The Pennsylvania State University
The Graduate School

EMPIRICAL STUDIES ON PLATFORM-DRIVEN AND USER-INITIATED METHODS FOR MISINFORMATION CORRECTION

A Dissertation in
Informatics
by
Haeseung Seo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2023

The dissertation of Haeseung Seo was reviewed and approved by the following:

Dongwon Lee
Director of Graduate Program
Professor of Information Sciences and Technology
Dissertation Co-Advisor, Co-Chair of Committee

Aiping Xiong
Assistant Professor of Information Sciences and Technology
Dissertation Co-Advisor, Co-Chair of Committee

Kelley Cotter
Assistant Professor of Information Sciences and Technology

Bu Zhong
Professor of Journalism and Communications

Abstract

Due to the proliferation of social media, the amount of online misinformation has been escalating and becoming a societal concern. It has caused social divisions through false political news and led people to fatal outcomes with incorrect treatment methods. While significant efforts have been invested in misinformation detection through human fact-checkers and machine-learning algorithms, the study on correcting misinformation and preventing its dissemination has not received adequate attention.

Misinformation correction refers to the act of informing users about the inaccuracy of specific information to alter their perception and/or behavior toward misinformation. We need to prioritize the exploration of effective misinformation correction methods for users since they serve as both consumers and disseminators of information on social media. It is vital to prevent users from unwittingly disseminating misinformation by ensuring their awareness of its inaccuracies.

In order to explore misinformation correction methods that can effectively reach more users, I conducted online human subject experiments through three studies that encompassed a total of eight experiments. I focused on two correction agents on social media and, accordingly, conducted research on misinformation correction through the following two approaches: (1) platform-driven correction and (2) user-initiated correction.

The first approach is from the social media platform perspective. I researched which types of platform warnings can effectively assist users in identifying misinformation. I found that, with the absence of source information, a machine learning-driven warning with explanations enhances users' capability to identify fake news. Additionally, I conducted a more thorough investigation into the explanatory section, drawing inspiration from the framing effect, and confirmed the effectiveness of negative framing. The second approach is from the users' perspective. I explored four types of user-initiated correcting comments with reliable sources. All these types substantiated the effectiveness of correcting comments.

The following research questions are addressed in this dissertation: (1) As a platform-driven correction, can machine learning warnings enhance the ability to discern misinformation? (2) Are platform-driven warnings with explanations more effective compared to a warning without explanation? What factors influence the effectiveness of platform-driven warnings? (3) As a user-initiated correction, can users effectively correct misinformation through comments?

My studies empirically demonstrated that a machine-learning warning with explana-

tions can effectively correct misinformation. Furthermore, I underscored the valuable role that a user can play in hindering misinformation dissemination by leaving correcting comments. My research holds significance in presenting effective misinformation correction methods through multiple iterations of experiments that recruited a substantial number of participants from a systematic point of view. In the future, I anticipate conducting studies incorporating elements from a broader range of real-life situations, building upon the foundation of this research.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My doctoral journey was a period of self-reflection where I confronted my limitations, both within and beyond the academic realm. The most significant achievement during this time was gaining a profound understanding of my shortcomings, and what I am most proud of is my determination not to give up on the journey to address those deficiencies. In those moments, I firmly planted the roots of my quest for truth. I want to express my heartfelt gratitude to those who gave me love and support during those times.

I am deeply grateful to my co-advisors, Dr.Dongwon Lee and Dr.Aiping Xiong, for their unwavering support and guidance throughout this journey. They were excellent advisors to me and admirable "academic parents." I am fortunate to have had the guidance from them. Through them, I learned what the meaning of research and the attitude of a researcher are up close. Furthermore, I learned what the life of a true researcher entails through their lives. They always respected my opinions and never hesitated to impart a wealth of knowledge and wisdom so that I could grow into an independent researcher. In particular, through Dr.Dongwon Lee, I was able to broaden my horizons in informatics and engage in intellectual growth alongside brilliant peers at the PIKE lab. Through Dr.Aiping Xiong, I learned from A to Z about human experimental research and ways of conducting research.

I extend my sincere thanks to my committee members, Dr.Bu Zhong and Dr.Kelley Cotter. Their continuous support and insightful feedback significantly contributed to my research. They provided meticulous advice to ensure the completeness and quality of my work. Their guidance has enriched my research journey. Through them, I gained a more comprehensive perspective on my research questions from a communication studies standpoint.

Furthermore, I would like to express my appreciation for my supportive colleagues. I am especially grateful to Sian Lee, a valued research collaborator. Throughout our collaboration on various projects, we have shared invaluable insights, reinforced each other's work, and collectively nurtured our academic growth. Limeng Cui, despite being a younger colleague, consistently stood as an intellectual and inspirational leader, providing kind support whenever needed. Her presence significantly enriched my doctoral journey, making it more interesting.

I would also like to express my gratitude to the numerous amazing Pike Lab colleagues, including Jason Zhang, Yiming Liao, Thai Le, and others, as well as fellow doctoral students. They provided constant, constructive input, encouragement, and support

throughout my research journey.

I must express my gratitude to my friends who have been instrumental in propelling me forward throughout my entire journey. Bora Nam, Mihee Kim, Sooyeon Park, Nayoung Park, Seongryung Kim, Minjeong Kim, Nayeon Kim, Pyeonghwa Kim, Jaehyun Park, Carla Mullen, Yanglan Ou, and many others have wholeheartedly supported my life choices and lent a sympathetic ear to my story. Without their love and encouragement, completing this journey would have been more challenging.

I wish to convey my genuine appreciation to my special motivator and soulmate, Fengting Yang. He is a seasoned researcher with a steadfast commitment to his work, exceptional research skills, diligence, and a visionary mindset. He has never hesitated to offer his advice and encouragement to me, consistently serving as a guiding colleague. Through him, I was able to grow not only as a researcher but also as a person. I am thankful for all the time I spent in Happy Valley, where I had the opportunity to study alongside him.

I wish to extend my deepest thankfulness to our family, who consistently stood by my side, offering boundless love and unwavering courage. To my beloved grandparents, who instilled in me the true essence of love; to my parents, the driving force and endless source of positivity in my life; to my lovely siblings who are always supportive; to Grandma Glory and Uncle Seungjoon, who provided steadfast support for my life in the United States; to my husband, the warm sunshine of my life; and to all our extended family members, I convey my heartfelt gratitude. It is due to the consistent support of my entire family that I stand here today, able to complete this dissertation. My heart overflows with gratitude.

The research presented in this dissertation was partly supported by the NSF award (#1742702, #1820609, #1915801, and #2121097), ORAU-directed R&D program in 2018, and Penn State SSRI Seed Grant. The findings and conclusions in this dissertation do not necessarily reflect the view of the funding agency.

Dedication

To my grandmother, Sangjin Lee; my grandfather, Changho Seo; my mother, Jenongboon Seo; my father, Kwangsoo Seo; my sister, Eunbi Cho; my brother, Jonghyun Seo; and my husband, Fengting Yang, for their unconditional love and support.

Chapter 1 Introduction

The development of social media has not only facilitated the dissemination of valuable information but has also exacerbated the spread of misinformation [1, 2]. Misinformation is not a unique phenomenon of our time; it has always been a societal issue in human history [3, 4]. Before the rise of social media, misinformation was spread through various means such as word of mouth, newspapers, television, books, etc [3–5]. However, the distinctive information characteristics of social media, such as the ease of information production for anyone, the rapid dissemination within one's network, and no proper fact-checking mechanisms [6, 7] have led to information spreading at an exponential rate. These unique features have made various forms of information on social media spread so quickly that they can become significant societal issues.

Such misinformation can arise across all kinds of topics, ranging from politics, public health, science, etc. Any misinformation that distorts facts can confuse people and deprive them of the right to know the truth. Among these, misinformation related to politics and public health can become a particularly serious social issue. For instance, in the period of the 2016 U.S. presidential election, numerous fake news stories were disseminated for political purposes and demonstrated their potential to divide society [1]. Also, during the recent COVID-19 pandemic, the world experienced an unprecedented period of turmoil, and the fear surrounding the novel virus led to the proliferation of false treatments, which costs lives in extreme cases [2].

In response to the deleterious consequences caused by misinformation, many organizations [8], researchers [9,10], and social media companies [11,12] have dedicated their efforts to various studies aimed at combating it. Fact-checking websites such as *snopes.com* and *politifact.com* are actively employing human fact-checkers to assess information veracity.

[1] https://www.ischool.berkeley.edu/news/2020/hany-farid-how-disinformation-dividing-nation
[2] https://www.bbc.com/news/world-53755067

A number of researchers have engaged in misinformation detection and developed powerful algorithms to identify such information automatically [13–15]. Their work spans a broad range of techniques, including machine learning, natural language processing, and the assessment of source credibility and social networks. Leading social media platforms like Facebook and Twitter (recently renamed as "X") have also committed substantial resources to curtail the spread of misinformation, including but not limited to adopting the efforts from fact-checking websites and misinformation detection algorithms [3].

Identifying misinformation is an essential step in reducing misinformation, but considering that the ultimate consumers of information are users, the importance of misinformation correction should not be underestimated. The efficacy of misinformation detection cannot be fully realized until end users actively recognize and acknowledge certain content as misinformation. Active research on misinformation correction should run in parallel with misinformation detection studies.

Misinformation correction refers to the act of informing users about the inaccuracy of specific information to alter their perception and/or behavior toward misinformation. Misinformation correction helps users distinguish the false messages delivered in online posts, rumors, fake news, etc, and it entails disseminating accurate information. Nevertheless, the research on misinformation correction has, to date, remained relatively restricted.

While some studies have explored misinformation correction and the effectiveness of warning messages [5, 16–18], there is limited experimental research addressing misinformation correction in the context of social media, platforms notorious for the rapid spread of false these days. Therefore, I conducted experiments to systematically investigate misinformation correction on social media and identify practical and effective strategies for real-world applications.

I explored platform-driven correction and user-initiated correction, focusing on the two key actors in social media: the platform and the user. Platforms, primarily represented by social media companies such as Facebook and Twitter [19], often correct misinformation through warning messages. Users, as central actors reflecting the identity of social media, have various means of expressing their opinions on social media [6, 7] and can engage in misinformation correction in different ways. Within user-initiated correction, I focused on comments as the most explicit and direct way to respond to misinformation posted.

Meanwhile, based on the source of correction, I divided misinformation sources into human fact-checkers and machine detectors. Human fact-checkers are exemplified by

[3]https://ai.meta.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/

fact-checking websites like *snopes.com* and *politifact.com*, where humans verify news articles one by one. Machine detectors refer to machine learning algorithms developed to automatically filter out fake news.

In platform-driven correction, I focused on warnings, which is one of the most representative forms of platform-driven correction. Social media platforms are actively engaged in the development of algorithms to filter out vast amounts of misinformation efficiently. However, research on whether users are receptive to such algorithm-based warnings (i.e., machine detectors-based correction) remains scarce. Existing studies [20,21] on social media warnings predominantly examined the effectiveness of human fact-checker warnings. It is necessary to investigate the efficacy of platform-driven correction in the form of machine-learning-driven warnings and compare it against fact-checker-driven warnings. In accordance with this, my first research question is:

RQ1. As a platform-driven correction, can machine learning warnings enhance the ability to discern misinformation?

Moreover, such machine-learning-driven warnings demand explainable artificial intelligence (XAI). XAI advocates users' right to know and promotes users' trust in AI judgments [22, 23]. When users face AI machine judgments, they may wonder about the underlying logic. A natural way to fill the gap is to provide the corresponding explanations along with AI-generated warnings (i.e., machine-learning-driven warnings). However, it is unclear how we can frame such explanations to optimize their effectiveness. The framing effect [24] suggests that different ways of conveying the same message can influence human decision-making processes differently. I investigated the effectiveness of the explanation-enhanced machine-learning-driven warnings and compared the effectiveness of positive and negative framing in explanations. Accordingly, I further ask the following questions:

RQ2. Are platform-driven warnings with explanations more effective than a warning without explanation? What factors influence the effectiveness of platform-driven warnings?

Despite the efforts that social media makes to prevent misinformation distribution, users, the direct customers of online information, play a key role in shaping social media. Research on ways to encourage users to participate in misinformation correction is desired. Many users have experienced misinformation correction, either directly or indirectly [25]. However, the effectiveness of such user-initiated correction is underexplored. The

effectiveness of corrections can vary based on the means users employ (e.g., posts, comments, likes, shares) and their framing of the correction messages. I focused on comments as the most direct and clear means to convey messages related to misinformation postings and aimed to find effective forms of correcting comments with the following research question.

