention (See Chapter 2 for an outline of these factors).

Admittedly, we cannot attend to all the complexities and nuances involved in the dynamics of misinformation in the bound of experimental settings. Thus, researchers are encouraged to leverage the insights from experimental research about the factors involved in misinformation, and examine the dynamics by which such factors my operate based on authentic online interactions in *observational study* settings. For example, informed by the results from the conducted experiment in Chapter 3 about the role of community responses, this dissertation designs and conducts an observational study, which leverages the concept of framing, and examines the broader scope of misinformation in terms of its impacts on the way people come to understand and respond to events around them. In the context of this dissertation, the approach taken in this study made it possible to explore how misinformation might manifest as broader shifts in how people interpret and view the world around

them, and play a role in processes by which framing evolves. Similarly, other researchers can leverage the approach taken in this study to examine how misinformation might contribute to people's understanding around other events that are targeted by misinformation, such as presidential election, or ongoing global conflicts, and expand knowledge around the broader scope of misinformation as a societal phenomenon in observational settings. In addition, researchers can employ this approach to attend to the potential driving factors involved in framings (e.g., distrust in authorities, prior beliefs, influencers' impacts, selective bias). Indeed, such settings cannot confirm causal relationships about the driving factors. Yet, they can provide rich understandings about the way these factors might interact within complex, real-world context, and also inform hypothesis for the design of future experimental studies to tease apart the nature and directions of the relationships between intentional factors involved. Furthermore, this dissertation acknowledges that while the concept of framing has been effectively employed in this dissertation to study misinformation from an ecological perspective, this concept and the approach taken in this dissertation may not represent the only possible solution. Researchers should explore other such concepts, as detailed in Chapter 2, that could similarly contribute to adopting an ecological approach to misinformation, and expand knowledge around misinformation as a broader, societal phenomenon.

The next section describes a key community-oriented aspect examined in this dissertation, which future research can utilize in both experimental and observational studies to study broader scope of misinformation.

6.1.1 Community Responses: A Key element for Understanding Community-Oriented Misinformation in Online Contexts

A significant challenge in studying community-oriented mechanisms involved in misinformation is the limited research identifying which community elements are most relevant and effective for integration into such studies. This dissertation provides valuable insights about community responses as a key community element, enabling the design of studies that expand our understanding of community mechanisms involved in misinformation, applicable to both experimental and observational approaches.

This dissertation identifies community responses as a key element that can be effectively

leveraged in designing experimental studies aimed at understanding community-oriented mechanisms of misinformation. Specifically, in an experimental study examining how misinformation influences perceptions of online community norms (Chapter 2), this study demonstrates the significant role of community responses to shared content in shaping norm perceptions. In this context, community responses exert a greater influence on perceptions about community's norms than the content itself. To investigate other community-oriented mechanisms (e.g., the majority illusion, pluralistic ignorance, peer influence reviewed in Chapter 2) using experimental approaches, researchers can similarly design studies that manipulate community responses. While this community element is promising, manipulation checks are essential to validate that observed effects are indeed occurred or mediated by community responses.

Similarly, this dissertation demonstrates the utility of community responses as a rich, informative community element to focus on in observational studies aimed at understanding the dynamics of community oriented mechanisms involved in misinformation. Specifically, to examine how misinformation is involved in the way people frame the world's events, and examine the dynamics and nuances involved in the interplay between misinformation and framing processes, this dissertation utilizes and demonstrates the significant of community responses in an observational study (Chapter 5). This study analyzes community responses based on both the posts shared by community members and the discussions that emerge in the comment sections, and reveals that the discussions provide a richer, more nuanced understanding of how a community come to understand an event (i.e., framing) and respond to it. Put differently, community responses offer a key means to understand how online communities evolve their understanding of global events, thereby respond to those events. Similarly, to observing other dynamics that misinformation might be involved in and/or contribute to within online communities, such as the spread of rumors, formation of beliefs in new conspiracy theories, evolutions of communities' norms and shifts in acceptable behaviors within a community, among others, community responses provide a valuable means for uncovering the underlying processes.

6.2 Social Media Designer, and Community Moderators

Conceptualizing misinformation as a broader societal phenomenon and adopting an ecological approach to misinformation, this dissertation offers insights into how social media designers, together with community moderators, can attend to the broader scope and impacts of misinformation.

This section first outlines how the approach presented in this dissertation, which focuses on the interplay between misinformation and framing processes, enables community moderators to adopt an ecological approach to misinformation, thereby facilitating moderation practices that transcend individualistic content moderation (Section 6.2.1.1). Next, it discusses the implications for social media designers in developing computational techniques that enable community moderators to implement these moderation practices to address misinformation impacts beyond pieces of content (Section 6.2.1.2). Finally, this section underscores the vital role of community members in the evolution of online communities, both in terms of the evolution of their communities' norms and the processes of meaning construction within their communities. This dissertation therefore encourages that social media designers to design and develop tools that might enable community members with actionable practices to address the broader scope of misinformation and mitigate its broader impacts within their communities (Section 6.2.2).

6.2.1 Entanglement of Framing and Misinformation: A Pathway for Adopting an Ecological Approach to Misinformation as a Societal Phenomenon

6.2.1.1 Monitoring Framing: Community Moderation beyond Content Moderation

Community moderators employ different strategies, and engage regularly with their communities to maintain the health of their communities (Seering et al., 2019; Gillespie, 2018; Gerrard, 2018; Jhaver et al., 2018). However, all these moderation practices are focused on individual incidents, and in the case of misinformation, focused on individual pieces of false

and misleading content.

This dissertation suggests that the approach that it presents to examine the interplay between misinformation and framing processes can similarly enable community moderators to address misinformation within their communities beyond addressing individual pieces of content. Specifically, community moderators, who play a great role in how their communities run (Seering et al., 2019; Jhaver et al., 2018)), can monitor the way their communities interpret and come to understand the world's events. By monitoring framing processes, community moderators can then engage with their communities to mitigate the evolution of framings that are based on misleading evidence, thereby mitigating the spread and impacts of misperceptions about events around them. In this way, community moderators can address misinformation and its broader impacts within their communities, effectively extending their moderation practices beyond an individualistic focus on pieces of content.

As illustrated in Chapter 5, the evolution of framings, and misperceptions that such framings contribute to and/or amplify, are not the result of only individual pieces of content. Rather, these processes are also influenced by "misleading evidence" and the way such evidence are used to create misleading narratives about the world's events. Specifically, misleading evidence, in which distrust in authorities plays a significant role, contributes to divergent framings of an event, that can further contribute to amplifying these misperceptions. For example, as detailed in Chapter 5, framing the pandemic as a "plandemic", or denial of the virus's existence, are not derived only from any individual piece or pieces of content, but rather from collective perceptions of involved entities, particularly distrust in these entities, and personal narratives that are treated as evidentiary when interpreting news about the pandemic. Thus, these misleading narratives, and misperceptions that these narratives might contribute to, cannot be simply addressed with individualistic focus on false and misleading content.

Instead, to address these broader scope of misinformation, community moderators can expand their moderation strategies and examine how their communities come to frame an event. Specifically, they can monitor potential deviations in evolution of framings that are evolved based on misperceived evidence and might further amplify such misperceptions. These strategies can enable community moderators to address the root causes of misperceptions

tions that are involved in development of misleading framings. For example, these strategies might enable community moderators to identify "misleading evidence" that is used in evolution of these deviated framings (e.g., referring to the US financial support for research in Wuhan lab as an "evidence" for U.S.-China collaborative effort to create bioweapons, or the emergency authorization of vaccines as a evidence that indicates inherent safety issues), and/or the way such "misleading evidence" or even "factual evidence" might be presented to validate and amplify misleading framings. Put concisely, understanding framing processes can provide insights about the factors that might be involved in shaping misperception about the world's events (e.g. evidence that makes sense within the context, but is not definitive (Klein et al., 2007), perceived reality that is deviated from the reality (Klein et al., 2007; Nickerson, 1998; Sleegers et al., 2019)), which might not be necessarily factually incorrect if investigated individually. In these cases, community moderators can engage with their communities and offering alternative framings based on actual evince and guide framing evolutions within their communities. Indeed, prior work suggests these moderators often serve as role models, and their linguistic patterns are imitated at higher rates than those of other community members (Seering et al., 2017). Thus, the framings they feed into their communities are likely to be similarly well-received by members within the community, and mitigate the spread of misleading framings to be the dominant framings.

Monitoring framings as a moderation practice, even with the assistance of computational tools, can arguably be more effortful than existing moderation practices (i.e., attending to individual incidents). However, prior work suggests community moderators are willing to invest such an effort to protect their communities from harmful behaviors (Seering et al., 2019). This commitment and willingness to engage directly is further illustrated by moderators' approach to algorithmic moderation tools. For example, although algorithmic tools (e.g., "AutoModerator", "AutoMod") exist for automated content moderation, recent study shows that moderators primarily rely on these tools for clearly objectionable content, such as malware or pornography links (Seering et al., 2019). For potentially controversial threads, community moderators prefer to personally review and intervene, demonstrating a desire to maintain nuanced control over their community's trajectory. These insights suggest that community moderators acknowledge the complexity of social interaction, as well as the

potential harm and consequences of misclassified cases when using automatic tools to moderate these interactions. Consequently, they prioritize direct engagement to address harmful interactions, even when it demands greater effort. Given community moderators value the the health of their communities and their notable commitment and efforts to moderate their communities suggested by prior work, they are likely willing to monitor framing evolution to mitigate the evolution of misleading framings and their spread within their communities.

However, given that framings manifest across multiple interconnected conversations within online communities, manual examination of framing evidence, as detailed in Chapter 4, becomes increasingly challenging, if not infeasible in larger communities. Thus, the effective monitoring of framing processes requires the use of appropriate tools facilitate framing analysis. The subsequent section explores how the computational model developed in this dissertation can empower moderators to monitor these framing processes. Furthermore, it provides insights for the development of future, new tools for this purpose, outlining key considerations for their design.

Admittedly, monitoring framing as a moderation practice may be less relevant in communities where misleading evidence is more readily accepted (e.g., r/conspiracy, r/AlternativeHealth). In these communities, misperceptions are often more closely related to individuals' existing beliefs rather than to misleading evidences, and faulty reasoning. Furthermore, given that moderators are often members of the community themselves, in these communities moderators may be less likely to recognize misleading framings as such. That said, monitoring framings can still be beneficial for communities that are vulnerable to misleading framings, but value truth and seek to protect the integrity of framing evolution within their communities.

6.2.1.2 Computational Tools to Facilitate Monitoring Framings

This section examines the implications for social media designers, providing them with guidance for designing computational tools that facilitate monitoring framings, a community moderation approach that extends beyond individualistic content moderation, as outlined in the preceding section. To do so, it first discusses opportunities to build upon the computational model designed and developed in this dissertation. Next, informed by the results

of the study conducted in Chapter 5, this section suggests social media designers to design tools to explore framing processes that might be critical, yet missing in a community. Finally, this section outlines important design considerations that this dissertation identifies as being overlooked in existing models that are developed for framing analysis, and the way the models presented in this dissertation accounts for these consideration. It discusses why and how these considerations should be similarly considered in the development of future such models for framing analysis, if platform designers wish to effectively monitor framing processes within online communities and make their platforms resilient to the broader impacts of misinformation.

Despite the efficacy of the model designed in this dissertation in facilitating exploratory analysis of framing, this model represents an initial effort in examining framing with with a dynamic, processual orientation using computational techniques. Therefore, there are great opportunities for social media designers to both improve and build upon this model. For example, social media designers can explores ways to enhance the model presented in this dissertation and make it more accessible, and more usable for community moderators seeking to monitor framings within their communities. To do so, social media designers are encouraged to conduct further human-subject studies, specifically with community moderators who are the "relevant human readers" of these models (Hoyle et al., 2021), and make this tool more inline with their expectations. In addition to improving the usability of the current model, social media designers should explore new tools that might incorporate additional linguistic features, such as catchphrases and metaphors, and/or consider implementing variable weighting for document sections (e.g., titles, introductions, bodies, conclusions) to reflect their potential framing influence.

Beyond designing tools that assist with analyzing framing, platform designers should explore tools that can monitor for *absent framings* within online communities, i.e., framings that are important but not not being evolved in a community. For example, analyzing framing processes in r/conspiracy community, this dissertation shows how the tragedy framing of the pandemic did not evolved in this community, and rather is rejected to reinvent the intentionality framing. Can social media designers examined tools that enable to see framing processes that are important, but missing in a community? Potential such tools can

inform community moderators about important aspects of an on going event that may be overlooked by their communities, empowering community moderators to offer framings that foster more comprehensive understandings about ongoing events within their communities. Arguably, such tools might be able to address some of the other community level mechanisms involved in the broader impacts of misinformation as well, such as familiarity bias, or formation of filter bubbles (See (Zuiderveen Borgesius et al., 2016) for more detail on this concept). More specifically, for example, by offering framings that might be completely ignored by community members, regardless of the intention beyond such ignorance, community moderators can contribute to what their communities observe and might mitigate potential familiarly bias and its consequences.

In addition to providing a concrete model with opportunities to build upon, and offering insights to design novel future such tools, this dissertation identifies and resolves two primarily concerns with existing techniques employed to analyze framing, which made those prior techniques less relevant, and insufficient in providing understanding about framing as processes involved in meaning constructions. The following passages describe these concerns, outline how they are addressed in the model presented in this dissertation, and suggest important design considerations for social media designers when designing any future such tools for monitoring framings.

First, prior computational techniques focuses on a single document, or sometimes even single sentences to identify frames. However, much insights from sociological research on framing suggests framing is a distributed process, and as such, framing evidences may be interwoven across multiple documents. Therefore, this dissertation argues that when investigating and designing computational techniques for assisting with examining framing, the unit of analysis should not be a single sentence or even a single document. Instead, to account for how framing evidence might be interwoven across multiple documents, and to address the aforementioned concern, the model designed in this dissertation do not enforce any individualistic unit of analysis. Put precisely, while results are organized by topics, the approach taken to leverage the results remains agnostic as to whether one topic or multiple topics constitute evidences of one framing. Similarly, social media designers aiming to design computational techniques to facilitate exploration of framings, as processes

of meaning constructions, should not enforce any individualistic unit of analysis. Instead, these models should be designed to examine framing evidences, as they appear within and across documents. The only prior computational techniques that move beyond isolated pieces of content in analyzing framing are topic modeling techniques (e.g., Card et al., 2016; DiMaggio et al., 2013; Ylä-Anttila et al., 2022). However, as they are currently designed and utilized in prior studies, these techniques fall short to offer insights into framing as a process, for the reason discussed below.

Second, prior computational techniques primarily focus on studying frames without exploring the interpretive packages that give meaning to an event (i.e., framing processes) (Gamson and Modigliani, 1989). As a result, such examinations offer limited insight into the processes by which people come to understand the world, which can be informative for community moderators aiming to address misleading framings within their communities. For example, by treating frames as discrete and distinct from the entities involved, previous computational work fall short in accounting for the various factors shaping people's understanding of an event, such as who is taking actions, what causes are at play, and who or what entities are impacted, and how people view and assess the moral aspects of the events. Indeed, these nuances are all parts of interpretive packages that (Gamson and Modigliani, 1989) argues are involved in processes by which people understand events, and are therefore important to attend to exploring framing processes. The computational model designed and developed in this dissertation accounts for not only what people say (i.e., the choice of words and words co-occurrence), but also how they talk about issues, and the entities that they see involved in the issues discussed. That is, by incorporating the grammatical relationships in which terms occur the presented model captures linguistic patterns that may be indicative of framing, by which it can attend to understanding about framing processes beyond the issues that are discussed. For instance, by providing evidence about how certain entities appear in documents, this model provides insights into the role of different entities at play, which is indeed an important aspect of framing processes. Capturing linguistic features that allow to attempt to relationship between topic terms beyond their co-occurrence, and might enable to interpret topics for understanding the nuances in language, should be similarly accounted for by social media designers aiming to examine

framings within their platforms.

In addition to the two concerns stated above, this dissertation posits that computational models which are designed to imposing pre-defined frames and seeking classification approaches to identifies those frames, as adopted in some prior work (e.g., Card et al., 2016; DiMaggio et al., 2013; Ylä-Anttila et al., 2022), is less effective, if at all, for understanding framing processes. Instead, it investigates designs computational models to find *evidence of framing languages*, by which researchers are able to explore framing processes. Put differently, this dissertation acknowledges the complexities inherent in framing processes, and only use the tool to assist researchers with identifying framing processes, without having the computational model to directly identify these framing processes. Social media designers should, similarly, design computational models with the goal of *assisting* community moderators in their analysis framing, rather than designing classification models or LLM-based models (which are fundamentally predictive models) to independently analyze and make inferences on their own.

To reiterate, social media designers aiming to design computational tools for framing analysis as a way to make their platform resilient to the broader impacts of misinformation, need to account for the discussed concerns, summarized as follows. First, recognizing that framing evidence is often distributed across multiple discussions, these tools should avoid imposing rigid analytical units. Second, understanding framing as processes of meaning construction requires to account for linguistic features that enable researchers to discern subtle nuances in how people understand events and attribute roles to different entities involved. Third, recognizing the complexities and nuances of language, as discussed throughout this dissertation, computational models are unlikely to directly interpret framing evidence or capture all involved nuances. Instead, these tools should be designed to support community moderators with their exploratory analysis of framing e.g., identifying linguistic patterns that may indicate framing), rather than making inferences about the framing processes themselves.

6.2.2 Community-Driven Interventions: Empowering Community Members to Mitigate Broader Impacts of Misinformation

this dissertation shows the importance of community members on how online communities evolve, both in terms of their role on how their communities' norms are perceived (Chapter 3), and in the way their communities come to understand the world's events (i.e., framing) (Chapter 5). Given this pivotal role of community members, it is essential to empower them to effectively mitigate the broader impacts of misinformation within their communities, specifically in addressing its effects on norm perceptions and framing evolutions, and help guide the healthy evolution of their communities. This section explores how social media designers can develop tools and interventions to support these objectives.

6.2.2.1 Community Members and their Role in Norms Evolution and Behaviors within their Community

This dissertation shows how community members can mitigate the broader impacts of misinformation on the way their community's norms are perceived (Chapter 3). More specifically, it demonstrates while the prevalence of false and misleading content within a community can influence perceptions about what is normative and acceptable within a community, community members can mitigate these impacts by responding to such content.

Therefore, given their impacts, it is important to empower community members with tools and features that can guide their efforts in mitigating the impacts of misinformation on how their community norms is persevered and how norms in their communities evolve. Such interventions might, for instance, ask a user to help address misleading content, and misleading reasoning in their community. Doing so via selective notifications could provide just-in-time responses to false and misleading content, not from an automated fact checker or from platform admins, but from other human community members.

In addition to developing these community-centered interventions, future research should also explore ways to encourage community members to leverage these interventions and join the effort of mitigating the spread and broader impacts of misinformation on their communities. For example, these interventions can directly inform community members

about their pivotal roles as a way to increase likelihood of their engagements with this effort. Additionally, community members can be prompted to consider the broader impacts of misleading content on how their community and its norm evolution. For instance, targeted community members could be encouraged to promote norms of sharing only reliable content, and encourage others to account for how the content they share might influence others, and to emphasize the value of legitimate information sharing.

These efforts do not necessarily need to be limited to responding to individual pieces of content. Rather, they can be designed to prompt behaviors that inoculate a community against misinformation spread and impacts. For example, in a recent work (Aghajari et al., 2024), the author and her collaborators demonstrate how community members can communicate about being vaccinated with a low-effort design element (i.e., adopting a vaccinated profile picture frame) and influence the perception of norms surrounding vaccination, an issue frequently targeted by misleading narratives (Silverman, 2021; Rao, 2021). The findings show that this low-effort, community-driven intervention not only helps protect perceived norms around vaccination from the broader impacts of misinformation, but also increases the likelihood that other community members will engage in efforts of addressing misleading content around vaccination.

Indeed, recent work shows that community members already perform significant moderation work, both by using site features, and by socially engaging in conversations within their community (Seering et al., 2019; Cullen and Kairam, 2022). Therefore, tools designed to engage community members in addressing the broader impacts of misinformation are likely to be well-received by community members. While engaging community members in addressing misinformation is a promising direction, further research is needed to explore how to effectively design around the role of community members. Key questions include which members should be targeted to participate in this effort? How should potential conflicts arising from these user engagements be managed? Addressing these questions presents valuable opportunities for future research to make significant contributions to research on misinformation.