RQ3. As a user-initiated correction, can users effectively correct misinformation through comments?

To tackle these research questions, I carried out three studies and addressed them in three separate chapters (See Figure 1.1). Chapter 3 demonstrates the first study that examined the effectiveness of machine-learning warnings as a platform-driven correction. I conducted two experiments on Amazon Mechanical Turk using twenty-four pieces of political news. In the first experiment, I found that human fact-checker warnings were effective when fake news had news source information provided. In the second experiment, I demonstrated that machine-learning warnings accompanied by explanations were effective when news source information was not provided. I also confirmed that the most trusted warning came from fact-checker warnings across two experiments.

Chapter 4 covers the second study, in which I extended the first research by exploring the effectiveness of AI explanations in platform-driven warnings. I conducted a total of three online experimental studies on Amazon Mechanical Turk using twenty-four pieces of COVID-19 misinformation. The experiment results revealed a tendency for negatively framed explanations to decrease participants' perceived accuracy ratings of fake news. Additionally, participants showed a high reliance on warning messages without explanations when presented within an AI system that declared high reliability. Besides this, all three experiments consistently showed that the most trusted warning was the warning message without an explanation.

Chapter 5 focuses on user-initiated correction in response to platform-driven correction. I examined the effectiveness of four types of user-initiated correcting comments using a total of 20 COVID-19 misinformation news. The results of three experiments showed that correction messages from health organizations, fact-checking websites, and individual users were all effective. Post-analysis results indicated that the effectiveness of correcting comments was independent of who corrected, and participants regarded the reliability of the correction as a critical criterion. Participants cared either "who wrote the comment" or "whether the comment included a reference URL" the most, depending on the conditions they belonged to. In other words, no matter whether it is an individual or an institution,

social media users can effectively engage in user-initiated correction as long as they demonstrate the reliability of fact-checking.

In Chapter 6, I conclude this dissertation by summarizing my research, discussing my contribution, implications, and limitations, and outlining future research directions.

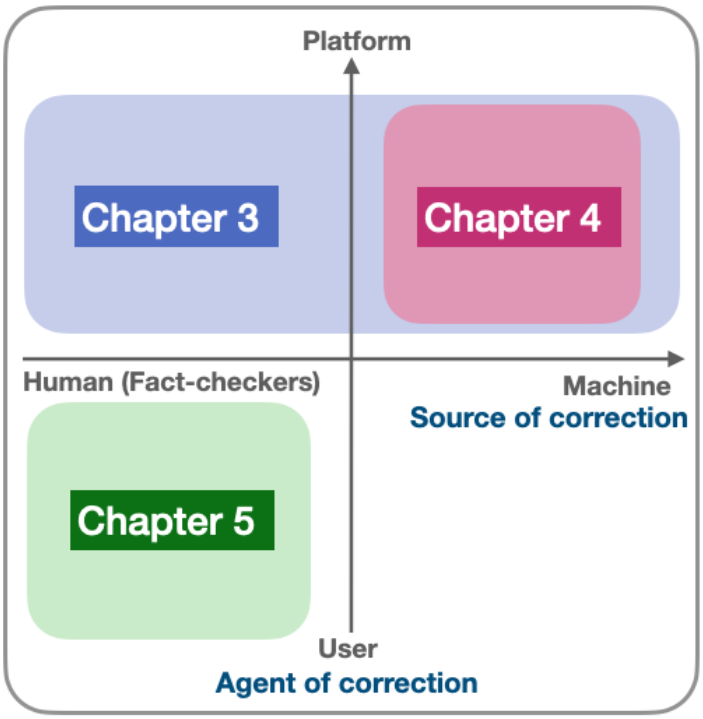

Figure 1.1: This four-quadrant graph illustrates the source and agent of correction on its axes, with corresponding chapters addressing different correction scopes. Chapter 3 explores platform-driven correction through human fact-checker warnings and machine-learning warnings. Chapter 4 discusses platform-driven warnings, focusing on machine-learning warnings. Chapter 5 focuses on user-initiated correction comments that reference the judgment of human fact-checkers.

Chapter 2
Related Works

2.1 Misinformation on Social Media

2.1.1 Definition of Misinformation

Misinformation has garnered increasing attention in various academic communities, particularly within the fields of communication [26,27], psychology [5,20], and information science/computer science [14, 28]. Comprehending the core nature of misinformation is pivotal in the context of today's information landscape, where the rapid dissemination of information through digital media platforms has become indispensable.

Despite the existence of various definitions, the term "misinformation" generally pertains to the dissemination of false or inaccurately crafted information, irrespective of whether there is a deliberate intent to mislead [29,30]. It often signifies information lacking clear evidence and expert consensus [26, 31]. Occasionally, the term "disinformation" is used interchangeably with "misinformation," but they bear distinct connotations. "Disinformation" specifically denotes information that is consciously fabricated with the intent to deceive [32], whereas "misinformation" encompasses inaccurate information shared without the deliberate intent to mislead [33, 34].

Despite this distinction, Treen et al. [35] contend that the nature of misinformation within the context of social media poses challenges in determining intent during the dissemination process. As a result, they advocate for a comprehensive definition of misinformation, encompassing inaccurate information shared regardless of intent. This expansive interpretation recognizes the intricacies of intent and context in the age of digital communication, where the line between misinformation and disinformation becomes hazy.

According to a recent study [36] that explored the definition of misinformation with

the insights of 150 experts, it was revealed that 43% of the experts defined misinformation as false and misleading information. Following this, 30% of the experts defined misinformation as false and misleading information that is disseminated unintentionally. An interesting observation is that a substantial portion of quantitative experts prefer the former definition, whereas most qualitative experts tend to favor the latter definition. In this dissertation, I adopt the flexible approach by the majority, defining misinformation as false and misleading information without regard to intention.

2.1.2 Types of Misinformation

From a more comprehensive viewpoint, misinformation is a multifaceted phenomenon that can be categorized into various distinct types, each with its own characteristics and implications. These categories encompass a wide range of deceptive or misleading information that can significantly impact public perception and understanding. The types of misinformation in this context include conspiracy theories, rumors, clickbait headlines, deepfakes, and fake news [28, 32, 36–38]. These categories are components within the larger misinformation landscape, requiring thorough analysis and specialized approaches for detection and prevention. Grasping the nuances of these misinformation forms is vital for fostering media literacy, critical thinking, and responsible information consumption.

Conspiracy theories, as revealed by recent studies [32, 36, 39], constitute a distinctive category of misinformation. Conspiracy theories typically center on covert and sometimes nefarious actions attributed to groups, often with scant corroborating evidence. They are marked by the widespread belief that individuals in positions of power purposefully mislead the public. Within this category, one can find a diverse array of narratives, ranging from allegations of clandestine government activities to assertions of concealed actions by influential entities.

Rumors, another facet of misinformation [36, 40], represent informal transmissions of information or narratives that have not been substantiated through credible sources. They circulate from one individual to another through various means, such as word of mouth, social media, or even traditional media outlets. These narratives frequently pertain to events, circumstances, or individuals, and their accuracy may vary significantly. Rumors can encompass truths, partial truths, or outright falsehoods.

Clickbait headlines, a prevalent form of digital misinformation [36, 41], are designed primarily to capture interest and entice visitors to follow a link to a specific webpage. They frequently use sensational words, alluring visuals, or overstated assertions to capture

one's interest. Clickbait is prevalent across a range of online platforms, spanning from social media to news websites. Grasping the methods employed in clickbait and the psychology driving its effectiveness is essential for devising approaches to reduce its impact.

Deepfakes, a more recent addition to the realm of misinformation [36, 42], result from the application of artificial intelligence (AI) techniques. These techniques involve the blending, amalgamation, substitution, and overlaying of images and video segments, resulting in the creation of counterfeit videos that convincingly emulate reality. Deepfakes carry the capability to mislead the public by producing authentic-looking videos where individuals appear to say or do things they never actually did. They present distinctive challenges in the battle against misinformation, necessitating advancements in both detection techniques and countermeasures.

Fake news, a term that gained prominence in recent years [3, 36], refers to falsified content that emulates the news media format but deviates from its organizational procedures and underlying objectives. Fake news can be fabricated with the intent to deceive, mislead, or exploit readers. It can be created and disseminated for various purposes, including political manipulation, profit, or to generate attention. Understanding the motivations behind the creation and dissemination of fake news is vital to mitigate misinformation online.

Among these types of misinformation, my research has predominantly focused on fake news, a key misinformation type that has propelled the public's acknowledgment of misinformation as a societal issue with the development of social media. Even the terms "fake news" and "misinformation" are sometimes used interchangeably without differentiation [43, 44]. In the following section, I will explore social media and fake news in greater detail.

2.1.3 Social Media and Fake News

The rise of misinformation has emerged as a consequential challenge, accelerated by the rapid expansion of social media platforms, as noted in previous studies [1, 2]. As social media continues to advance, the dissemination of misinformation has become increasingly effortless, amplifying its impact through social media, which often serves as the epicenter of its propagation [1].

Within the realm of social media, one prevalent manifestation of misinformation is fake news, which closely mimics the structural elements of authentic news articles. Fake news, as a subcategory of misinformation [3,30], is characterized as fabricated information

that imitates the format of news media but lacks the same rigorous editorial processes and intent [3]. Alternatively, it can encompass news articles that are intentionally and verifiably false, capable of misleading readers by emulating the presentation of legitimate news reports [1, 45]. Fake news often originates in motivation driven by financial gain or ideological motives [28].

The prominence of fake news has grown in parallel with the expansion of social media and mobile communication since 2008 [46], reaching global notoriety, particularly in the aftermath of the 2016 U.S. presidential election [1, 3, 47, 48]. Notably, during this period, some of the most widely circulated fake news stories on platforms like Facebook began to outperform even the most popular mainstream news articles in terms of shares [49].

Social media, with its distinctive characteristics that set it apart from traditional media, inherently provides an environment conducive to the dissemination of fake news. A critical distinction lies in the absence of third-party oversight, fact-checking, and editorial judgments, which are inherent in traditional media [1]. Social media platforms also do not impose the same professional standards expected of journalists, who are bound by the responsibility to deliver thoroughly objective information [50]. Instead, social media empowers users to share content, ostensibly granting them the freedom of expression [51]. Driven by the freedom afforded to users and the extensive reach inherent to information diffusion [51], these individuals find themselves largely unrestricted in their information-related conduct on social media platforms.

Within the dynamic realm of social media, where a plethora of posts flood the digital landscape daily, discerning the authenticity of shared information becomes a formidable and intricate challenge [28, 52]. Furthermore, the prevalence of echo chambers [53] and filter bubbles [54] on social media confines users to information niches that align with their preexisting interests, resulting in a progressively narrower perspective. This, in turn, can diminish their ability to assess information objectively from an external standpoint [1, 55] and can make them spread misinformation in line with their preferences, often without thoroughly assessing the accuracy of the information they share [56]. In the meantime, the low initial expenses associated with entering the market and creating content intensify the comparative profitability of the small, short-term methods frequently adopted by purveyors of fake news. This diminishes the motivation to establish a lasting reputation for delivering high-quality content [1]. With the characteristics of social media, misinformation has easily spread on social media platforms to the extent of being recognized as a societal issue [48].

In this dissertation, I use the terms “misinformation” and “fake news” interchangeably

to refer to false information disseminated through social media. This choice is influenced by the common practice in the literature where these terms are often used interchangeably, as highlighted in previous research [44]. Also, drawing a clear demarcation between these two concepts can be a complex task. Since my research centers on fake news, a prominent type of misinformation that garners attention on social media, I use the terms 'misinformation' and 'fake news' interchangeably to refer to false information presented in a news-like format, irrespective of the intent behind it.

2.1.4 Severity of Misinformation Problem

The infiltration of numerous fake news stories into social media platforms emphasizes the seriousness of the misinformation problem. In critical instances, this issue can give rise to significant societal challenges. For example, the widespread dissemination of false information during pivotal societal events, such as elections and the COVID-19 pandemic, has not only sowed political discord within societies [57] but has also, in tragic cases, led to significant loss of life [58].

One compelling example of the political chaos spurred by misinformation is a fabricated news report falsely asserting that the Pope had endorsed Donald Trump for President[1]. This misinformation primarily propagated through social media during the intense political campaign leading up to the U.S. presidential election, garnering significant global attention [45, 49]. The Pope, as a highly influential figure in global religious leadership, further amplified the impact of this deceptive news article, fomenting political turmoil. The incident, characterized by the alleged endorsement from a prominent religious authority, injected additional complexity and division into an already fiercely contested electoral landscape.

The COVID-19 pandemic is another crucial societal event where misinformation has caused confusion within society. People grappled with unprecedented fear and anxiety as they sought to acquire information about a virus previously unknown to them [59, 60]. The challenge lay in distinguishing authentic information from the countless undisclosed facts and the ever-evolving data [59,60]. In a landscape rife with information of uncertain accuracy, individuals found themselves overwhelmed and, in some instances, compelled to place their trust in what they had. Tragically, this reliance on misinformation led to dire consequences [58]. At least 5, 800 people were hospitalized for following incorrect treatments, while many other cases involved individuals consuming methanol or cleaning

[1]https://web.archive.org/web/20161115024211/http://wtoe5news.com/us-election/pope-francis-shocks-world-endorses-donald-trump-for-president-releases-statement/

products containing alcohol, leading to fatal outcomes [2].

Hence, there is a pressing need for proactive measures to mitigate the widespread dissemination of fake news, which has the potential to induce significant societal confusion. In the forthcoming section, I will delve into a detailed discussion of approaches aimed at mitigating misinformation, offering valuable insights into tackling this complex issue.

2.2 Mitigating Misinformation

Researchers have explored two approaches aimed at alleviating the adverse effects of misinformation: 1) misinformation detection and 2) misinformation correction. Misinformation detection involves the identification of content deemed to be misinformation [28], while misinformation correction entails informing the recipient of the information about the falsehood that has been uncovered [5].

Identifying misinformation can be accomplished by two groups: human fact-checkers and machine detectors. Human fact-checkers may be individuals, but it is primarily non-profit organizations specializing in fact-checking, such as *snope.com* and *politifact.com* [21, 61]. On the other hand, machine detectors predominantly refer to computer-based algorithms developed by platform companies or researchers [28, 62].

Considering the extensive volume of online information, it is challenging for people to individually detect various forms of manipulated misinformation [13–15]. Consequently, a plethora of misinformation detection methods have emerged in recent years, primarily led by computer scientists. These techniques encompass diverse features, spanning both single-modal and multi-modal approaches.

Single-modal strategies predominantly focus on analyzing textual elements within news articles. For instance, some methods involve quantifying assertive language, which is often more prevalent in untrustworthy sources [63]. Others assess the coherence between a news article's topic sentence and its main body [64].

In contrast, multi-modal approaches integrate features from a wide array of sources, including the content of news articles, the identities of those who disseminate news, the publishers of news content, and the patterns of news propagation in networks. These features may comprise various textual attributes, such as the content of news articles and user comments [65], or they may encompass diverse data types, including combinations of text, images, or videos [66–68].

Nonetheless, it is essential to acknowledge that the primary actors responsible for

[2] https://www.bbc.com/news/world-53755067

acquiring and disseminating information on social media are ordinary users. Consequently, equipping these users with the necessary tools to identify and prevent the spread of false information is paramount. This underscores the importance of research endeavors dedicated to proactive misinformation correction. So far, the field of misinformation research has predominantly revolved around detection, with relatively fewer studies focusing on correction. I have particularly directed my research efforts toward misinformation correction while recognizing its pivotal role within the broader scope of misinformation mitigation. In the forthcoming section, I will provide an in-depth exploration of misinformation correction.

2.3 Misinformation Correction

While it may be challenging to pinpoint an exact definition of misinformation correction, it can be broadly grasped as the act of informing users about the inaccuracy of specific information to alter their perception and/or behavior regarding misinformation.

Existing experimental studies on misinformation correction have yielded mixed results, potentially due to variations in the specific misinformation correction formats utilized in each study [16]. While some studies showed that correction can significantly reduce participants' misinformation belief [26, 69], the other studies showed negative effects [31, 70], or combined effects simultaneously [71, 72].

It is crucial to comprehend why this task is challenging to identify effective methods for correcting misinformation better [73–75]. The primary obstacle often lies in what is known as the continued influence effect of misinformation. This phenomenon refers to the lasting impact of misinformation on people's beliefs and judgments, even after corrections have been provided [18, 76]. Several factors contribute to the continued influence effect, including mental models, retrieval failure, fluency, familiarity, and reactance [5, 29–31, 75, 77]. Notably, in some cases, attempts to correct misinformation can unintentionally exacerbate the problem, leading to what is referred to as backfire effects [5, 31].

To address this challenge, Lewandowsky and his colleagues [5] proposed what to improve misinformation correction. Their strategies include the use of alternative accounts, repeated retractions, emphasizing factual information, and providing pre-exposure warnings [5]. In a meta-analysis conducted by Walter and his colleagues [16], various effective correction methods were identified, such as enhancing source credibility, employing fact-checking procedures, and issuing general warnings.

Meanwhile, it is important to note that most of the existing studies on misinformation correction have not been tailored to the nuances of social media contexts. As previously discussed, recent misinformation-related issues primarily revolve around social media, and given the unique media characteristics of social platforms [1], it is imperative to conduct research that thoroughly examines the effectiveness of misinformation correction within this social context.

Therefore, I aim to explore effective misinformation correction methods within the realm of social media, taking into account the unique characteristics of these platforms. Specifically, I focus on two key actors involved in the act of misinformation correction within social media: platforms and users. Centering my approach around these actors, I categorize social media correction into platform-driven correction and user-initiated correction.

2.3.1 Platform-Driven Correction

Social media has been criticized as a primary breeding ground for the spread of fake news [1, 78]. In response to this criticism, social media platforms have made efforts to reduce misinformation. Prominent social media platforms such as Facebook or Twitter (recently renamed as "X") have implemented misinformation correction by displaying warnings [79, 80] or related articles [81] to users in connection with misinformation.

Several studies confirmed the effect of warnings on fake news in correcting misinformation [21, 79, 82]. Pennycook et al. found a warning can reduce people's willingness to share fake news [79] and decrease the perceived accuracy rating of fake news [83] even though unlabeled false stories can be perceived as genuine due to the implied truth effect. Methods of presenting warnings include a general warning message that cautions against misinformation in general [21], specific warnings for each news item [79], warning approaches that guide towards factual information through related stories [27], etc.

Overall, existing studies assume scenarios in which human fact-checkers detect misinformation and subsequently provide warnings. However, social media platforms are actively developing machine learning algorithms to adapt to the evolving and diverse forms of misinformation. Instances of these detectors identifying fake news and appending warnings to them are becoming increasingly common. Therefore, in the pursuit of identifying effective platform-driven correction strategies within the social media landscape, it becomes imperative to examine whether machine-learning warnings can effectively resonate with users compared to the impact of traditional human fact-checker warnings. In light of this, Chapter 3 delves into an experimental study designed to evaluate the

effectiveness of machine-learning warnings as a response to fact-checker warnings within the framework of platform-driven correction.

Additionally, as AI technology continues to advance, the expectations for platform-driven correction are on the rise [22, 23, 84–87]. In an era where algorithm-based decision-making prevails online, users are increasingly seeking explanations for the underlying reasons behind these decisions [84, 88]. This demand aligns with the principles of explainable artificial intelligence (XAI), which focuses on making AI systems' outcomes more comprehensible for humans and providing them with understandable explanations about how the system operates [22, 23, 84–86]. To meet these recent needs, advanced machine learning techniques are essential in enhancing the transparency and comprehensibility of AI systems [87]. When algorithmic reasoning lacks clarity or transparency, it can lead to user distrust in AI systems [88]. Considering these current demands, machine learning warnings should not neglect the inclusion of explanations [84]. Consequently, Chapter 4 extends beyond the machine learning warnings introduced in Chapter 3 and explores research on machine learning warnings that are accompanied by explanations.

2.3.2 User-Initiated Correction

User-initiated correction refers to actions taken by individual social media users to rectify instances of misinformation. Given that users play a central role in disseminating information on social media platforms [89, 90], it is plausible that user-initiated correction necessitates even more proactive research efforts compared to platform-driven correction.

Many users are already directly or indirectly involved in the process of misinformation correction on social media. A study conducted by Bode and Vraga (2021) provides insight into people's experiences with COVID-19 misinformation correction on social media. Out of 1,094 participants, 34% reported witnessing the correction of others' misconceptions, while 22% actively engaged in correcting misinformation propagated by others. This demonstrates the active role users play in the correction process.

User-initiated correction can manifest through various means, such as comments or posts [26, 91]. Analyzing the linguistic aspects of misinformation correction reveals that users frequently employ formal language when delivering corrections on social media platforms [92]. In addition, linguistic cues used during fact-checking tend to convey information more positively, reducing the uncertainty associated with a particular fact [72].

Research on user-initiated correction has been explored from various perspectives, with notable contributions from Vraga and Bode [26, 69, 93], which will be discussed in

Chapter 5. In their recent experiments [26, 69, 93], Vraga and Bode investigated user-initiated correction within a simulated social media environment, specifically through user comments. Their findings suggested that social correction, which involves users actively correcting misinformation, can be effective when it includes sufficient source information. This information comprises elements like the source's logo (representing organizations such as *snopes.com* and CDC), a refuting headline, a supporting sentence, and reference links from these authoritative organizations [69, 93]. They also determined that the platform (Facebook or Twitter) where user-initiated corrections are published is irrelevant [93], and both social and algorithmic corrections are equally effective in mitigating misperceptions [69].

Additionally, they conducted a comparative analysis of social corrections and corrections provided by reputable organizations like the CDC. Interestingly, they found correction effects when the CDC corrected misinformation and also when the CDC corrected after an individual user had done so [26]. In this particular study, they simplified the comment manipulation process by including a debunking sentence with a reference link from the CDC. However, it is worth noting that they did not observe a correction effect from a single individual user alone [69,93]. Instead, correction effects were observed when individual users referred to different sources [69, 93].

In Chapter 5, I expanded upon the research influenced by Vraga and Bode, delving deeper into the examination of the effects of user-initiated correction on social media. My primary objective was to provide substantial evidence regarding the impact of single user-initiated correction, a factor that has not been conclusively validated, by utilizing a more extensive range of news materials and a larger participant base in my experiments.

Chapter 3
Platform-Driven: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation

In this chapter, I present the initial experimental study conducted on platform-driven corrections. The primary focus of this study is to assess the effectiveness of warning messages as a prominent form of platform-driven correction. Warnings can be generated by either human fact-checkers or machine-learning algorithms, but limited research has addressed the effectiveness of warning messages produced by machine-learning algorithms. To bridge this deficiency, we compare the effects between fact-checking warnings (i.e., human fact-checkers warnings) and machine-learning warnings (i.e., machine detector warnings). Additionally, we introduce three distinct designs for machine-learning warnings. The chapter encompasses the findings of two separate experiments.

3.1 Introduction

We currently live in a historical era called the "information age". The advent of modern information technology fundamentally changes the ways people access, communicate, and share information. Specifically, the rise of the Internet and, more recently, social media platforms (e.g., Facebook, Twitter) have made it possible for individuals to produce, consume, and share diverse multi-modal information (e.g., text, picture, video). With the boundary between information source and information receiver becoming blurred and often invisible, then, issues arise with regard to the quantity and quality of the information to which people are exposed [94]. Especially, it must be acknowledged that people are not necessarily good at evaluating the quality of online information. *Fake*

News often refers to (intentionally) false stories or fabricated information written and published for various incentives, including political agenda or financial gain [95–97]. In recent years, the spread of fake news has been identified as a major risk for individuals and society [28]. For instance, fake news has fostered people's bias and false belief of climate change [34] and greatly influenced elections and democracies [98].

Two venues of approaches have been investigated to mitigate the negative impacts of fake news: (1) computation-based detection and prevention of fake news, and (2) decision-aid methods to warn users when a piece of fake news has been identified. Among the latter venue of approaches (the focus of this study), attaching warnings to the news that was suspicious or fact-checked to be fake news was implemented to discourage users' consumption and belief in fake news. One such example once was used by Facebook. While some studies showed that exposure to a fact-checking warning under Facebook-style headlines reduced the perceived accuracy of fake news compared to a control condition [21], other studies did not [79], motivating our study.

Also, with more fact-checking work being done by machine learning algorithms [28], one interesting but rarely investigated question-related to both venues is: *After computational methods detect fake news, how to convincingly present the result to users to make informed decisions consequently?* To answer this intriguing question, we investigate the following research questions:

1. RQ1: Will the presence of a fact-checking warning increase participants' fake news detection relative to a control condition in which there is no warning?

2. RQ2: Will the presence of automatic fake news detection results using machine learning algorithms increase participants' fake news detection relative to the control condition?