6.2.2.2 Community Members and their Role in Framing Evolution within their Community

This dissertation shows that community members play a significant role on processes of framing evolutions within their communities (Chapter 5). For instance, by providing evidence (including factual or perceived, and leading or misleading evidence) and contributing to interpretations of the world's events, community members can influence how their communities come to understand those events. While some community members may adopt and promote misleading narratives that disrupt and diverge from reality, others within the same community may hold perceptions that are based on actual evidence on the event under discussion, offering alternative framings of the same event. These members, thus, can contribute to framing evolutions within their communities, mitigate the formation and spread of misleading framings, and prevent potential deviated framings to become the dominant framings through which others, especially vulnerable others, within their communities develop their understandings of world's event.

Social media designers are encouraged to equip community members with interventions to actively contribute to framing processes, especially when monitoring framing processes (discussed in the preceding section) reveals the development of framings that are disruptive and diverge from reality, and evolved using misleading evidence and faulty reasoning. Such interventions can effectively guide community members' efforts in addressing misleading framings, thereby enhancing the impacts of their efforts. For instance, in the case of framing the COVID-19 pandemic, the evidence around this event reported in the news suggests the pandemic is indeed a global tragedy, with actual evidence from the death tolls and number of infected people. Yet, some community members might come to discuss this event as a hoax (possibly even in a community wherein false and misleading content is not prevalent), and/or diminish its magnitudes, considering the COVID-19 virus as a typical form of flue. However, in the same community, there could be people who acknowledge the actual evidence, and understand the pandemic as a tragedy (informed by news, personal experience, or the experiences of those around them), which requires every individuals to follow the guidance by the health authorities to help mitigate the spread and impacts of

the virus.

Could these members effectively mitigate the impacts of misperception that misleading framing of the pandemic might cause? Motivated by the insights from this dissertation on the role of community members on framing evolutions, this dissertation asserts these community members can potentially mitigate these broader impacts of misleading framings within their communities, if their efforts are guided and are made more visible. For example, selective notifications could request community members to provide just-in-time engagement with their community in response to misleading framings, and help their community make their arguments based on "factual evidence" and not "misleading evidence", and/or help clarify problems with faulty reasoning and offer alternative reasoning. Even if those involved in the formation of misleading narratives are not directly protected by such community efforts (e.g., due to strong beliefs in conspiracy theories), these community oriented initiatives can still potentially influence vulnerable individuals who are trying to make sense of the events around them.

Following the arguments presented in the preceding section, the design of low-effort interventions offers a potentially effective strategy for enhancing community engagement. For example, when reaching to targeted community members to contribute to framing about an event, platform moderators can provide evidence of misleading framings, inform these targeted members about the broader impacts of such misleading framings, and offer resources to these members to provide alternative, and leading framings within their community. Given that framing evidence can be intertwined across documents, it is perhaps effective to ask community members to both engage in individual discussions and contribute to the framings happening in those discussions, and to share leading framing in the form of post to reach broader community and mitigate the influence of misleading framings. Future research is required to examine which of these efforts could be more effective, and how to deploy the combinations of these interventions.

Indeed, empowering community members to mitigate the formation of misleading framings not only influences how communities understand and respond to world events, but also potentially shapes the perception and evolution of community norms. These perception about community's norms can in turn further contribute to the way communities run, both

in terms of the types of content that is shared, the framings that evolves around the content shared in a community, and the way community members respond to those framings (Aghajari et al., 2023c, 2024).

Bibliography

2006. *Hearing the other side: Deliberative versus participatory democracy.* Cambridge University Press.

2018. How Twitter is fighting spam and malicious automation. https://blog.twitter.com/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation

2018. How whatsapp helped turn an Indian village into a lynch mob. https://www.bbc.com/news/world-asia-india-44856910

2019a. Designing new ways to give context to news stories. https://newsroom.fb.com/news/2018/04/inside-feed-article-context/

2019b. Helping ensure news on Facebook is from trusted sources. https://about.fb.com/news/2018/01/trusted-sources/

2019. Information operations directed at Hong Kong. https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong

2020. Fake news and its impact on the economy. https://priorityconsultants.com/blog/fake-news-and-its-impact-on-the-economy/.

2020. Partly false claim: Vaccines contain toxic levels of aluminum, polysorbate 80, yeast and other substances. https://www.reuters.com/article/uk-factcheck-vaccines-toxic-substances-idUSKBN22H2OP

2021. Removing coordinated inauthentic behavior. https://about.fb.com/news/2020/09/removing-coordinated-inauthentic-behavior-russia/

2022. How Americans came to Distrust Science. https://bostonreview.net/articles/andrew-jewett-science-under-fire/

2022. Latent variable analysis [R package lavaan version 0.6-12]. https://cran.r-project.org/web/packages/lavaan/

2022. Research guides. https://guides.temple.edu/fakenews. (Accessed on 10/05/22).

2022. Top 10 U.S. newspapers by circulation. https://www.agilitypr.com/resources/top-media-outlets/top-10-daily-american-newspapers/

Devon A Abdallah and Christine M Lee. 2021. Social norms and vaccine uptake: College students' COVID vaccination intentions, attitudes, and estimated peer norms and comparisons with influenza vaccine. *Vaccine* 39, 15 (2021), 2060–2067.

Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*. 36–43.

Zhila Aghajari. 2023. Adopting an ecological approach to misinformation: Understanding the broader impacts on online communities. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 417–420.

Zhila Aghajari, Eric PS Baumer, and Dominic DiFranzo. 2023a. What's the Norm Around Here? Individuals' Responses Can Mitigate the Effects of Misinformation Prevalence in Shaping Perceptions of a Community. (2023).

Zhila Aghajari, Eric PS Baumer, Allison Lazard, Nabarun Dasgupta, and Dominic DiFranzo. 2024. Investigating the Mechanisms by which Prevalent Online Community Behaviors Influence Responses to Misinformation: Do Perceived Norms Really Act as a Mediator?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–14.

Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. 2023b. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. (2023).

Zhila Aghajari, Eric P. S. Baumer, and Dominic DiFranzo. 2023c. What's the Norm Around Here? Individuals' Responses Can Mitigate the Effects of Misinformation Prevalence in Shaping Perceptions of a Community. (2023).

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

Kimberley R Allison, Kay Bussey, and Naomi Sweller. 2019. 'I'm going to hell for laughing at this' Norms, Humour, and the Neutralisation of Aggression in Online Communities. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of LDA generative models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I 20*. Springer, 67–82.

Ashley A Anderson, Sara K Yeo, Dominique Brossard, Dietram A Scheufele, and Michael A Xenos. 2018. Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research* 30, 1 (2018), 156–168.

Craig A Anderson, Mark R Lepper, and Lee Ross. 1980. Perseverance of social theories: the role of explanation in the persistence of discredited information. *Journal of personality and social psychology* 39, 6 (1980), 1037.

Janna Anderson and Lee Rainie. 2020. The Future of Truth and Misinformation Online. `https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/`

Simge Andı and Jesper Akesson. 2020. Nudging Away False News: Evidence from a Social Norms Experiment. *Digital Journalism* 9, 1 (2020), 106–125.

Nicolas M Anspach. 2017. The new personal influence: How our Facebook friends influence the news we read. *Political Communication* 34, 4 (2017), 590–606.

Ahmer Arif, John J Robinson, Stephanie A Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A closer look at the self-correcting crowd: Examining

corrections in online rumors. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 155–168.

R Armitage. 2021. Online 'anti-vax'campaigns and COVID-19: censorship is not the solution. *Public Health* 190 (2021), e29.

Robin Aronow. 2016. Semantics: Thematic Roles. `https://www.linguisticsnetwork.com/semantics-thematic-roles/`

AssociatedPress. 2021. Twitter launches crowd-sourced fact-checking project. `https://apnews.com/article/twitter-launch-crowd-sourced-fact-check-589809d4c9a7eceda1ea8293b0a14af2`

American Library Association et al. 2009. Association of College and Research Libraries. *Information literacy competency standards for higher education* (2009), 2–3.

Charles K Atkin. 1985. Informational utility and selective exposure to entertainment media. *Selective exposure to communication* (1985), 63–91.

Patricia Aufderheide. 2018. Media literacy: From a report of the national leadership conference on media literacy. In *Media literacy in the information age*. Routledge, 79–86.

Stef Aupers. 2012. 'Trust no one': Modernization, paranoia and conspiracy culture. *European Journal of Communication* 27, 1 (2012), 22–34.

Frederick T Bacon. 1979. Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory* 5, 3 (1979), 241.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

David Bamman, Ted Underwood, and Noah A Smith. 2014. A bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 370–379.

John A Bargh, Katelyn YA McKenna, et al. 2004. The Internet and social life. *Annual review of psychology* 55, 1 (2004), 573–590.

Corey H Basch, Aleksandar Kecojevic, and Victoria H Wagner. 2020. Coverage of the COVID-19 pandemic in the online versions of highly circulated US daily newspapers. *Journal of community health* 45 (2020), 1089–1097.

Melisa Basol, Jon Roozenbeek, Manon Berriche, Fatih Uenal, William P McClanahan, and Sander van der Linden. 2021. Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society* 8, 1 (2021), 20539517211013868.

Melisa Basol, Jon Roozenbeek, and Sander van der Linden. 2020. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition* 3, 1 (2020).

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*. 1472–1482.

Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 conference on Designing interactive systems*. 93–102.

Eric PS Baumer, Xiaotong Xu, Christine Chu, Shion Guha, and Geri K Gay. 2017. When subjects interpret the data: Social media non-use as a case for adapting the Delphi Method to CSCW. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 1527–1543.

Ian Maynard Begg, Ann Anas, and Suzanne Farinacci. 1992. Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General* 121, 4 (1992), 446.

Robert D Benford and David A Snow. 2000. Framing processes and social movements: An overview and assessment. *Annual review of sociology* (2000), 611–639.

Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021a. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

Md Momen Bhuiyan, Hayden Whitley, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021b. Designing Transparency Cues in Online News Platforms to Promote Trust: Journalists'& Consumers' Perspectives. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–31.