3. RQ3: What is the best way to communicate the result of machine learning based on fake news detection?

In our study, we proposed new machine-learning warnings in response to an emphasis on "algorithm transparency" [99–101]. A *Fact-Checking* warning that was used in the study of [21] was also used to see whether we could replicate their results. Using a between-subjects design, we conducted two online experiments on Amazon Mechanical Turk (MTurk), in each of which the immediate, short-term, and long-term effectiveness of three warnings against a control condition was evaluated in three phases, respectively. Across two experiments, 1,176 MTurk workers completed three interrelated decision

tasks of *recognition*, *detection*, and *sharing* to different news (half real and half fake) in each phase. In addition to the analysis of decision rates, we used a *signal-detection theory* (SDT) [102, 103] approach assessing individuals' susceptibility and bias at detecting fake news.

Across all phases of Experiment 1, participants showed limited recognition and cautious sharing decisions in general. Compared to the control condition, participants increased their correct detection of both fake and real news in the *Fact-Checking* condition but not the others. In Experiment 2, when the news source, a cue that most participants used to identify news' legitimacy, was removed from each news headline, similar results were obtained for the recognition and sharing tasks. But the effect of the *Fact-Checking* warning obtained in Experiment 1 disappeared. Instead, compared to the control condition, a *Machine-Learning-Graph* warning increased participants' sensitivity in differentiating fake and real news.

Our work makes the following three key contributions:

1. We proposed and evaluated the use of warnings to communicate the results of machine-learning-detected fake news to users. Across three machine-learning warnings, only the machine-learningrning-Graph warning that includes the detailed results of machine-learning-based detection increased individuals' correct detection of fake news, suggesting that a transparent machine learning algorithm is critical to improve people's fake news detection.

2. Our results showed that the *Fact-Checking* warning increased participants' correct detection of both fake and real news when the source was included in news headlines but not when the source was excluded. Participants showed more trust on the *Fact-Checking* warning even though the best detection performance was obtained with the *Machine-Learning-Graph* warning, suggesting promoting users' fake news detection does not necessarily promoting users' trust on the warning.

3. We introduced a SDT approach to investigate individuals' fake news detection and obtained that the *Machine-Learning-Graph* warning increased participants' sensitivity to differentiate fake from real news but not the *Fact-Checking* warning.

These contributions bridge the two venues of fake news mitigation and should help researchers and practitioners improve their understanding of people's decision-making in facing fake news and propose usable and transparent algorithms to address fake news problems.

3.2 Related Work

3.2.1 Human Fake News Detection

Within experimental settings, a few factors have been investigated to understand their impact on people's belief in and willingness to share fake news on social media. Pennycook et al. [79] conducted online studies examining the influence of warning and repetition. In their Experiments 2 and 3, participants were asked to evaluate different pieces of news in multiple stages. In stage 1, participants were asked to indicate whether they were to share news headlines (half fake and half real) on social media. Also, half of the participants were randomly assigned to a warning condition, in which all fake news stories were flagged with a caution symbol and the text "Disputed by 3rd Party Fact-Checkers". The rest half were assigned to a control condition in which no warning was presented. After a distracting stage, in stage 3, participants were asked to rate the familiarity and accuracy of real and fake news headlines (a half from stage 1 and a half from a new set of headlines). Each participant in Experiment 3 was also invited to return for a follow-up session one week later, in which the same headlines were seen in stage 3 and a new set of headlines were presented. Results showed that repeated headlines were rated as more "real" than novel headlines regardless of headlines' legitimacy and warning. The increased accuracy perception obtained with a single exposure lasted even after a week, regardless of the warning. Although the main effect of warning and its interaction with news legitimacy were significant in Experiment 2, neither term was significant in Experiment 3.

Clayton et al. [21] conducted an online study to investigate the effect of warning further. To eliminate confounding variables, they removed the source from all news headlines. In one condition, they implemented a "Fact-Checking" warning similar to that in [79] but specified the third parties' names within the warning. 413 participants in the condition indicated their perceived accuracy and likelihood to "Like" or share nine news (six fake, four of which with a warning, and three real). Compared to a control condition, participants' perceived accuracy of fake news with the warning was reduced, indicating the effectiveness of using a warning to reduce participants' belief in fake news.

A comparison between the studies of [79] and [21] revealed several critical differences, which may cause the ineffectiveness of the warning in Experiment 3 of [79], but the effect obtained by [21]. First, warnings were presented at the familiarity phase of [79] but the evaluation phase of [21]. Thus, Clayton et al. [21] evaluated the effect of warning, but

Pennycook et al. [79] evaluated its short-term effect. Second, the source was removed for each news headline used by [21], which may increase participants' reliance on using warnings to assess the legitimacy of news headlines. Also, the 3rd party names were specified in [21], which may increase individuals' trust in the warning. Accordingly, in our work, we investigated a warning like [21] during the assessment phase but varied the presence and absence of the source to understand how it impacts individuals' belief in fake news with warnings. Besides, we evaluated the immediate, short-term, and long-term effects of the warning in different phases and asked participants to indicate their trust level in the warning.

3.2.2 Computational Fake News Detection

In recent years, much attention has been made to detect fake news using computational means (e.g., [13–15]), especially using various features such as single-modal [63, 64] and multi-modal features [66, 67]. The single-modal methods mainly focus on analyzing the textual contents of news, for example, counting the number of assertive words which are shown more in trusted sources [63] or evaluating the consistency between topic sentence and main text [64]. Meanwhile, multi-modal methods include features derived from various sources, such as contents of news, users who posted news, publishers of news, or how news has propagated in a network. For instance, those features can be several textual features, including news contents and user's comments [65] or different data types, including a combination of text, image, or video [66–68].

In addition, to provide the accountability of algorithmic solutions, researchers have started offering details about the inner mechanisms of machine learning algorithms [104]. With more fact-checks done by machine learning algorithms [28], we study how to present the result *after* the detection of fake news occurs. Specifically, the machine-learning warnings in our study were not generated by machine-learning algorithms. Instead, we used hypothetical evaluation metrics (e.g., accuracy) and multi-modal features (e.g., text, picture) of machine learning algorithms within various warning signs (e.g., one with the wording "Machine Learning") to leverage the advancements in computational solutions.

3.2.3 Signal Detection Theory

Accuracy measure, such as the number of correct identification of fake news, is incomplete in understanding individuals' vulnerability to fake news because they ignore factors, such as the influence of real news. Accordingly, in our work, in addition to measures of

decision rates, we use SDT [105] to understand individuals' detection in response to fake news. SDT has been implemented for investigating decision-making in the context of perceptual uncertainties and risk [102], such as susceptibility to a phishing email and web pages [106, 107].

In SDT, participants' responses are defined as two normal distributions of pieces of evidence, representing both *signal* and *noise*. The difference between the means of signal and noise distributions reflects participants' sensitivity (d'), e.g., their ability to tell whether a piece of news is fake. Independent from d', SDT also allows a measure of participants' response criterion (c), e.g., their tendency to treat a piece of news as fake. In the context of fake news detection, the signal will be fake news to detect, and the noise will be real news. If the news is fake and the decision for the news judgment is suspicious, the trial is a H: hit. If there is a piece of real news but is judged suspicious, it is a FA: false alarm. If fake news is misjudged as non-suspicious, it is a miss. Finally, if real news is judged as non-suspicious, it is a correct decision. d' and c are derived as follows:

$$d' = z(H) - z(FA) \tag{3.1}$$

$$c = -0.5[z(H) + z(FA)] \tag{3.2}$$

Therefore, using SDT, the evaluation of how well a participant detects fake news will be not influenced by whether the participant is biased or not.

3.3 Method

The question of whether machine learning warning reduces individuals' fake news susceptibility has consequences for a broader perspective on the deployment of transparent machine learning algorithms. In this paper, we conducted two experiments investigating the three *RQs* by examining participants' recognition, detection, and willingness to share fake and real news[1].

We conducted a between-subjects online study investigating the effect of two machine-learning and one fact-checking warning in mitigating fake news. In addition to the three warning conditions, a control group (CON) in which no warning was presented was also included in the study. Participants made recognition, detection, and sharing decisions on

[1]The detailed data from all our experiments is available for download at `http://pike.psu.edu/download/websci19/`

fake and real news in three phases. In Phase 1, participants got warnings on fake news trials except for those participants in *CON*. After a distraction task of filling demographic information, Phase 2 started, in which participants did the same task as Phase 1 without warning to evaluate the short-term effect of the warning. One week later, we invited each participant back to Phase 3 to do the same task as Phase 2 to evaluate the long-term effect of the warning. Half of the trials in Phase 3 were news headlines that were already presented in Phases 1 and 2, which were used to investigate participants' decisions of repeated fake news.

The study was conducted on Amazon MTurk, and all participants were (1) at least 18 years old, (2) located in the United States, and (3) with a human intelligence task approval rate above 95%. Participants were allowed to participate in the study once. Our online study was programmed using Qualtrics. This and the following study were approved by the Institutional Research Board of The Pennsylvania State University.

Materials. We created 24 news headlines in the format of Facebook posts, consisting of a picture, source, header, and a short description (see Figure 3.1). 12 were verified fake news from *snope.com* and *politifact.com*, well-known third-party fact-checking websites. The other 12 news headlines were real news chosen from major news media, such as *huffpost.com* and *reuters.com*. The 24 pieces of news were divided into three groups (half real and half fake in each group). For each condition, a Latin-square design was implemented to balance the order of the groups across three phases.

In Experiment 1, we proposed three warnings: Fact-Checking (i.e., human fact-checkers) (*FC*), Machine-Learning (*ML*), and Machine-Learning-Accuracy (*MLA*). Each warning was attached to the bottom of the fake news in the study. Figure 3.1 gives a depiction of the warning design and the content of each warning. The two machine-learning warnings were the same, except a hypothetical value, 97%, was described in the *MLA* warning to indicate the accuracy of the machine-learning algorithm.

In Experiment 2, to increase the transparency of machine learning algorithms, we proposed a Machine-Learning-Graph (*MLG*) warning in which factors that a machine learning algorithm considers during the fact-checking are provided under "Disputed by a Machine Learning Algorithm" label. Because participants identified the news source as the most influential factor in their judgment of the news headlines' legitimacy, we also assessed the robustness of the effect of the *FC* warning from Experiment 1 by removing the source information. We also included *CON* and *ML* without sources to provide baselines for evaluation.

Figure 3.1: Warnings presented in Experiment 1, top row: A piece of fake news with Fact-Checking (FC) warning, center row: Machine-Learning (ML) warning, and bottom row: Machine-Learning-Accuracy (MLA) warning.

The selected news was released from April to June 2018, and the topic of news was limited to politics because 1) political news is one type of the most popular news that most individuals will read every day, so most people have a certain sense to judge its credibility without professional knowledge; 2) the negative effect caused by fake political news has become a critical issue in our daily life [98]. For example, in the 2016 American presidential election period, a piece of news titled "Pope Francis Shocks World, Endorses Donald Trump for President" [2] shook the world and commoved voters. Therefore, we believe political news should be treated as one of the top priority news types in solving fake news problems.

Procedure. Figure 3.2 illustrates the flow chart of Experiment 1. Participants were randomly assigned to one of the four conditions. After participants made an informed consent, Phase 1 started. Eight different pieces of news (half fake) were presented one at a time in a randomized order. Participants were instructed to view the headline first and then decide whether they have heard about the news (i.e., *Yes, Unsure, No*). Then, participants were asked to judge the accuracy and decide their willingness to share the news on a 5-point Likert scale, respectively (1 means "Very inaccurate" or "I would never share news like this one", 5 means "Very accurate" or "I would love to share news like this one").

After Phase 1, participants completed a demographic questionnaire that asked for age, gender, race, etc., as a distraction. Then Phase 2 started, in which participants completed the same three tasks with another set of eight news as Phase 1, except that the warning labels were *removed*. At the end of Phase 2, participants completed additional questions about their computer skill, social media experience, interest in politics, factors that impact their decisions on three tasks, and their trust in the warning on a 5-point Likert scale (1 means they did not trust the warning at all, 5 means they trust the warning a great deal). Phase 3 was conducted one week after Phases 1 and 2. Each participant received emails inviting him/her to evaluate a set of 16 pieces of news (half real and half fake) as in Phase 2. The given news included a new set of eight news, and four from Phase 1 and another four from Phase 2. Each participant was compensated for \$0.5 for the completion of Phases 1 and 2, and participants who finished Phase 3 received an extra \$0.5.

[2] https://www.snopes.com/fact-check/pope-francis-donald-trump-endorsement/

3.4 Results

3.4.1 Experiment 1

We recruited 800 MTurk workers on July 27, 2018. After removing nine incomplete submissions, 44 responses with both duplicate GPS coordinates (longitude and latitude provided by Quarltrics) and IP addresses, 178 responses with duplicate GPS coordinates but different IP addresses (rationales adopted from [108]), and 17 responses submitted within 3 minutes (median completion time is about 7 minutes), the numbers of participants that we accepted for the three warning conditions were 132, 136, and 138, respectively. The number of participants recruited in the *CON* condition was 146. In total, 552 participants (55.2% female) were included for data analyses. Participants' average age was 39, with 75% between 20 to 40 years. 55% of participants were college students or professionals who had a bachelor's or higher degree. The demographic distributions were similar among the four conditions.

For our analysis, selection rates of "Yes" for the recognition task were calculated for fake news and real news, respectively. For the detection task, choices of "Very inaccurate" and "Inaccurate" for fake news, and choices of "Accurate" and "Very accurate" for real news, were counted and coded as correct. The selection ratio of "Probably yes" and "I would love to share news like this one" of the sharing task were counted for fake and real news, respectively. For each task, we also measured participants' selection rates of the "Unsure" option.

For each phase, specified decision rates (range from 0 to 1) of each participant for each task were transferred into arcsine values and then entered into 2 (news' legitimacy: *fake*, *real*) $\times$ 2 (condition: *CON*, one warning label) mixed analysis of variances (ANOVAs), with a significance level of .05. At Phase 3, we included eight news from Phases 1 and 2, so repetition (*repeated*, *non-repeated*) was added as another within-subject factor for the tests.

Because the proportion of successful fake news *detection* ignores the influence from real news, we also used the SDT examining participants' sensitivity (d') and response bias (c) based on their correct detection of fake news (H) and incorrect detection of real news (FA). To accommodate H and FA rates of 0 or 1, a log-linear correction added 0.5 to the number of H, 0.5 to the number of FA, 1 to the number of signals (fake news), and 1 to the number of noise (real news) [106, 109]. Although the true d' values were underestimated by the log-linear correction [109], the relative differences across the

conditions should reflect differences apparent in the raw accuracy data. Measures of d' and c of detection decisions from Phases 1 and 2 were submitted to two-sample t-tests. At Phase 3, ANOVAs were conducted with repetition added as a within-subject factor.

Phase 1: Effect of warning. Table 1 lists the specified decision rates of each task for each condition in each phase, as well as the SDT measures for the detection task.

Recognition decisions. Across all phases, participants recognized more real news (34.6%) than fake ones (4.6%), $F_s > 99.29, p_s < .001, \eta^2_{ps} > .459$, and were more unsure about the recognition of real news (20.4%) than fake news (7.5%), $F_s > 32.64, p_s < .001, \eta^2_{ps} > .248$. No term involved *condition* was significant except the unsure recognition in *FC* (10.9%) was smaller than in *CON* (14.3%), $F_{(1,276)} = 4.31, p = .039, \eta^2_p = .015$. So, we focus on the analyses of detection and sharing decisions in the following parts but return to recognition decisions in the General Discussion.

Detection decisions. Analyses of correct detection decisions revealed that the main effects of news legitimacy were significant across all comparisons, $F_s > 160.56, p_s < .001, \eta^2_{ps} > .368$. Regardless of conditions, participants correctly detected more fake news (74.4%) than real news (40.7%). Relative to *CON* (56.2%), the overall correct detection rate was higher for *FC* (62%), $F_{(1,276)} = 5.99, p = .015, \eta^2_p = .021$, but not the other conditions (*ML*: 55%, *MLA*: 57.1%), $F_s < 1.0$. However, the two-way interaction of news legitimacy and the condition was not significant, $F < 1.0$. Thus, the *FC* warning not only increased participants' correct detection of fake news but also increased their correct detection of real news, suggesting that participants may rely on the presence and absence of the warning to judge the legitimacy of news headlines.

Across all comparisons, participants were more unsure in detecting real news (35.8%) than fake news (18.9%), $F_s > 56.92, p_s < .001, \eta^2_{ps} > .169$, which made sense since the warning label was presented with fake news only. Relative to *CON* (28.6%), only participants in *FC* (22.4%) showed less unsure about their detection, $F_{(1,276)} = 6.05, p = .015, \eta^2_p = .014$, but not the other conditions (*ML*: 28.1%, *MLA*: 29.9%), $F_s < 1.0$. Also, the reduced unsure detection rate (about 6%) of the *FC* warning was almost equal to the increased correct detection rate of the *FC* warning (about 6%), suggesting that participants relied on the *FC* warning to make decisions mainly when they were uncertain about the news' legitimacy. The main effect of the condition did not interact with news legitimacy, $F < 1.0$, indicating the effect of *FC* was similar between fake and real news.

SDT measures. When warning was present, participants showed minimal bias toward detecting news as fake across all conditions ($c = 0.02$). Compared to *CON* ($d' = 1.17$), participants' sensitivity to differentiate fake and real news were similar for all warnings (*FC*: $d' = 1.37$, $t_{(276)} = 1.80, p = .073$; *ML*: $d' = 1.12$, $t < 1.0$; and *MLA*: $d' = 1.34$, $t_{(282)} = 1.50, p = .135$).

Sharing decisions. Participants' overall willingness to share the news was low (see Table 1), but their willingness to share real news (14.9%) was higher than that of fake news (5.9%), $F_s > 41.78, p_s < .001, \eta^2_{ps} > .130$. Neither the main effect of condition (*CON* vs. *FC* vs. *ML* vs. *MLA*: 10.1% vs. 10.2% vs. 11.7% vs. 9.6%) nor its interaction with news legitimacy were significant, $F_s < 3.51$.

Participants were more unsure about sharing real news (13.5%) than fake news (6.9%), $F_s > 28.89, p_s < .001, \eta^2_{ps} > .095$. Compared to *CON* (9.5%), participants in *FC* (5.6%) were less unsure about their decisions, $F_{(1,276)} = 6.46, p = .012, \eta^2_p = .016$, but not participants in *ML* (12.5%) or *MLA* (12.9%) conditions, $F_s < 1.0$. Consistent with the results of unsure detection decisions, the *FC* warning also reduced participants' uncertainty during sharing decision-making.

Phase 2: Short-term effect of warning. Specified decision rates and SDT measures for Phase 2 are listed in Table 1.

Detection decision. As in Phase 1, the main effect of news legitimacy was significant, $F_s > 149.60, p_s < .001, \eta^2_{ps} > .352$. When the warning was absent in Phase 2, participants' correct detection of fake news (70.1%) was still better than that of real news (38.8%). For unsure option selection, the main effect of news legitimacy was also significant, $F_s > 67.55, p_s < .001, \eta^2_{ps} > .197$. Same as in Phase 1, participants showed less unsure of fake news (23.3%) than that of real news (38.6%). Regardless of the warning's presence or absence, more uncertainty at detecting real news than fake news probably was not due to the lack of decision aid for real news trials. No other terms were significant or approached significance.

SDT measures. When the warning was absent in Phase 2, across all conditions, participants showed similar sensitivity ($d' = 1.17$) and minimal bias toward detecting news as real ($c = 0.09$), see Table 1. Neither measure showed a difference across conditions, $t_s \leq 1.35$. Taking the results of the detection decision and SDT measures together, participants' reasonably accurate detection of fake news but not real news seems mainly

due to their uncertainty of real news.

Sharing decision. Without warnings, participants were more willing to share real news (13.7%) than fake news (6.9%), $F_s > 24.25, p_s < .001, \eta^2_{ps} > .079$, and were more unsure about sharing of real news (15%) than fake news (9.1%), $F_s > 12.56, p_s < .001, \eta^2_{ps} > .043$. No term involved condition was significant.

Phase 3: Long-term effect of warning. A total of 225 participants returned for Phase 3. Return rates (*CON*: 41.8%, *FC*: 42.4%, *ML*: 42.7%, *MLA*: 36.2%) and demographics were similar across conditions. Decision results and SDT measures for Phase 3 also are shown in Table 1.

Detection decisions. Correct detection of fake news (67.2%) was still better than that of real news (42.1%), $F_s > 50.46, p_s < .001, \eta^2_{ps} > .305$. And the main effect of news legitimacy interacted with repetition across all comparisons, $F_s > 20.84, p_s < .001, \eta^2_{ps} > .151$. Participants correctly detected more fake news, which was presented in Phase 3 only (73.2%) than those from Phases 1 and 2 (61.2%). However an opposite pattern was obtained for the real news: participants correctly detected less real news, which was presented in Phase 3 only (38.7%) than those from prior phases (45.4%).

For unsure option selection, both the main effect of news legitimacy and its interaction with repetition were significant across all comparisons, $F_s > 4.03, p_s < .047, \eta^2_{ps} > .033$. As in the prior two phases, participants were more unsure of detecting real news (34.5%) than fake news (23.5%). Besides, participants' uncertainty selection difference between repeated and non-repeated pieces of news was larger for fake news (repeated: 27.0%, non-repeated: 20.0%) than for real news (repeated: 33.4%, non-repeated: 35.7%).

SDT measures. Across conditions, there were no differences for both d' and c for the detection decisions, $F_s < 1.0$. But participants were biased to judge repeated pieces of news as real (c = 0.22) and non-repeated news as fake (c = -0.28), $F_s > 116.58, p_s < .001, \eta^2_{ps} > .517$. Also, participants tended to be less sensitive for repeated news (d' = 0.99) than non-repeated news (d' = 1.16), with the effect of repetition was significant for *FC* and *ML*, $F_s > 3.94, p_s > .049, \eta^2_{ps} > .033$, but not *MLA*, $F_{(1,109)} = 2.95, p = .088, \eta^2_p = .026$.

Sharing decisions. One week later, the willingness to share real news (16.7%) was still higher than that of fake news (9.7%), $F_s > 14.34, p_s < .001, \eta^2_{ps} = .109$. No other

effects were significant, except there was a main effect of repetition for the group of *FC*, $F_{(1,115)} = 4.14, p = .044, \eta_p^2 = .035$. Participants' willingness to share news was reduced for *FC* (11.4%) than for *CON* (15.8%).

For the unsure option, only the main effects of news legitimacy were significant, $F_s > 16.47, p_s < .001, \eta_{ps}^2 = .131$. Again, participants showed more unsure to share real news (15%) than fake news (8.1%).

Post-session questions. 72.6% participants did not have a major or work experience in computer-related fields, and 97.5% of participants did not show concern about using computers successfully in diverse situations. 73.2% of participants indicated that they used social media, such as Facebook and Twitter, daily or a few times a week. 82.6% of participants had an interest in politics.

When asked participants to confirm factors that impact their decisions on news' credibility and sharing on social media, Most participants selected source as the most influential factor for their detection (59.2%) and sharing (46.7%) decisions. Overall, participants did not demonstrate significant trust in the warnings; 31.8% expressed "a great deal" or "a lot" of trust, 30.8% reported a moderate level of trust, while 37.4% exhibited little or no trust. A chi-squared test showed that participants' trust on warning varied across conditions, $\chi_{(2)}^2 = 7.27$, $p = .026$, mainly due to more trust obtained for *FC* (40.2%) than *ML* (25%), $p_{adj.} = .023$.

Summary. In Experiment 1, we proposed two machine-learning warnings and evaluated their effects and one fact-checking warning to help individuals mitigate fake news. In Phase 1, relative to *CON* in which no warning was present, better detection results were obtained for the *FC* warning but not the *ML* and *MLA* warnings. The *FC* warning improved the correct detection of both fake and real news, suggesting that participants may use the presence and absence of warning as the criterion to make their detection decision, which is in agreement with the more trust obtained for the *FC* warning in post-session questions. When no warnings were displayed with fake news in Phases 2 and 3, the effect of *FC* disappeared. The *FC* warning did not show any short-term or long-term effect in helping participants detect fake news, probably because there were no details to inform participants about why the fake news was labeled. Although machine learning is a buzzword, participants showed less trust in the two machine-learning warnings than the fact-checking warning, suggesting that they may not necessarily understand what it is and, consequently, distrust its use for fake news

detection.

3.4.2 Experiment 2

We recruited extra 800 MTurk workers on October 16, 2018. The requirements to participate in this study were the same as in Experiment 1. Furthermore, any participants who had already participated in the previous study were excluded.