Md Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael A Horning, and Tanushree Mitra. 2018. FeedReflect: A tool for nudging users to assess news credibility on twitter. In *Companion of the 2018 ACM conference on computer supported cooperative work and social computing*. 205–208.

Cristina Bicchieri. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Cristina Bicchieri, Enrique Fatas, Abraham Aldama, Andrés Casas, Ishwari Deshpande, Mariagiulia Lauro, Cristina Parilli, Max Spohn, Paula Pereira, and Ruiling Wen. 2021. In science we (should) trust: Expectations and compliance across nine countries during the COVID-19 pandemic. *PloS one* 16, 6 (2021), e0252892.

Nick Bilton. 2019. The downfall of Alex Jones shows how the internet can be saved. `https://www.vanityfair.com/news/2019/04/the-downfall-of-alex-jones-shows-how-the-internet-can-be-saved`

George D Bishop and David G Myers. 1974. Informational influence in group discussion. *Organizational behavior and human performance* 12, 1 (1974), 92–104.

Ladislav Bittman. 1985. *The KGB and Soviet disinformation: an insider's view*. Washington: Pergamon-Brassey's.

David Blei and John Lafferty. 2006. Correlated topic models. *Advances in neural information processing systems* 18 (2006), 147.

David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

Leticia Bode and Emily K Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication* 65, 4 (2015), 619–638.

Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI interprets the probes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1077–1086.

George J. Borjas. 2016. Yes, immigration hurts American workers. `https://www.politico.com/magazine/story/2016/09/trump-clinton-immigration-economy-unemployment-jobs-214216/`

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 1–14.

Samantha Bradshaw, Lisa-Maria Neudert, and Philip N Howard. 2018. Government responses to malicious use of social media. *NATO StratCom Centre of Excellence, Riga, Working Paper* (2018).

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

J Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* (1996).

Hronn Brynjarsdottir, Maria Håkansson, James Pierce, Eric Baumer, Carl DiSalvo, and Phoebe Sengers. 2012. Sustainably unpersuaded: how persuasion narrows our vision of sustainability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 947–956.

Tom Buchanan. 2020. Why do people spread false information online? The effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *Plos one* 15, 10 (2020), e0239666.

Tom Buchanan and Vladlena Benson. 2019. Spreading Disinformation on Facebook: Do Trust in Message Source, Risk Propensity, or Personality Affect the Organic Reach of "Fake News"? *Social Media+ Society* 5, 4 (2019), 2056305119888654.

Cody Buntain, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2021. YouTube recommendations and effects on sharing across online social platforms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

Eugene Burnstein and Amiram Vinokur. 1973. Testing two classes of theories about group induced shifts in individual choice. *Journal of Experimental Social Psychology* 9, 2 (1973), 123–137.

Leonardo Bursztyn, Georgy Egorov, and Stefano Fiorin. 2020. From extreme to mainstream: The erosion of social norms. *American Economic Review* 110, 11 (2020), 3522–48.

David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & quantity* 56, 3 (2022), 1391–1412.

Sahara Byrne and Philip Solomon Hart. 2009. The Boomerang Effect A Synthesis of Findings and a Preliminary Theoretical Framework. *Annals of the International Communication Association* 33, 1 (Jan. 2009), 3–37. https://doi.org/10.1080/23808985.2009.11679083

Joseph N Cappella and Kathleen Hall Jamieson. 1994. Broadcast adwatch effects: A field experiment. *Communication Research* 21, 3 (1994), 342–365.

Dallas Card, Justin H Gross, Amber Boydstun, and Noah A Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 1410–1420.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. 675–684.

Gizem Ceylan and Norbert Schwarz. 2020. Look What I Am Re-Sharing: How Self-Presentation Goals Impact What Consumers Spread on Social Networks. *ACR North American Advances* (2020).

Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science* 28, 11 (2017), 1531–1546.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems* 22 (2009).

Emily Chen, Herbert Chang, Ashwin Rao, Kristina Lerman, Geoffrey Cowan, and Emilio Ferrara. 2021. COVID-19 misinformation and the 2020 US presidential election. *The Harvard Kennedy School Misinformation Review* (2021).

Xinran Chen. 2016. The influences of personality and motivation on the sharing of misinformation on social media. *IConference 2016 Proceedings* (2016).

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1217–1230.

Noam Chomsky. 1997. What makes mainstream media mainstream. *Z magazine* 10, 10 (1997), 17–23.

Dennis Chong. 1994. Tolerance and social adjustment to new norms and practices. *Political Behavior* 16, 1 (1994), 21–53.

Adrienne Chung and Rajiv N Rimal. 2016. Social norms: A review. *Review of Communication Research* 4 (2016), 1–28.

Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annual review of psychology* 55, 1 (2004), 591–621.

Robert B Cialdini, Carl A Kallgren, and Raymond R Reno. 1991. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In *Advances in experimental social psychology*. Vol. 24. Elsevier, 201–234.

Robert B Cialdini, Raymond R Reno, and Carl A Kallgren. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology* 58, 6 (1990), 1015.

Robert B. Cialdini and Melanie R. Trost. 1998. Social influence: Social norms, conformity and compliance. In *The handbook of social psychology* (4 ed.), Daniel T. Gilbert, Susan T. Fiske, and Gardner Lindzey (Eds.). McGraw-Hill, 151–192.

Jacopo Ciani Sciolla. 2023. The normative challenges of data scraping: legal hurdles and steps forward. *i-lex* 16, 2 (2023), i–viii.

Justin T Clapp, Jody A Roberts, Britt Dahlberg, Lee Sullivan Berry, Lisa M Jacobs, Edward A Emmett, and Frances K Barg. 2016. Realities of environmental toxicity and their ramifications for community engagement. *Social Science & Medicine* 170 (2016), 143–151.

Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior* 42, 4 (2020), 1073–1095.

Kevin Coe, Kate Kenski, and Stephen A Rains. 2014. Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication* 64, 4 (2014), 658–679.

Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational Journalism: A Call to Arms to Database Researchers. In *5th Biennial Conference on Innovative Data Systems Research (CIDR'll), ACM.*

Jonas Colliander. 2019. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior* 97 (2019), 202–215.

Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.

Maria Constantinou, Andreas Kagialis, and Maria Karekla. 2021. A Systematic Review and Meta-Analysis of the Efficacy of Acceptance and Commitment Therapy for Social Anxiety Disorder. *Behavioral Sciences* 11, 7 (2021), 98. https://doi.org/10.3390/bs11070098

Josh Constine. 2017. Facebook fights fake news with links to other angles. https://techcrunch.com/2017/08/03/facebook-related-articles/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAGkjbvnLbViPDaKtfsQZuRdpcmsVLSFLP9egNExbvgObqXa8_QOM_jXOWpCX6vtZXEnVS5tzwdqWZSpnzL6_M6w67dQj8U5bheaA93Ch_pqUia_EAcEIqfJcrGNJgS48eIAGdIoam0-nPoWIRfVQxIln5vHobK3wGAXy0xQ3AByx

Darel Cookson, Daniel Jolley, Robert C Dempsey, and Rachel Povey. 2021. "If they believe, then so shall I": Perceived beliefs of the in-group predict conspiracy theory belief. *Group Processes & Intergroup Relations* 24, 5 (2021), 759–782.

Paul T Costa Jr and Robert R McCrae. 2008. *The Revised Neo Personality Inventory (neo-pi-r)*. Sage Publications, Inc.

Joseph Cox and Jason Koebler. 2019. Facebook Bans White Nationalism and White Separatism. https://www.vice.com/en_us/article/nexpbx/facebook-bans-white-nationalism-and-white-separatism

Amanda LL Cullen and Sanjay R Kairam. 2022. Practicing moderation: Community moderation as reflective practice. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW1 (2022), 1–32.

Paul G Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66 (2016), 4–19.

Yvonne AW De Kort, L Teddy McCalley, and Cees JH Midden. 2008. Persuasive trash cans: Activation of littering norms by design. *Environment and Behavior* 40, 6 (2008), 870–891.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses.. In *Lrec*, Vol. 6. 449–454.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559.

David Detmer. 2003. Challenging postmodernism: Philosophy and the politics of truth. (2003).

Morton Deutsch and Harold B Gerard. 1955. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology* 51, 3 (1955), 629.

Nicholas Dias, Gordon Pennycook, and David G Rand. 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* 1, 1 (2020).

Nicholas Dias, Gordon Pennycook, and David G. Rand. 2021. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media: HKS Misinformation Review. `https://misinforeview.hks.harvard.edu/article/emphasizing-publishers-does-not-reduce-misinformation/`

Nicholas DiFonzo, Martin J Bourgeois, Jerry Suls, Christopher Homan, Noah Stupak, Bernard P Brooks, David S Ross, and Prashant Bordia. 2013. Rumor clustering, consensus, and polarization: Dynamic social impact and self-organization of hearsay. *Journal of Experimental Social Psychology* 49, 3 (2013), 378–399.

Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. 2018. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics* 41, 6 (2013), 570–606.

Karthik Dinakar, Jackie Chen, Henry Lieberman, Rosalind Picard, and Robert Filbin. 2015. Mixed-initiative real-time topic modeling & visualization for crisis counseling. In *Proceedings of the 20th international conference on intelligent user interfaces*. 417–426.

Carl DiSalvo, Phoebe Sengers, and Hrönn Brynjarsdóttir. 2010. Mapping the landscape of sustainable HCI. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1975–1984.

Mine Dogucu and Mine Çetinkaya-Rundel. 2021. Web scraping in the statistics and data science curriculum: Challenges and opportunities. *Journal of Statistics and Data Science Education* 29, sup1 (2021), S112–S122.

Nicole Doughty, Helen M Paterson, Carolyn MacCann, and Lauren A Monds. 2017. Personality and memory conformity. *Journal of Individual Differences* (2017).

Karen M Douglas and Robbie M Sutton. 2015. Climate change: Why the conspiracy theories are dangerous. *Bulletin of the Atomic Scientists* 71, 2 (2015), 98–106.

Karen M. Douglas, Robbie M. Sutton, and Aleksandra Cichocka. 2017. The Psychology of Conspiracy Theories. *Current Directions in Psychological Science* 26, 6 (Dec. 2017), 538–542. https://doi.org/10.1177/0963721417718261

James N Druckman. 2001. The implications of framing effects for citizen competence. *Political behavior* 23, 3 (2001), 225–256.