Materials and procedures of Experiment 2 were identical to Experiment 1 except as noted. First, we removed the source for all the 24 news headlines used in Experiment 1. Second, for the *MLG* condition, we added an extra bar chart below the warning label to represent factors that our hypothetical multi-modal machine learning algorithm considers (e.g., [65–68]). Three factors, "Source Reliability", "Content Trustfulness", and "Picture/Video Truthfulness", were listed from top to bottom. A filled bar graph accompanies each factor, and the length of each bar indicates values that the machine learning algorithm derived for the evaluation of the factor. The shorter the filled blue bar, the less reliable or accurate the news (see Figure 3.3).

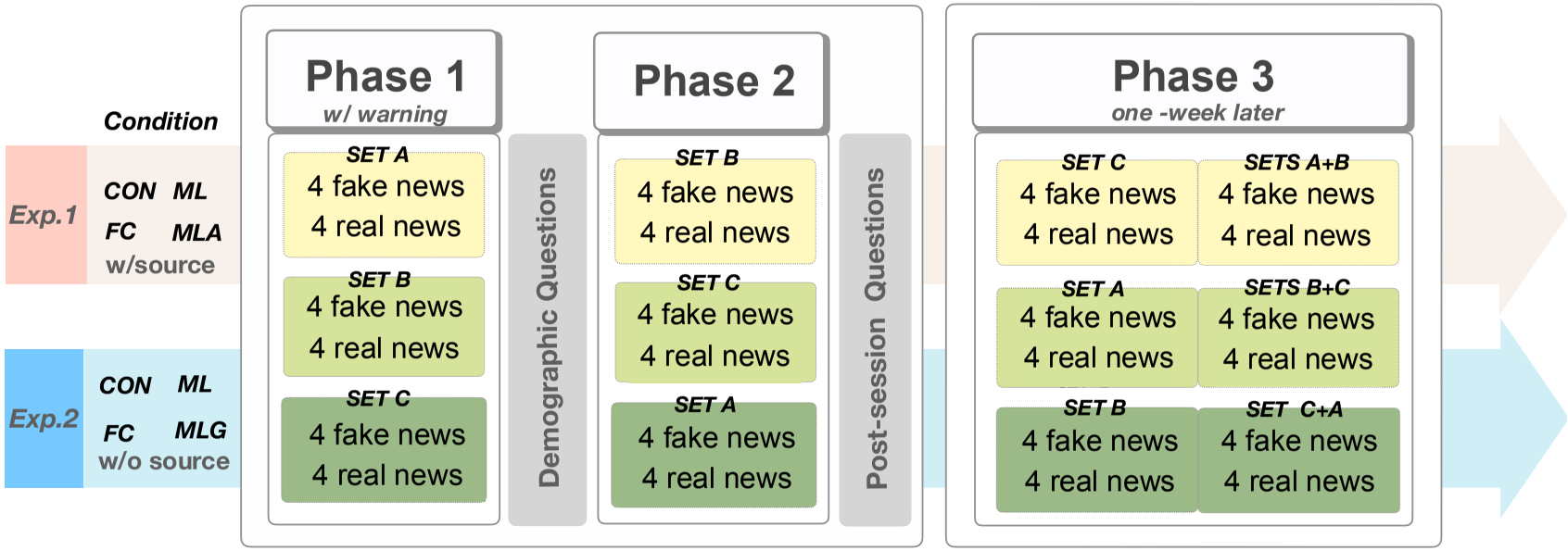

Figure 3.2: A flow chart showing the experimental design of each phase for both Experiments 1 and 2.

Table 3.1: Recognition, unsure recognition, correct detection, unsure detection, d', c, sharing, and unsure sharing results of fake and real news of each condition in each phase for Experiments 1 and 2. Sub. means subject, recog. means recognition.

Decision	Exp.1											Exp.2									
	Cond.	Sub. No.	Phase 1		Phase 2		Sub. No.	Phase 3 (New)		Phase 3 (Repeated)		Cond.	Sub. No.	Phase 1		Phase 2		Sub. No.	Phase 3 (New)		Phase 3 (Repeated)	
			Fake	Real	Fake	Real		Fake	Real	Fake	Real			Fake	Real	Fake	Real		Fake	Real	Fake	Real
Recog.	**CON**	146	4.3%	33.0%	3.9%	31.5%	61	4.9%	31.6%	11.1%	44.7%	**CON**	153	5.1%	35.8%	5.4%	34.6%	47	2.7%	23.4%	14.9%	39.4%
	FC	132	3.0%	30.3%	3.8%	28.4%	58	2.2%	29.0%	14.7%	45.1%	**FC**	160	7.5%	33.9%	6.3%	28.4%	60	4.6%	31.3%	16.7%	45.4%
	ML	136	5.9%	38.8%	5.7%	30.3%	58	9.1%	32.3%	17.7%	41.8%	**ML**	160	4.1%	31.3%	5.2%	29.4%	45	1.1%	25.6%	13.3%	47.2%
	MLA	138	4.0%	28.4%	3.8%	33.2%	50	2.0%	31.0%	9.0%	44.0%	**MLG**	151	3.6%	37.1%	3.8%	31.5%	54	2.3%	25.0%	13.4%	43.1%
Recog_Unsure	**CON**	146	8.4%	20.2%	8.7%	24.5%	61	6.6%	21.7%	16.0%	23.4%	**CON**	153	8.7%	19.6%	9.8%	23.5%	47	9.6%	23.4%	17.0%	27.7%
	FC	132	6.1%	15.7%	9.3%	18.9%	58	5.4%	23.2%	10.3%	19.6%	**FC**	160	6.1%	21.9%	10.6%	27.8%	60	9.6%	22.9%	15.4%	19.6%
	ML	136	6.4%	19.9%	12.9%	23.5%	58	8.2%	18.1%	18.1%	25.9%	**ML**	160	8.1%	23.6%	8.9%	25.6%	45	7.2%	20.6%	17.8%	21.7%
	MLA	138	9.1%	25.9%	9.8%	21.7%	50	7.0%	21.0%	15.5%	21.5%	**MLG**	151	6.8%	19.4%	10.6%	22.5%	54	11.1%	27.8%	19.0%	25.0%
Detection	**CON**	146	72.6%	39.7%	71.1%	39.7%	61	79.5%	38.9%	60.7%	43.9%	**CON**	153	70.1%	44.8%	69.0%	41.0%	47	72.3%	38.3%	61.2%	43.6%
	FC	132	79.2%	45.1%	72.0%	40.9%	58	73.7%	42.0%	62.9%	49.1%	**FC**	160	73.9%	42.5%	68.1%	39.4%	60	70.4%	42.9%	65.0%	51.7%
	ML	136	71.7%	38.4%	65.3%	35.7%	58	67.2%	36.6%	59.1%	42.2%	**ML**	160	73.8%	39.8%	65.0%	39.5%	45	72.8%	34.4%	63.9%	43.9%
	MLA	138	74.5%	39.7%	72.1%	38.9%	50	72.0%	37.0%	62.5%	47.0%	**MLG**	151	78.3%	43.0%	67.7%	37.6%	54	76.9%	35.2%	67.1%	43.1%
Detection_Unsure	**CON**	146	21.6%	35.6%	22.4%	38.7%	61	15.2%	34.8%	27.9%	34.0%	**CON**	153	21.1%	34.2%	23.4%	40.2%	47	16.5%	39.9%	28.7%	33.0%
	FC	132	13.8%	31.1%	20.3%	34.5%	58	20.5%	36.6%	25.9%	31.3%	**FC**	160	18.4%	37.2%	22.8%	40.2%	60	21.3%	40.8%	20.4%	29.6%
	ML	136	20.4%	35.8%	27.6%	42.6%	58	23.3%	35.3%	27.2%	35.3%	**ML**	160	19.4%	37.7%	25.5%	43.3%	45	23.9%	42.8%	28.3%	35.0%
	MLA	138	19.7%	40.4%	22.8%	38.6%	50	21.5%	36.0%	27.0%	33.0%	**MLG**	151	17.7%	35.8%	27.2%	41.1%	54	18.1%	39.8%	26.4%	30.6%
d'	**CON**	146	1.17		1.22		61	1.30		0.94		**CON**	153	1.20		1.22		47	1.20		0.92	
	FC	132	1.37		1.14		58	1.28		1.06		**FC**	160	1.33		1.19		60	1.34		1.13	
	ML	136	1.12		1.07		58	0.96		0.89		**ML**	160	1.26		1.17		45	1.21		1.03	
	MLA	138	1.34		1.23		50	1.11		1.05		**MLG**	151	1.41		1.14		54	1.26		0.97	
c	**CON**	146	0.04		0.09		61	-0.07		0.22		**CON**	153	0.11		0.15		47	0.08		0.19	
	FC	132	-0.04		0.04		58	0.06		0.21		**FC**	160	0.06		0.15		60	0.17		0.21	
	ML	136	0.03		0.16		58	0.06		0.23		**ML**	160	0.04		0.23		45	0.06		0.18	
	MLA	138	0.07		0.07		50	0.01		0.21		**MLG**	151	-0.004		0.14		54	-0.03		0.08	
Sharing	**CON**	146	6.5%	13.7%	6.9%	13.2%	61	9.8%	19.7%	13.1%	20.5%	**CON**	153	7.8%	15.0%	6.5%	15.0%	47	7.4%	11.2%	7.4%	15.4%
	FC	132	4.5%	15.9%	5.3%	13.6%	58	7.6%	12.1%	9.8%	16.10	**FC**	160	9.7%	15.2%	10.5%	20.2%	60	11.7%	13.3%	10.8%	15.8%
	ML	136	7.9%	15.4%	7.7%	14.9%	58	11.2%	18.5%	11.2%	13.40	**ML**	160	5.0%	13.6%	7.0%	12.0%	45	5.6%	8.9%	4.4%	9.4%
	MLA	138	4.7%	14.5%	6.9%	13.2%	50	8.5%	17.0%	5.5%	16.00	**MLG**	151	5.5%	15.4%	4.5%	12.6%	54	6.5%	12.5%	7.9%	15.3%
Sharing_Unsure	**CON**	146	6.2%	12.8%	9.1%	14.0%	61	7.4%	16.0%	8.6%	13.9%	**CON**	153	7.4%	13.1%	8.3%	16.7%	47	5.9%	10.1%	12.2%	11.2%
	FC	132	3.2%	8.0%	6.4%	10.0%	58	4.5%	12.1%	4.9%	15.2%	**FC**	160	5.6%	14.5%	9.7%	12.2%	60	7.9%	15.4%	6.3%	13.3%
	ML	136	8.8%	16.2%	11.6%	20.2%	58	9.1%	15.1%	10.8%	19.0%	**ML**	160	7.0%	10.9%	4.8%	12.5%	45	8.3%	9.4%	6.7%	11.1%
	MLA	138	9.2%	16.7%	9.4%	15.6%	50	6.5%	13.5%	13.5%	15.5%	**MLG**	151	6.8%	12.7%	11.3%	14.1%	54	9.7%	15.3%	7.4%	8.8%

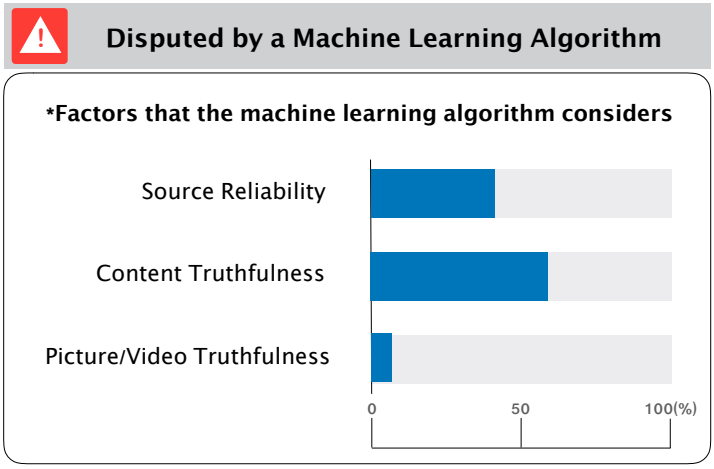

Figure 3.3: Machine-Learning-Graph (MLG) warning of Experiment 2. In Experiment 2, to increase the transparency of machine learning algorithms, we proposed a Machine-Learning-Graph (MLG) warning in which factors that a machine learning algorithm considers during the fact-checking are provided under the "Disputed by a Machine Learning Algorithm" label.