Brianna Dym and Casey Fiesler. 2018. Vulnerable and online: Fandom's case for stronger privacy norms and tools. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 329–332.

Alice H Eagly and Shelly Chaiken. 1993. *The psychology of attitudes.* Harcourt brace Jovanovich college publishers.

Adam M Enders, Joseph E Uscinski, Casey Klofstad, and Justin Stoler. 2020. The different forms of COVID-19 misinformation and their consequences. *The Harvard Kennedy School Misinformation Review* (2020).

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory* 390 (1993), 397.

Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–11.

Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

Facebook. 2013. About Fact-Checking on Facebook. `https://www.facebook.com/business/help/2593586717571940?id=673052479947730`. Online; accessed 13 January 2022.

Facebook. 2018. Removing Coordinated Inauthentic Behavior From China. `https://about.fb.com/news/2019/08/removing-cib-china/`. Online; accessed 13 January 2022.

Facebook. 2020. Expanding Facebook's U.S. Fact-Checking Program and Supporting the Fact-Checking Ecosystem. `https://www.facebook.com/journalismproject/fact-checking-expansion-and-investment-2020`

Facebook. 2021. How fact-checking works. https://transparency.fb.com/features/how-fact-checking-works/. Online; accessed 13 January 2022.

FacebookNewsroom. 2017. Replacing Disputed Flags With Related Articles. https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/

Andrej Findor, Matej Hruška, John A Gould, Roman Hlatky, Zuzana Tomková, and Miroslav Sirota. 2021. Framing Effects, Social Norm Perception, and Tolerance of Lesbian and Gay Individuals: Experimental Evidence from Slovakia. *Journal of Homosexuality* (2021), 1–25.

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

Brian J Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, et al. 2001. What makes web sites credible? A report on a large quantitative study. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 61–68.

Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. 2003. How do users evaluate the credibility of Web sites? A study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*. 1–15.

Thomas E Ford, Erin R Wentzel, and Joli Lorion. 2001. Effects of exposure to sexist humor on perceptions of normative tolerance of sexism. *European Journal of Social Psychology* 31, 6 (2001), 677–691.

Krisandra S Freeman and Jan H Spyridakis. 2004. An examination of factors that affect the credibility of online health information. *Technical communication* 51, 2 (2004), 239–263.

Adrien Friggeri, Lada Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor cascades. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

Romy Froehlich and Burkhard Rüdiger. 2006. Framing political public relations: Measuring success of political communication strategies in Germany. *Public Relations Review* 32, 1 (2006), 18–25.

Lorenzo Gagliardi. 2023. The role of cognitive biases in conspiracy beliefs: A literature review. *Journal of Economic Surveys* (2023).

William A Gamson. 1989. News as framing: Comments on Graber. *American behavioral scientist* 33, 2 (1989), 157–161.

William A Gamson and Andre Modigliani. 1989. Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology* 95, 1 (1989), 1–37.

Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–16.

Megan Garber. 2012. The newest factchecker: Reddit. `https://www.theatlantic.com/technology/archive/2012/10/the-newest-factchecker-reddit/263238/`

R Kelly Garrett. 2009a. Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of computer-mediated communication* 14, 2 (2009), 265–285.

R. Kelly Garrett. 2009b. Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate. *Journal of Communication* 59, 4 (2009), 676–699. `https://doi.org/10.1111/j.1460-2466.2009.01452.x`

R Kelly Garrett and Shannon Poulsen. 2019. Flagging Facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings. *Journal of Computer-Mediated Communication* 24, 5 (2019), 240–258.

R Kelly Garrett and Brian E Weeks. 2013. The promise and peril of real-time corrections to political misperceptions. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1047–1058.

Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), 721–741.

Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20, 12 (2018), 4492–4511.

Sayantari Ghosh, Saumik Bhattacharya, Shagata Mukherjee, and Sujoy Chakravarty. 2024. Promote to protect: data-driven computational model of peer influence for vaccine perception. *Scientific Reports* 14, 1 (2024), 306.

Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moder ation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

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

Ted Goertzel. 1994. Belief in conspiracy theories. *Political psychology* (1994), 731–742.

Erving Goffman. 1974. *Frame analysis: An essay on the organization of experience.* Harvard University Press.

Jack Goodman and Flora Carmichael. 2020. Coronavirus: Bill Gates 'microchip'conspiracy theory and other vaccine claims fact-checked. *BBC News* 30 (2020).

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101, suppl_1 (2004), 5228–5235.

James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.

Andrew M Guess and Benjamin A Lyons. 2020. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform* 10 (2020).

Kirk Hallahan. 1999. Seven models of framing: Implications for public relations. *Journal of public relations research* 11, 3 (1999), 205–242.

Felix Hamborg, Norman Meuschke, and Bela Gipp. 2017. Matrix-based news aggregation: exploring different news perspectives. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 1–10.

Michael Hameleers and Toni GLA van der Meer. 2020. Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research* 47, 2 (2020), 227–250.

Lawrence C Hamilton, Joel Hartter, and Kei Saito. 2015. Trust in scientists on climate change and vaccines. *Sage Open* 5, 3 (2015), 2158244015602752.

Noriko Hara, Jessica Abbazio, and Kathryn Perkins. 2019. An emerging form of public engagement with science: Ask Me Anything (AMA) sessions on Reddit r/science. *PloS one* 14, 5 (2019), e0216789.

Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

Lynn Hasher, David Goldstein, and Thomas Toppino. 1977. Frequency and the conference of referential validity. *Journal of verbal learning and verbal behavior* 16, 1 (1977), 107–112.

Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015. The quest to automate fact-checking. In *Proceedings of the 2015 computation+ journalism symposium*.

Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. Share, like, recommend: Decoding the social media news consumer. *Journalism studies* 13, 5-6 (2012), 815–824.

Benjamin D Horne, Dorit Nevo, Sibel Adali, Lydia Manikonda, and Clare Arrington. 2020. Tailoring heuristics and timing AI interventions for supporting news veracity assessments. *Computers in Human Behavior Reports* 2 (2020), 100043.

Amin Hosseiny Marani, Joshua Levine, and Eric PS Baumer. 2022. One Rating to Rule Them All? Evidence of Multidimensionality in Human Assessment of Topic Labeling Quality. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 768–779.

Carl I Hovland and Walter Weiss. 1951. The influence of source credibility on communication effectiveness. *Public opinion quarterly* 15, 4 (1951), 635–650.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Joop J Hox and Timo M Bechger. 1998. An introduction to structural equation modeling. (1998).

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in neural information processing systems* 34 (2021), 2018–2033.

Y Linlin Huang, Kate Starbird, Mania Orand, Stephanie A Stanek, and Heather T Pedersen. 2015. Connected through crisis: Emotional proximity and the spread of misinformation

online. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 969–980.

Robert Huckfeldt, Paul Allen Beck, Russell J Dalton, and Jeffrey Levine. 1995. Political environments, cohesive social groups, and the communication of public opinion. *American Journal of Political Science* (1995), 1025–1054.

Robert Huckfeldt, Jeanette Morehouse Mendez, and Tracy Osborn. 2004. Disagreement, ambivalence, and engagement: The political consequences of heterogeneous networks. *Political Psychology* 25, 1 (2004), 65–95.

Elle Hunt. 2017. Disputed by multiple fact-checkers': Facebook rolls out new alert to combat fake news. *The Guardian* 21 (2017).

Juan-José Igartua and Lifen Cheng. 2009. Moderating effect of group cue while processing news on immigration: Is the framing effect a heuristic process? *Journal of Communication* 59, 4 (2009), 726–749.

Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: Computationally-derived social signals from account histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

R. Imhoff and M. Bruder. 2014. Speaking (un-) truth to power: Conspiracy mentality as a generalised political attitude. *European Journal of Personality* 28, 1 (2014), 25–43. `https://doi.org/10.1002/per.1930`

H Innes and M Innes. 2021. De-platforming disinformation: conspiracy theories and their control. *Information, Communication & Society* (2021), 1–19.

Mike Isaac. 2016. How facebook's fact-checking partnership will work. `https://www.nytimes.com/2016/12/15/technology/facebook-fact-checking-fake-news.html`

Shanto Iyengar. 1996. Framing responsibility for political issues. *The Annals of the American Academy of Political and Social Science* 546, 1 (1996), 59–70.

Justin Jager, Diane L Putnick, and Marc H Bornstein. 2017. II. More than just convenient: The scientific merits of homogeneous convenience samples. *Monographs of the society for research in child development* 82, 2 (2017), 13–30.

Farnaz Jahanbakhsh, Amy X Zhang, Adam J Berinsky, Gordon Pennycook, David G Rand, and David R Karger. 2021. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–42.

Youngseung Jeon, Bogoan Kim, Aiping Xiong, Dongwon Lee, and Kyungsik Han. 2021. ChamberBreaker: Mitigating the Echo Chamber Effect and Supporting Information Hygiene through a Gamified Inoculation System. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–26.

Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

Neil F Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. 2020. The online competition between pro-and anti-vaccination views. *Nature* 582, 7811 (2020), 230–233.

Ridley Jones, Lucas Colusso, Katharina Reinecke, and Gary Hsieh. 2019. r/science: Challenges and opportunities in online science communication. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

S Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. 2021. Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist* 65, 2 (2021), 371–388.

The Wall Street Journal. [n. d.]. Breaking news, business, Financial Economic News, World News and Video. https://www.wsj.com/

Dan M Kahan. 2012. Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision making* 8 (2012), 407–24.

Dan M Kahan. 2017. Misconceptions, misinformation, and the logic of identity-protective cognition. (2017).

Joseph Kahne and Benjamin Bowyer. 2017. Educating for democracy in a partisan age: Confronting the challenges of motivated reasoning and misinformation. *American Educational Research Journal* 54, 1 (2017), 3–34.

Daniel Kahneman and Amos Tversky. 1984. Choices, values, and frames. *American psychologist* 39, 4 (1984), 341.

Jonas Kaiser, Birte Fähnrich, and Laura Heintz. 2023. Ups and downs on "r/science"—exploring the dynamics of science communication on Reddit. *Journal of Science Communication* 22 (2023).