Using the same criterion as Experiment 1, we got a total of 624 (54.9% female) valid responses, with 153, 160, 160, and 151, for *CON*, *FC*, *ML*, and *MLG*, accordingly. Participants' average age was 39.5 years. 54.2% of participants hold a bachelor's or higher degree. For each task in each phase, specified decision rates and SDT measures as a function of the condition were calculated for each participant. Analyses of the decision rates and SDT measures were conducted in the same way as Experiment 1.

Phase 1: Effect of warning. Table 1 lists the specified decision rates and SDT measures. Same as Experiment 1, participants recognized more real news (34.5%) than fake news (5.1%) regardless of conditions or phases, $F_s > 103.94, ps < .001, \eta^2_{ps} > .512$, and were more unsure about recognizing real news (21.1%) than fake news (7.4%), $F_s > 19.83, ps < .001, \eta^2_{ps} > .181$. Again, we focus on the analyses of detecting and sharing decisions in the following parts but return to recognition decisions in the General Discussion.

Detection decisions. The main effect of news legitimacy was significant across all comparisons, $F_s > 130.56, ps < .001, \eta^2_{ps} > .296$. Participants correctly detected more fake news (74.0%) than real news (42.5%). Moreover, for *MLG*, compared to *CON*, there was a two-way interaction of news legitimacy and condition, $F_{(1,302)} = 5.48, p = .020, \eta^2_p = .018$. Those participants made more correct decisions on fake news with the *MLG* warning (78.3%) than without warning (70.1%), but their correct decision on real news was similar between the two conditions (*CON*: 44.8%, *MLG*: 43.0%), suggesting the effectiveness of *MLG* in reducing participants' fake news susceptibility. Relative to *CON*, neither the main effect of the condition nor its interaction with the condition was significant for the correct detection with *FC* or *ML* warnings, $F_s < 3.05$.

For unsure detection, compared to the *CON*, only the main effect of news legitimacy was significant across all comparisons, $F_s > 74.86, ps < .001, \eta^2_{ps} > .194$, Participants were more uncertain about the accuracy of real news (36.2%) than fake news (19.2%).

SDT measures. When warning was present, compared to *CON* ($d' = 1.20$), participants' sensitivity to differentiate fake and real news was better for *MLG* ($d' = 1.41$), $t_{(302)} = 1.98, p = .048$, but not other conditions (*FC*: $d' = 1.33$, $t_{(311)} = 1.17, p = .242$, and *ML*: $d' = 1.26$, $t < 1.0$). Relative to *CON* ($c = 0.11$), participants showed similar bias for each warning [*MLG* (c = -0.004): $t_{(302)} = -1.84, p = .067$; *FC* ($c = 0.06$): $t < 1.0$; and and *ML* ($c = 0.04$): $t_{(311)} = -1.05, p = .294$].

Sharing decisions. Compared to *CON*, only the main effect of news legitimacy was significant for both sharing and unsure decisions for all warnings, $F_s > 15.31, ps < .001, \eta^2_{ps} > .047$. In general, participants were more willing to share real news (14.8%) than fake news (7.0%), and they also showed more uncertainty in sharing real news (12.8%) than fake news (6.7%).

Phase 2: Short-term effect of warning. Decision results and SDT measures of each task are shown in Table 1.

Detection decisions. When the warning was absent after a short distraction task, the main effect of news legitimacy was still significant across all comparisons, $F_s > 119.49, ps < .001, \eta^2_{ps} > .278$. Participants correct detection of fake news (67.4%) was better than that of real news (39.4%). However, the effect of *MLG* obtained in Phase 1 disappeared, $F < 1.0$. For unsure option selection, participants still showed more unsure for real news (41.2%) than fake news (24.7%) across all comparisons, $F_s > 68.11, ps < .001, \eta^2_{ps} > .184$.

SDT measures. When the warning was absent, neither measure showed difference across conditions, $t_s \leq -1.29, ps \geq .197$.

Sharing decisions. Sharing decisions also showed the same pattern as prior results (see Table 1). Participants were more willing to share real news (15.0%) than fake news (7.2%), $F_s > 36.69, ps < .001, \eta^2_{ps} > .106$. For unsure option selection, participants also showed more uncertainty about sharing real news (13.8%) than fake news (8.5%), $F_s > 18.69, ps < .001, \eta^2_{ps} > .058$. Moreover, the effect of warning was revealed in all comparisons. Relative to *CON*, participants who saw the *MLG* warning in Phase 1, increased their uncertainty about sharing fake news but reduced their uncertainty about sharing real news, $F_{(1,302)} = 4.14, p = .043, \eta^2_p = .014$. Participants who saw the *FC* warning in Phase 1 showed a similar pattern as participants in *MLG*, $F_{(1,311)} = 3.87, p = .050, \eta^2_p = .012$. But their increased susceptibility of fake news was numerically smaller than that of *MLG*, and reduced susceptibility of real news was numerically larger than that of *MLG*. And for participants in *ML*, they reduced their uncertainty of sharing both fake and real news, $F_{(1,311)} = 4.61, p = .033, \eta^2_p = .015$.

Phase 3: Long-term effect of warning. After one week, 206 participants returned for Phase 3. Return rates (*CON*: 30.7% , *FC*: 37.5%, *ML*: 28.1%, *MLG*: 35.8%) and

demographics were similar across conditions. Decision results were also shown in Table 1.

Detection decisions. Same as Experiment 1, participants still correctly detected more fake news (68.8%) than real news (41.9%) one week later, $F_s > 37.17, ps < .001, \eta^2_{ps} > .261$, and their correct detection pattern varying as a function of repetition, $F_s > 9.76, ps < .002, \eta^2_{ps} > .098$. Across all conditions, participants' correct detection of repeated fake news (64.4%) was smaller than their correct detection of non-repeated fake news (73.1%). However, participants correctly detected more repeated real news (45.9%) than non-repeated real news (38.0%). Although participants in the *MLG* condition showed numerically better results in detecting fake news, the long-term effects of *MLG* were not significant, $F_s < 1.0$.

Participants were more unsure about the selection of real news (36.4%) than fake news (23.0%), $F_s > 18.80, ps < .001, \eta^2_{ps} > .173$. Although the main effect of repetition was not significant, it interacted with the news legitimacy, $F_s > 10.41, ps < .002, \eta^2_{ps} > .104$. Participants were more unsure about detecting fake news from real news that was repeated than those that were non-repeated.

SDT measures. Same as Experiment 1, there were no differences for both d' and c for the detection decisions across conditions, $F_s < 1.0$. Nevertheless, participants showed less sensitivity for the repeated news headlines ($d' = 1.01$) than for the non-repeated news headlines ($d' = 1.26$), $F_s > 4.61, ps < .035, \eta^2_{ps} > .045$. They also tended to be biased to judge repeated news as real ($c = 0.16$) than non-repeated news ($c = 0.07$), with the effect of repetition was significant for *ML* and *MLG*, $F_s > 5.55, ps < .021, \eta^2_{ps} > .058$, but not *FC*, $F_{(1,105)} = 3.51, p = .064, \eta^2_p = .032$.

Sharing decisions. Same as prior phases, participants showed more willingness to share real news (12.6%) than fake news (8.0%), $F_s > 8.42, ps < .005, \eta^2_{ps} > .074$. Participants only showed more unsure about sharing real news than fake news for the comparison between *CON* and *FC*, $F_{(1,105)} = 9.95, p = .002, \eta^2_p = .087$. The two-way interaction of repetition and condition was significant for the comparison between *CON* and *MLG*, $F_{(1,99)} = 8.62, p = .004, \eta^2_p = .080$. Compared to *CON*, participants in *MLG* condition showed more uncertainty at sharing news that was non-repeated but less uncertainty at sharing news that was repeated.

Post-session questions. Overall results of post-session questions in Experiment 2 were similar to those from Experiment 1. 72.4% of participants did not have a major or

work experience in computer-related fields, and 98.2% of participants were not concerned about using computers successfully in diverse situations. 74.2% of participants indicated that they used social media, such as Facebook and Twitter, more than a few times a week, and 84.8% of participants had an interest in politics.

When asked how much their trust in the warning when evaluating the accuracy of news during the study, participants did not show much trust in warnings in general, with 16.8% giving "a great deal" or "a lot" trust, 28.2% indicating their trust was moderate, and 55% showed a little or no trust. Participants' trust level also varied across warnings, $\chi^2_{(2)} = 34.40$, $p < .001$. Participants showed more trust for *FC* (30.6%) than *ML* (7.5%), $p_{adj.} < .001$, and *MLG* (11.9%), $p_{adj.} < .001$, respectively.

Summary After removing the source within each news headline, in Experiment 2, we did not obtain the effect of the *FC* warning as in Experiment 1. Compared to *CON* in which no warning was presented, the *MLG* warning improved participants' detection of fake news and increased their sensitivity to differentiate fake and real news, while the *ML* warning did not. When warnings were absent in Phases 2 and 3, neither the main effect of warning nor its interaction with other factors were significant for detection decisions. However, the effects of *MLG* and *FC* were revealed in participants' increased uncertainty of sharing fake news but reduced uncertainty in sharing real news in Phase 2, suggesting a short-term effect for both warnings. Although participants showed better fake news detection with *MLG* in Phase 1, their trust on the *MLG* warning was less than that of the *FC* warning, suggesting that participants' better detection of fake news with *MLG* in Phase 1 was mostly due to their reliance on the factors that presented within the warning.

3.5 General Discussion

Across two experiments, we proposed three machine-learning warnings and evaluated their effects and a fact-checking warning in helping individuals mitigate fake news. Both decision rates and SDT measures showed the effect of *MLG* warning in helping participants differentiate fake news from real ones. When no warnings were displayed in Phase 2, although the *MLG* warning did not impact individuals' detection decisions, participants increased their uncertainty in sharing fake news but reduced their uncertainty in sharing real news, suggesting a short-term effect of the warning.

We obtained that the effect of *FC* warning increased participants' correct detection

of both fake and real news when the source was included in news headlines but not when sources were excluded. Although the *FC* warning did not impact individuals' detection decisions when the source was excluded, they increased participants' uncertainty in sharing fake news and reduced their uncertainty in sharing real news when the warning was not displayed in Phase 2, suggesting a short-term effect. Thus, our results did not replicate [21], but are somewhat consistent with [79], showing a small effect of the warning. With the *FC* warning, participants increased not only the correct detection of fake news to which the warning was attached but also the correct detection of real news, suggesting that participants probably relied on the presence and absence of the warning to make the detection decision.

3.5.1 Limited Effect of Warning Labels

All the warnings that have been implemented in current and prior studies (e.g., [79]) revealed a small effect on mitigating fake news. One possible reason is that all those proposed warnings are passive, which indicates misinformation to participants without interrupting their primary task. i.e., viewing news headlines and obtaining new information. Prior studies on cybersecurity, e.g., phishing [110, 111], showed that participants ignored passive security indicators and relied instead mainly on the website contents to decide the trustworthiness of a web page. The results of current Experiment 1 showed a similar pattern, in that participants mainly relied on the source of news to make the news legitimacy decisions even when the warning labels were present. Therefore, one way to improve the effectiveness of the warning is to make it active, which will capture users' attention and force users to choose one of the options that were presented by the warning [112–114].

However, a zero-day exploit of fake news will leave no opportunity for automatic detection and prevention, and people need to make a decision on their own [115]. Therefore, the ability to tell fake news from real news is an important skill for individuals to acquire. Training is one promising approach to address individuals' inability to differentiate fake and real news. Also, prior studies in cybersecurity provided evidence that knowledge gained from training enhanced the effectiveness of phishing warnings [116]. Therefore, another way to improve the effect of warning is to embed training within the warning and use each warning as an opportunity to train users on how to mitigate fake news.

3.5.2 Better Recognition of Real News

A point to note about the present study is that overall, participants recognized more real news than fake news, and also showed more uncertainty in recognizing real news than fake news. "Recognition" and "unsure" decisions represent two distinct processes for recognition memory, *recollection*, and *familiarity* [117]. The distinction is that people could recognize a piece of news as familiar but not be able to recollect where he or she previously saw it.

Across three phases, participants appeared reasonably accurate at detecting fake news, but their correct detection of real news was less than chance. SDT measures did not show that participants were biased in judging news as fake, thus the poor detection of real news was mainly due to participants' more uncertainty about detecting real news than fake news. A further Pearson correlation analysis revealed that the unsure recognition of real news had a statistically significant positive linear relationship with the unsure detection of real news for both experiments, $p_s < .001$. The strength of the association was approximately moderate for Experiment 1, $r = .294$, and there was a small correlation $r = .245$ for Experiment 2.