Aimée A Kane, Sara Kiesler, and Ruogu Kang. 2018. Inaccuracy Blindness in Collaboration Persists, even with an Evaluation Prompt. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–9.

Alireza Karduni, Isaac Cho, Ryan Wesslen, Sashank Santhanam, Svitlana Volkova, Dustin L Arendt, Samira Shaikh, and Wenwen Dou. 2019. Vulnerable to misinformation? Verifi!. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 312–323.

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 137–146.

Hye Kyung Kim, Jisoo Ahn, Lucy Atkinson, and Lee Ann Kahlor. 2020. Effects of COVID-19 misinformation on information seeking, avoidance, and processing: A multicountry comparative study. *Science Communication* 42, 5 (2020), 586–615.

Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 324–332.

Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.

Jan Kirchner and Christian Reuter. 2020. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW2 (2020), 1–27.

Baris Kirdemir, Joseph Kready, Esther Mead, Muhammad Nihal Hussain, Nitin Agarwal, and Donald Adjeroh. 2021. Assessing bias in YouTube's video recommendation algorithm in a cross-lingual and cross-topical context. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 71–80.

Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. (2007).

Esther Michelsen Kjeldahl and Vincent F Hendricks. 2018. The sense of social influence: pluralistic ignorance in climate change: social factors play key roles in human behavior. Individuals tend to underestimate how much others worry about climate change. This may inhibit them from taking collective climate action. *EMBO reports* 19, 11 (2018), e47185.

Colin Klein, Peter Clutton, and Adam G Dunn. 2019. Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum. *PloS one* 14, 11 (2019), e0225098.

Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data–frame theory of sensemaking. In *Expertise out of context*. Psychology Press, 118–160.

CJason Koebler. 2018. Deplatforming Works. `vice.com/en/article/bjbp9d/do-social-media-bans-work`

Alex Zhi-Xiong Koo, Min-Hsin Su, SangWon Lee, So-Yun Ahn, and Hernando Rojas. 2021. What Motivates People to Correct Misinformation? Examining the Effects of Third-person Perceptions and Perceived Norms. *Journal of Broadcasting & Electronic Media* (2021), 1–24.

Yong Ming Kow, Yubo Kou, Xitong Zhu, and Wang Hin Sy. 2019. "Just My Intuition": Awareness of Versus Acting on Political News Misinformation. In *International Conference on Information*. Springer, 469–480.

Hamutal Kreiner and Eyal Gamliel. 2021. Framing fake news: Asymmetric attribute-framing bias for favorable and unfavorable outcomes. *Journal of Experimental Psychology: Learning, Memory, and Cognition* (2021).

Sarah E Kreps, Jillian L Goldfarb, John S Brownstein, and Douglas L Kriner. 2021. The relationship between US adults' misconceptions about COVID-19 vaccines and vaccination preferences. *Vaccines* 9, 8 (2021), 901.

Ziva Kunda. 1990. The case for motivated reasoning. *Psychological bulletin* 108, 3 (1990), 480.

Jim A Kuypers. 2010. Framing analysis from a rhetorical perspective. *Doing news framing analysis: Empirical and theoretical perspectives* (2010), 286–311.

Maria Knight Lapinski and Rajiv N Rimal. 2005. An explication of social norms. *Communication theory* 15, 2 (2005), 127–147.

Richard R Lau and Mark Schlesinger. 2005. Policy frames, metaphorical reasoning, and support for public policies. *Political Psychology* 26, 1 (2005), 77–114.

David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. 2017. Combating fake news: An agenda for research and action. (2017).

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies* 105 (2017), 28–42.

Sune Lehmann and Yong-Yeol Ahn. 2018. *Complex spreading phenomena in social systems.* Springer.

Kristina Lerman, Xiaoran Yan, and Xin-Zeng Wu. 2016. The" majority illusion" in social networks. *PloS one* 11, 2 (2016), e0147617.

Leo Levy. 1960. Studies in conformity behavior: a methodological note. *The Journal of Psychology* 50, 1 (1960), 39–41.

Stephan Lewandowsky, Ullrich KH Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition* 6, 4 (2017), 353–369.

Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

Stephan Lewandowsky and Klaus Oberauer. 2016. Motivated rejection of science. *Current Directions in Psychological Science* 25, 4 (2016), 217–222.

Björn Lindström, Simon Jangard, Ida Selbing, and Andreas Olsson. 2018. The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General* 147, 2 (2018), 228.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*. 504–514.

Claudia Claudia Wai Yu Lo. 2018. *When all you have is a banhammer: the social and communicative work of Volunteer moderators*. Ph. D. Dissertation. Massachusetts Institute of Technology.

Milton Lodge and Charles S Taber. 2013. *The rationalizing voter*. Cambridge University Press.

John Lofland, David Snow, Leon Anderson, and Lyn H Lofland. 2022. *Analyzing social settings: A guide to qualitative observation and analysis*. Waveland Press.

Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* 5, 3 (2021), 337–348.

Christopher Lucas, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. 2015. Computer-assisted text analysis for comparative politics. *Political Analysis* 23, 2 (2015), 254–277.

Alex Luscombe, Kevin Dick, and Kevin Walby. 2022. Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity* 56, 3 (2022), 1023–1044.

Robert Luzsa and Susanne Mayr. 2021. False consensus in the echo chamber: Exposure to favorably biased social media news feeds leads to increased perception of public support for own opinions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 15, 1 (2021).

Diane M Mackie. 1986. Social identification effects in group polarization. *Journal of Personality and Social Psychology* 50, 4 (1986), 720.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

Amin Hosseiny Marani and Eric PS Baumer. 2023. A Review of Stability in Topic Modeling: Metrics for Assessing and Techniques for Improving Stability. *Comput. Surveys* (2023).

Gary Marks and Norman Miller. 1987. Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological bulletin* 102, 1 (1987), 72.

Cameron Martel, Gordon Pennycook, and David G Rand. 2020. Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications* 5, 1 (2020), 1–20.

Maryville University. [n. d.]. What is Mainstream Media? `https://online.maryville.edu/blog/what-is-mainstream-media/#what-is`

Gina M Masullo and Jiwon Kim. 2021. Exploring "angry" and "like" reactions on uncivil Facebook comments that correct misinformation in the news. *Digital Journalism* 9, 8 (2021), 1103–1122.

Philipp K Masur, Dominic James DiFranzo, and Natalya Natalie Bazarova. 2021. Behavioral Contagion on Social Media: Effects of Social Norms, Design Interventions, and Critical Media Literacy on Self-Disclosure. (2021).

J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.

Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences* 114, 48 (2017), 12714–12719.

Scott E Maxwell, Harold D Delaney, and Ken Kelley. 2017. *Designing experiments and analyzing data: A model comparison perspective*. Routledge.

John McAlaney, Bridgette Bewick, and Clarissa Hughes. 2011. The international development of the 'Social Norms' approach to drug education and prevention. *Drugs: education, prevention and policy* 18, 2 (2011), 81–89.

D Harrison McKnight and Charles J Kacmar. 2007. Factors and effects of information credibility. In *Proceedings of the ninth international conference on Electronic commerce*. 423–432.

Meta. 2020. Helping Fact-checkers Identify False Claims Faster. `https://about.fb.com/news/2019/12/helping-fact-checkers/`. [Online; accessed 12-April-2023].

Panagiotis Takas Metaxas, Samantha Finn, and Eni Mustafaraj. 2015. Using twittertrails.com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*. 69–72.

Miriam J Metzger and Andrew J Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics* 59 (2013), 210–220.

Nicholas Micallef, Mihai Avram, Filippo Menczer, and Sameer Patil. 2021. Fakey: A Game Intervention to Improve News Literacy on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

Dale T Miller and Cathy McFarland. 1991. When social comparison goes awry: The case of pluralistic ignorance. (1991).

Randall K Minas, Robert F Potter, Alan R Dennis, Valerie Bartelt, and Soyoung Bae. 2014. Putting on the thinking cap: using NeuroIS to understand information processing biases in virtual teams. *Journal of Management Information Systems* 30, 4 (2014), 49–82.

Martin Moore and Damian Tambini. 2018. *Digital dominance: the power of Google, Amazon, Facebook, and Apple*. Oxford University Press.

Patricia Moravec, Randall Minas, and Alan R Dennis. 2018. Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business Research Paper* 18-87 (2018).

Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing* 1, 2 (2018), 1–18.

Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

Luke Munn. 2020. Angry by design: toxic communication and technical architectures. *Humanities and Social Sciences Communications* 7, 1 (2020), 1–11.

Alexander Muscat and Jonathan Duckworth. 2018. WORLD4: Designing ambiguity for first-person exploration games. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. 341–351.

Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. *Policy* 9 (2017), 4–233.

Ryosuke Nagura, Yohei Seki, Noriko Kando, and Masaki Aono. 2006. A method of rating the credibility of news documents on the web. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 683–684.

An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018a. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 189–199.

An T Nguyen, Aditya Kharosekar, Matthew Lease, and Byron Wallace. 2018b. An interpretable joint graphical model for fact-checking from crowds. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

HUBERT J O'GORMAN. 1975. Pluralistic ignorance and white estimates of white support for racial segregation. *Public Opinion Quarterly* 39, 3 (1975), 313–330.

Hubert J O'Gorman and Stephen L Garry. 1976. Pluralistic ignorance—A replication and extension. *Public Opinion Quarterly* 40, 4 (1976), 449–458.

Chitu Okoli. 2015. A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems* 37, 1 (2015), 43.

R Scott Olds and Dennis L Thombs. 2001. The relationship of adolescent perceptions of peer norms and parent involvement to cigarette and alcohol use. *Journal of School Health* 71, 6 (2001), 223–228.

James M Olson and Mark P Zanna. 1979. A new look at selective exposure. *Journal of Experimental Social Psychology* 15, 1 (1979), 1–15.

Ton Oostveen, Ronald Knibbe, and Hein De Vries. 1996. Social influences on young adults' alcohol consumption: norms, modeling, pressure, socializing, and conformity. *Addictive behaviors* 21, 2 (1996), 187–197.

J Milošević orević, S Mari, M Vdović, and A Milošević. 2021. Links between conspiracy beliefs, vaccine knowledge, and trust: Anti-vaccine behavior of Serbian adults. *Social Science & Medicine* 277 (2021), 113930.

Barbara Ortutay. 2017. Replacing Disputed Flags With Related Articles. `https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/`. [Online; accessed 12-April-2023].

Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2020. Partisan polarization is the primary psychological motivation behind "fake news" sharing on Twitter. (2020).

Clifton M Oyamot Jr, Melinda S Jackson, Emily L Fisher, Grace Deason, and Eugene Borgida. 2017. Social norms and egalitarian values mitigate authoritarian intolerance toward sexual minorities. *Political Psychology* 38, 5 (2017), 777–794.

Zhongdang Pan and Gerald M Kosicki. 1993. Framing analysis: An approach to news discourse. *Political communication* 10, 1 (1993), 55–75.

Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Hee Sun Park and Sandi W Smith. 2007. Distinctiveness and influence of subjective norms, personal descriptive and injunctive norms, and societal descriptive and injunctive norms on behavioral intent: A case of two behaviors critical to organ donation. *Human Communication Research* 33, 2 (2007), 194–218.

Namkee Park, Hyun Sook Oh, and Naewon Kang. 2012. Factors influencing intention to upload content on Wikipedia in South Korea: The effects of social norms and individual differences. *Computers in Human Behavior* 28, 3 (2012), 898–905.

Josh Pasek. 2018. Don't trust the scientists! Rejecting the scientific consensus "conspiracy.". *Conspiracy theories and the people who believe them* (2018), 201–213.

Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.

Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.

Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

Gordon Pennycook and David G Rand. 2020. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality* 88, 2 (2020), 185–200.

Gordon Pennycook and David G Rand. 2021. Examining false beliefs about voter fraud in the wake of the 2020 Presidential Election. *The Harvard Kennedy School Misinformation Review* (2021).

H Wesley Perkins and Alan D Berkowitz. 1986. Perceiving the community norms of alcohol use among students: Some research implications for campus alcohol education programming. *International journal of the Addictions* 21, 9-10 (1986), 961–976.

Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.

Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2021. What Makes People Join Conspiracy Communities? Role of Social Factors in Conspiracy Engagement. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–30.

Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2022. Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 770–781.

Lara SG Piccolo, Alisson Puska, Roberto Pereira, and Tracie Farrell. 2020. Pathway to a Human-Values Based Approach to Tackle Misinformation Online. In *International Conference on Human-Computer Interaction*. Springer, 510–522.

Sara Pluviano, Caroline Watt, and Sergio Della Sala. 2017. Misinformation lingers in memory: failure of three pro-vaccination strategies. *PloS one* 12, 7 (2017), e0181640.

Francesca Polletta and Jessica Callahan. 2019. Deep stories, nostalgia narratives, and fake news: Storytelling in the Trump era. *Politics of meaning/meaning of politics: Cultural sociology of the 2016 US presidential election* (2019), 55–73.

Alexandrin Popescul, Lyle H Ungar, David M Pennock, and Steve Lawrence. 2013. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. *arXiv preprint arXiv:1301.2303* (2013).

Sunthud Pornprasertmanit, Patrick Miller, Alexander Schoemann, Yves Rosseel, C Quick, M Garnier-Villarreal, et al. 2013. semTools: Useful tools for structural equation modeling. R package version 0.4-6. *Retrieved July* 17 (2013), 2013.

Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *European Conference on Information Retrieval*. Springer, 810–817.

Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. Alto: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1158–1169.

Hafizh A Prasetya and Tsuyoshi Murata. 2020. A model of opinion and propagation structure polarization in social media. *Computational Social Networks* 7, 1 (2020), 1–35.

Deborah A Prentice and Dale T Miller. 1996. Pluralistic ignorance and the perpetuation of social norms by unwitting actors. In *Advances in experimental social psychology*. Vol. 28. Elsevier, 161–209.

Stephen Prochaska, Kayla Duskin, Zarine Kharazian, Carly Minow, Stephanie Blucker, Sylvie Venuto, Jevin D West, and Kate Starbird. 2023. Mobilizing manufactured reality: How participatory disinformation shaped deep stories to catalyze action during the 2020 US presidential election. *Proceedings of the ACM on human-computer interaction* 7, CSCW1 (2023), 1–39.

Neha Puri, Eric A Coomes, Hourmazd Haghbayan, and Keith Gunaratne. 2020. Social me-

dia and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Human Vaccines & Immunotherapeutics* (2020), 1–8.

Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on Facebook. *Available at SSRN 2795110* (2016).

George A Quattrone and Amos Tversky. 1988. Contrasting rational and psychological analyses of political choice. *American Political Science Review* 82, 3 (1988), 719–736.

Matthew Rabin and Joel L Schrag. 1999. First impressions matter: A model of confirmatory bias. *The quarterly journal of economics* 114, 1 (1999), 37–82.

Ankita Rao. 2021. Guardian News and Media. (2021, January 5). US pharmacist who tried to ruin Covid vaccine doses is a conspiracy THEORIST, police say. `https://www.theguardian.com/us-news/2021/jan/04/wisconsin-pharmacist-covid-19-vaccine-doses-steven-brandenburg.`

Yasmeen Rashidi, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. 2020. " It's easier than causing confrontation": Sanctioning Strategies to Maintain Social Norms and Privacy on Social Media. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (2020), 1–25.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1650–1659.

Susan M Reiter and William Samuel. 1980. Littering as a function of prior litter and the presence or absence of prohibitive signs 1. *Journal of Applied Social Psychology* 10, 1 (1980), 45–55.

Rajiv N Rimal and Maria K Lapinski. 2015. A re-explication of social norms, ten years later. *Communication Theory* 25, 4 (2015), 393–409.

Rajiv N Rimal, Maria K Lapinski, Rachel J Cook, and Kevin Real. 2005. Moving toward a theory of normative influences: How perceived benefits and similarity moderate the

impact of descriptive norms on behaviors. *Journal of health communication* 10, 5 (2005), 433–450.

Rajiv N Rimal and Kevin Real. 2003. Understanding the influence of perceived norms on behaviors. *Communication Theory* 13, 2 (2003), 184–203.

Alan Ritter, Oren Etzioni, et al. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. 424–434.

Margaret E Roberts, Brandon M Stewart, Edoardo M Airoldi, K Benoit, D Blei, P Brandt, and A Spirling. 2014. Structural topic models. *Retrieved May* 30 (2014), 2014.

Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *Journal of Public Health* (2021), 1–10.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.

Nathaniel Rodriguez, Johan Bollen, and Yong-Yeol Ahn. 2016. Collective dynamics of belief evolution under cognitive coherence and social conformity. *PLoS one* 11, 11 (2016), e0165910.

Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science* 7, 10 (2020), 201199.

Jon Roozenbeek and Sander van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5, 1 (2019), 1–10.

Hans Rosenberg, Shahbaz Syed, and Salim Rezaie. 2020. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation

during the COVID-19 pandemic. *Canadian journal of emergency medicine* 22, 4 (2020), 418–421.

Yves Rosseel. [n. d.]. Lavaan: An R package for structural equation modeling. `https://www.jstatsoft.org/article/view/v048i02`

Marten Scheffer, Denny Borsboom, Sander Nieuwenhuis, and Frances Westley. 2022. Belief traps: Tackling the inertia of harmful beliefs. *Proceedings of the National Academy of Sciences* 119, 32 (2022), e2203149119.

Dietram A Scheufele. 1999. Framing as a theory of media effects. *Journal of communication* 49, 1 (1999), 103–122.

Dietram A Scheufele. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass communication & society* 3, 2-3 (2000), 297–316.

Dietram A Scheufele and Nicole M Krause. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences* 116, 16 (2019), 7662–7669.

Donal Schön and Martin Rein. 1994. Frame reflection: Toward the resolution of intractable policy controversies. *Basic Book* (1994).

Christine M Schroeder and Deborah A Prentice. 1998. Exposing pluralistic ignorance to reduce alcohol use among college students 1. *Journal of Applied Social Psychology* 28, 23 (1998), 2150–2180.

Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1245–1254.

Norbert Schwarz, Lawrence J Sanna, Ian Skurnik, and Carolyn Yoon. 2007. Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in experimental social psychology* 39 (2007), 127–161.

Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 111–125.

Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New media & society* 21, 7 (2019), 1417–1443.

Colleen M Seifert. 2002. The continued influence of misinformation in memory: What makes a correction effective? In *Psychology of learning and motivation*. Vol. 41. Elsevier, 265–292.

Arloc Sherman, Danilo Trisi, Chad Stone, Shelby Gonzales, and Sharon Parrott. 2019. *Immigrants Contribute Greatly to US Economy, Despite Administration's ZPublic Chargey Rule Rationale*. JSTOR.

Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. 2017. Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *new media & society* 19, 8 (2017), 1214–1235.

Jieun Shin and Kjerstin Thorson. 2017. Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication* 67, 2 (2017), 233–255.

Jacob Silverman. 2021. Vaccine Denialism Is the Right Wing's Favorite New Conspiracy Theory. `https://newrepublic.com/article/161830/vaccine-conspiracy-theory-republicans-qanon`

Nathaniel Sirlin, Ziv Epstein, Antonio A Arechar, and David G Rand. 2021. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School Misinformation Review* (2021).

Willem WA Sleegers, Travis Proulx, and Ilja van Beest. 2019. Confirmation bias and misconceptions: Pupillometric evidence for a confirmation bias in misconceptions feedback. *Biological psychology* 145 (2019), 76–83.

Steven Sloman, Steven A Sloman, and Philip Fernbach. 2018. *The knowledge illusion: Why we never think alone*. Penguin.

Alison Smith, Tak Yeon Lee, Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Niklas Elmqvist, and Leah Findlater. 2017. Evaluating visual representations for topic understanding and their effects on manually generated topic labels. *Transactions of the Association for Computational Linguistics* 5 (2017), 1–16.

Joanne R Smith, Michael A Hogg, Robin Martin, and Deborah J Terry. 2007. Uncertainty and the influence of group norms in the attitude–behaviour relationship. *British Journal of Social Psychology* 46, 4 (2007), 769–792.

Laura M Smith, Linhong Zhu, Kristina Lerman, and Zornitsa Kozareva. 2013. The role of social media in the discussion of controversial topics. In *2013 International Conference on Social Computing*. IEEE, 236–243.

David A Snow, E Burke Rochford Jr, Steven K Worden, and Robert D Benford. 1986. Frame alignment processes, micromobilization, and movement participation. *American sociological review* (1986), 464–481.

Melodie Yun-Ju Song and Anatoliy Gruzd. 2017. Examining sentiments and popularity of pro-and anti-vaccination videos on YouTube. In *Proceedings of the 8th international conference on social media & society*. 1–8.

F Spezzano, A Shrestha, JA Fails, and BW Stone. 2021. That's fake news! Investigating how readers identify the reliability of news when provided title, image, source bias, and full articles. *Proceedings of the ACM on Human Computer Interaction journal* 5 (2021).

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

Kate Starbird, Emma Spiro, Isabelle Edwards, Kaitlyn Zhou, Jim Maddock, and Sindhuja Narasimhan. 2016. Could this be true? I think so! Expressed uncertainty in online

rumoring. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 360–371.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. In *Handbook of latent semantic analysis*. Psychology Press, 439–460.

Jonathan Stray. 2017. Defense against the dark arts: networked propaganda and counter-propaganda. *Tow Center for Digital Journalism. https://medium. com/tow-center/defense-against-the-darkartsnetworked-propaganda-and-counter-propaganda-deb7145aa76a* (2017).

Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. *Journal of communication* 60, 3 (2010), 556–576.

Sharifa Sultana and Susan R Fussell. 2021. Dissemination, Situated Fact-checking, and Social Effects of Misinformation among Rural Bangladeshi Villagers During the COVID-19 Pandemic. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–34.

S Shyam Sundar. 1998. Effect of source attribution on perception of online news stories. *Journalism & Mass Communication Quarterly* 75, 1 (1998), 55–68.

S Shyam Sundar, Silvia Knobloch-Westerwick, and Matthias R Hastall. 2007. News cues: Information scent and cognitive heuristics. *Journal of the American society for information science and technology* 58, 3 (2007), 366–378.

Cass R Sunstein. 1999. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper* 91 (1999).

Cass R Sunstein and Adrian Vermeule. 2009. Conspiracy theories: Causes and cures. *Journal of Political Philosophy* 17, 2 (2009), 202–227.

Robbie M Sutton and Karen M Douglas. 2022. Agreeing to disagree: reports of the popularity of Covid-19 conspiracy theories are greatly exaggerated. *Psychological medicine* 52, 4 (2022), 791–793.

Briony Swire, Adam J Berinsky, Stephan Lewandowsky, and Ullrich KH Ecker. 2017a. Processing political misinformation: comprehending the Trump phenomenon. *Royal Society open science* 4, 3 (2017), 160802.

Briony Swire, Ullrich KH Ecker, and Stephan Lewandowsky. 2017b. The role of familiarity in correcting inaccurate information. *Journal of experimental psychology: learning, memory, and cognition* 43, 12 (2017), 1948.

Charles S Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *American journal of political science* 50, 3 (2006), 755–769.

Piotr Tarka. 2018. An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & quantity* 52, 1 (2018), 313–354.

Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. Impact of rumors and misinformation on COVID-19 in social media. *Journal of preventive medicine and public health* 53, 3 (2020), 171–174.

Eric Taylor, Katherine E Atkins, Jan Medlock, Meng Li, Gretchen B Chapman, and Alison P Galvani. 2016. Cross-cultural household influence on vaccination decisions. *Medical Decision Making* 36, 7 (2016), 844–853.

Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N Bazarova. 2019. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

The YouTube Team. 2019. Continuing our work to improve recommendations on YouTube. `https://blog.youtube/news-and-events/continuing-our-work-to-improve/`

Temi. 2025. Temi - Transcription Services. `https://www.temi.com/` Accessed: 2025-02-19.

Hanneke A Teunissen, Renske Spijkerman, Mitchell J Prinstein, Geoffrey L Cohen, Rutger CME Engels, and Ron HJ Scholte. 2012. Adolescents' conformity to their peers'

pro-alcohol and anti-alcohol norms: The power of popularity. *Alcoholism: Clinical and experimental research* 36, 7 (2012), 1257–1267.

The Economic Time. 2020. Twitter: No more fake news: Twitter will label tweets that contain harmful, misleading content on coronavirus. https://economictimes.indiatimes.com/magazines/panache/no-more-fake-news-twitter-will-label-tweets-that-contain-harmful-misleading-content-o articleshow/75688104.cms. Online; accessed 13 January 20201.

The Washington Post. 2013. Fact Checker. https://www.washingtonpost.com/news/fact-checker/. Online; accessed 13 January 2022.

Emily Thorson. 2015. Identifying and correcting policy misperceptions. *Unpublished Paper, George Washington University. Available at http://www. americanpressinstitute. org/wp-content/uploads/2015/04/Project-2-Thorson-2015-Identifying-Political-Misperceptions-UPDATED-4-24. pdf* (2015).

Sabine Trepte and Leonard Reinecke. 2011. *Privacy online: Perspectives on privacy and self-disclosure in the social web*. Springer.

Gleb Tsipursky and Zachary Morford. 2018. Addressing behaviors that lead to sharing fake news. *Behavior and Social Issues* 27 (2018), AA6–AA10.

Amos Tversky and Daniel Kahneman. 1985. *The framing of decisions and the psychology of choice*. Springer.

Twitter. 2021. Permanent suspension of @realDonaldTrump. https://blog.twitter.com/en_us/topics/company/2020/suspension.

Irfan Ullah, Kiran S Khan, Muhammad J Tahir, Ali Ahmed, and Harapan Harapan. 2021. Myths and conspiracy theories on vaccines and COVID-19: Potential effect on global vaccine refusals. *Vacunas* 22, 2 (2021), 93–97.

Ted Underwood. 2019. *Distant horizons: digital evidence and literary change*. University of Chicago Press.

Upwork. 2025. Upwork: The Online Talent Marketplace. https://www.upwork.com Accessed: 2025-02-07.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language* 67 (2021), 101151.

Sander van Der Linden, Jon Roozenbeek, and Josh Compton. 2020. Inoculating against fake news about COVID-19. *Frontiers in psychology* 11 (2020), 2928.

Baldwin Van Gorp. 2010. Strategies to take subjectivity out of framing analysis. In *Doing news framing analysis*. Routledge, 100–125.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*. 18–22.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

Dror Walter and Yotam Ophir. 2019. News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures* 13, 4 (2019), 248–266.

Eva Walther, Herbert Bless, Fritz Strack, Patsy Rackstraw, Doris Wagner, and Lioba Werth. 2002. Conformity effects in memory as a function of group size, dissenters and uncertainty. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 16, 7 (2002), 793–810.

Claire Wardle, Hossein Derakhshan, et al. 2018. Thinking about 'information disorder': formats of misinformation, disinformation, and mal-information. *Ireton, Cherilyn; Posetti, Julie. Journalism,'fake news'& disinformation. Paris: Unesco* (2018), 43–54.

Wei Wei. 2018. The normalization project: The progress and limitations of promoting LGBTQ research and teaching in mainland China. *Journal of homosexuality* (2018).

Philipp Wicke and Marianna M Bolognesi. 2020. Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PloS one* 15, 9 (2020), e0240010.

Sam Wineburg and Sarah McGrew. 2017. Lateral reading: Reading less and learning more when evaluating digital information. (2017).

Magdalena Wojcieszak, Stephan Winter, and Xudong Yu. 2020. Social norms and selectivity: Effects of norms of open-mindedness on content selection and affective polarization. *Mass Communication and Society* 23, 4 (2020), 455–483.

Sijia Xiao, Coye Cheshire, and Amy Bruckman. 2021. Sensemaking and the Chemtrail Conspiracy on the Internet: Insights from Believers and Ex-believers. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–14.

Tuukka Ylä-Anttila, Veikko Eranti, and Anna Kukkonen. 2022. Topic modeling for frame analysis: A study of media debates on climate change in India and USA. *Global Media and Communication* 18, 1 (2022), 91–112.

Amos Yong. 2004. The Spirit bears witness: pneumatology, truth, and the religions. *Scottish journal of theology* 57, 1 (2004), 14–38.

Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*. 603–612.

Bi Zhu, Chuansheng Chen, Elizabeth F Loftus, Qinghua He, Chunhui Chen, Xuemei Lei, Chongde Lin, and Qi Dong. 2012. Brief exposure to misinformation can lead to long-term false memories. *Applied Cognitive Psychology* 26, 2 (2012), 301–307.

Bi Zhu, Chuansheng Chen, Elizabeth F Loftus, Chongde Lin, Qinghua He, Chunhui Chen, He Li, Robert K Moyzis, Jared Lessard, and Qi Dong. 2010. Individual differences in false memory from misinformation: Personality characteristics and their interactions with cognitive abilities. *Personality and Individual differences* 48, 8 (2010), 889–894.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11, 3 (2016), e0150989.

Frederik J Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H De Vreese, and Natali Helberger. 2016. Should we worry about filter bubbles? *Internet policy review* 5, 1 (2016), 1–16.

Zyte. 2021. Scrapy: A Fast and Powerful Scraping and Web Crawling Framework. Accessed: 2024-09-26.

Biography

Zhila Aghajari received her B.Sc. in Computer Engineering from Shahid Chamran University of Ahvaz and her M.Sc. in Artificial Intelligence from Nhajeh Nasir Toosi University of Technology. She received her Ph.D. degree in Computer Science at Lehigh University. During her PhD, Zhila's research focused on the societal dimensions of misinformation and its broader impacts. Her work has been published in premier venues, including the ACM Conference on Human Factors in Computing Systems (CHI) and the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW). Zhila also has served as a teaching assistant for multiple undergraduate and graduate courses, including Algorithms and Data Structures. Zhila's professional experience includes a research scientistic internship at eBay, where she led the design and development of a multi-stage framework that leverages LLMs and Generative AI image models, to automatically generate images for different categorizes for eBay online marketplace. She also developed an innovative automated evaluation pipeline to evaluate these automatically generated images using a multi-modal large language model. During her PhD, Zhila has received honors such as NSF funding for her research proposal at CSCW Doctoral Consortium, Upsilon Pi Epsilon membership, International Honor Society for Computing. Zhila also served as the CSE Department Representative, the Graduate Student Senate, as well as the Vice President of Women in Science and Engineering (WiSE).