Consistent with [118], our results showed that participants' willingness to share news was low in general and was lower for fake news than real news. Moreover, our study revealed that participants were more uncertain about sharing real news than fake news. For both experiments, Pearson correlation analysis showed a significant positive association between unsure recognition and the unsure sharing, $p_s < .003$, but with a small correlation, $r = .268$ for Experiment 1, and $r = .119$ for Experiment 2.

Altogether, the better recognition and more uncertainty of real news suggest that participants may have been exposed to those pieces of real news previously, and their familiarity (uncertainty) with real news seems to have impaired their evaluation of news' accuracy and their sharing decisions.

3.5.3 Effect of Repetition

At Phase 3, for those pieces of news that were repeated, participants showed better recognition. Moreover, the increase of recognition was more evident for real news than fake news, suggesting the repetition increased recollection of real news than fake news. Consistent with the effect of repetition obtained by [79], SDT measures further revealed that participants were less sensitive and more biased to judge news headlines as real for news from prior phases than the news that was presented in Phase 3 only. Participants'

unsure detection was also increased for the repeated pieces of news; however, the increase was more evident for fake news than real news, suggesting that the repetition mainly increased participants' familiarity (uncertainty) with fake news. Therefore, our study provided evidence that the repetition probably impacts the detection of fake and real news differently.

Human memory has been described as an optimization of information retrieval, which uses the statistics derived from past experience to estimate which knowledge will be currently relevant [119]. Besides allowing individuals to remember objects and events that they have actual experience, human memory systems are subject to distortion, bias, and the creation of illusions [74, 120]. Combining the overall better recognition of real news, increased recollection of repeated real news and increased familiarity with repeated fake news, our study further indicates the important role that memory plays in individuals' belief in fake news. Further research should be conducted to explore the extent to which memory affects individuals' belief in fake news.

3.5.4 Limitations

In our experiments, the effectiveness of warning labels was evaluated with a convenience sample of Amazon MTurk workers, who tended to be young, more educated, and more tech-savvy than the general public. Thus, the generalization of current findings to participants with different demographics needs to be further examined. In addition, the experiment design is limited in its ecological validity. We considered a more ecologically valid method, such as providing a social media interface during the study, but we decided to present news headlines to exclude extraneous variables that may have an effect on the outcomes, which increased our confidence in the internal validity of the obtained results. Note that such a design was the same as the prior studies [21, 79], which made our results comparable to the prior ones.

Another possible confound was that participants may have experienced the fact-checking warning previously but not the machine learning warnings. Better performance was only obtained for the *MLG* warning but not *ML* and *MLA* warnings, indicating that novelty may not be critical. Finally, all news headlines in our study are politically related, so generalizing the findings to other types of misinformation needs to be further investigated. Finally, in this study, we did not consider participants' political stance as a factor due to our main interest in warning labels, and the prior study showed that the partisan bias did not significantly affect participants' susceptibility to fake news [20]. However, recently Gao et al. [121] obtained results indicating that the warning labels are

more effective for participants in the liberal group than participants in the conservative group. Therefore, to understand whether a pre-existing political stance interacts with a machine-learning warning, future studies can consider the political stance of participants as an extra factor and measure how it impacts participants' belief in fake news.

3.6 Conclusion

In this work, we conducted two online experiments to understand the impact of machine-learning warnings on reducing individuals' fake news susceptibility. Each experiment consisted of three phases examining participants' recognition, detection, and sharing of fake news, respectively. Across three machine-learning warnings, the Machine-Learning-Graph warning increased participants' sensitivity in differentiating fake from real news, but participants showed limited trust in it. Our study results imply that a transparent machine learning algorithm (that explains the detailed results) may be critical to improving individuals' fake news detection but not necessarily to increasing their trust.

Chapter 4
Platform-Driven: Effects of AI Explanations on Misinformation Detection with a Warning

In this chapter, I delve into a second study on platform-driven corrections, building upon the research introduced in the previous chapter, which examined the effectiveness of machine-learning graph warnings. In the current research, I explore machine warnings more in-depth with explanations. The expression style in the explanation part of the warning is inspired by the framing effect and categorized into positively framed explanations and negatively framed explanations. Through a total of three experiments, I aim to elucidate which form of machine warning with explanation is more effective.

4.1 Introduction

In the context of the COVID-19 pandemic, the overflow of misinformation calls for urgent measures to reduce such misinformation [25]. Many cases have presented how detrimental health-related misinformation is as much as people can sometimes die from the wrong treatment for COVID-19.[1] To mitigate the rapid spread of misinformation on social media [30], companies such as Meta and Twitter have created warning systems to debunk fake news.[2] Previous works [79, 122] have shown that a debunking warning label plays an effective role in mitigating fake news.

Meanwhile, active efforts have been devoted to effectively detecting fake news [28,

[1] https://www.bbc.com/news/world-53755067

[2] https://www.facebook.com/journalismproject/programs/third-party-fact-checking/new-ratings, https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

123, 124]. Beyond improving detection models, recent research interest has expanded to explainable artificial intelligence (XAI) to provide an explanation of how an AI system detects fake news to news consumers [125, 126].

The value of the XAI for misinformation mitigation lies in helping users not accept or disseminate it. Empirical studies have been conducted to examine the effectiveness of AI explanations in influencing humans' misinformation detection [80, 127–129]. For example, Lu et al. [129] showed that presenting AI-based credibility indicators is effective in nudging participants into aligning their misinformation detection with the AI model's prediction. Seo et al. found that ML warning with explanation increased participants' ability to detect fake news more than the warning only when a news source was not provided.

Despite the promising empirical findings, most of the existing research examines the effectiveness of explanations through options to interact with the AI system, different human behavior (i.e., misinformation detection or sharing), or other factors (e.g., social influence). However, humans' perception and acceptance of an explanation are often shaped by how the problem is framed [24]. Moreover, prior work [80] presented that although AI explanations help people detect misinformation better, participants' trust in the AI system decreased. To fill the gap, in this work, we investigate whether explaining how an AI system debunks misinformation can improve the effectiveness of misinformation warnings. We focus on COVID-19 fake news, considering its timely importance. Specifically, we examine the following research questions(RQs).

- **RQ 1.** Will a misinformation warning with explanations improve humans' ability to detect fake news compared to the warning only? If so, will a positive framing work better than a negative framing for the explanations?
- **RQ 2.** Will the effect of misinformation warnings with explanations depend on the AI system's reliability?
- **RQ 3.** Will the misinformation warning with explanations increase humans' trust in the AI system?

We conducted three online experiments by recruiting Amazon Mechanical Turk workers ($N = 2,692$). In Experiment 1, we investigated the effect of explaining how an AI system debunks fake news on humans' detection of misinformation with a warning (**RQ1**). We proposed credibility explanations in both positive framing (i.e., POS) and negative framing (i.e., NEG) and examined the framing effect in Experiment 2 (**RQ1**).

In Experiment 3, we explored the impact of the AI system's reliability (i.e., whether the AI systems will make a lot of mistakes) (**RQ2**). We also evaluated humans' trust in AI systems (**RQ3**).

The results of our experiments suggest that the credibility explanations under negative framing (i.e., NEG) decrease humans' perceived accuracy ratings of fake news. Such results underline the necessity to consider the framing effect in examining the effectiveness of AI explanations in human misinformation detection. However, we find that humans do not always depend upon the warning or the warning with explanations for misinformation detection. They tend to think about miss errors (i.e., false negatives) of the AI systems. Moreover, the system's reliability is critical to address such suspicion. Those results highlight the importance of informing users of the AI system's reliability and possible error types. Finally, although humans' misinformation detection can be influenced by explanations, they show more trust in the warning itself. Moreover, such trust increases when the AI system's reliability becomes higher. Our main contributions are summarized as follows.

- We empirically examined the framing effect on misinformation warning with explanations upon humans' misinformation detection.
- We presented evidence showing the effectiveness of high system reliability in humans' misinformation detection and their trust in AI systems.
- We provided implications for researchers and practitioners in designing AI explanations to mitigate misinformation on social media platforms.

4.2 Related Work

4.2.1 Explainable Artificial Intelligence (XAI) and Misinformation Correction

Much recent work has been conducted to examine factors and designs enabling users to accept AI systems [130–133]. In line with those studies, XAI aims to make AI results more interpretable with explanations [22, 23, 85]. Many misinformation detection systems have been developed [62], but there is little research on how to display the detection results in a way that users can understand. Several related studies explored the effect of explanation in terms of AI evaluation on fake news [126] or investigated the effect of

warning with a certain type of explanation [128, 134]. However, those studies either did not explain concretely how an AI system derived each prediction (i.e., local) or how the AI system behaves in general (i.e., global). To fill the gap, our study explored the effect of a warning with an explanation showing factors that an AI system evaluates for fake news detection. Based on Seo et al. [80], which confirmed the effect of a transparent machine learning warning, we designed our warning with an explanation to extend the research scope along with the framing effect theory and the impact of an AI system's reliability.

4.2.2 Credibility for Misinformation Correction

Even though it is hard to find universal agreement on the concept of credibility across different fields [135], credibility has been used as a major criterion to measure the quality of information on the web [136] as well as traditional mass media [137]. Credibility can be understood in terms of believability, trust, reliability, accuracy, fairness, and objectivity [135]; therefore, it is naturally emphasized in detecting misinformation. A number of studies have investigated the credibility of information on social media [135, 138–141]. Among them, Savolainen et al. suggested a conceptual framework of credibility by dividing two approaches including the credibility of the author and the credibility of mis/disinformation content. Meanwhile, Molina et al. [37] organized features of real news and fake news, which are used for evaluation of news veracity. Based on these studies, we created our credibility factors for an explanation.

4.2.3 Framing Effects

Explaining the basis of the AI system's judgment is ultimately to help users understand the explanation effectively to increase the user's trust in the system [142]. In order for users to be receptive to explanations, an effective explanation needs to be provided [23]. Charts are often used to explain the determinants of an AI system effectively [143, 144], which leads to the question of how to visualize charts more effectively. One approach could be to consider different framing options. Tversky and Kahneman addressed the framing effect first by explaining that people's willingness to take risks can depend on how options are presented [24]. In the privacy domain, Choe et al. [145] investigated the effect of the privacy rating in the context of the framing effects through visual representations of an app to warn the level of the app's privacy protection. The study confirmed the effect of a positively framed rating icon. On the contrary, in the health

communication domain, Rosenblatt et al. [146] found that negatively framed warnings are more effective than positively framed warnings. Greene et al. [147]'s study presented no effects of providing a general warning about the dangers of online misinformation regardless of framing. Despite the different targets and framing designs, these studies throw insights that can be applied to our study. From the framing effect point of view, we propose to compare the negative framing and positive framing of the chart explanation to supplement the warning message against fake news.

4.2.4 Importance of Reliability

Previous studies showed that the effect of warning did not guarantee trust in the warning system. One of the factors could be the reliability of the system. Reliability is regarded as one factor of trust [148]. Several researchers [148, 149] demonstrated the impact of reliability information in building users' trust. Dzindolet et al. [148] confirmed that trust in automation can increase with information about why a decision of an automated system might err. In Chancey et al. [149]'s study, the high-reliability system got more trust than the low-reliability system. Furthermore, Kocielnik et al. [150] explored the impacts of different types of errors an AI system makes. These studies show how users react differently depending on how trustworthy an AI system is. In this context, we investigated the impact of reliability information on a warning system by comparing the impact of the high-reliability system and that of the low-reliability system. To the best of our knowledge, our study is the first study to cover the reliability information of the system in terms of explanation-based-warnings.

4.2.5 Trust in the AI System

The concept of trust is defined in various fields in different ways [149, 151, 152]. For example, Mayer et al. defined trust as a willingness to accept vulnerability. Trust in information systems indicates self-assurance by assessment of risks and alternatives [142]. Furthermore, trust is the one that can have an impact on reliance on automation [153]. Machine learning researchers also pay attention to the importance of trust, which is linked to justification issues of models [101, 154, 155]. However, building trust in algorithmic systems can be a challenging task [86].

An explanation increases trust by contributing to the transparency of the AI systems [154]. Seo et al. [80] conducted user studies testing the effects of warnings with and without explanation. The study stated machine learning graph warning increased participants'

Items	Options	EXP.1 (N=710)	EXP.2 (N=1014)	EXP.3 (N=968)
Gender	Male	51.8%	42.2%	38.7%
	Female	47.6%	57.4%	60.1%
	Prefer not to answer	0.6%	0.4%	1.1%
Age	18-29	21.3%	24.3%	20.0%
	30-39	34.2%	33.0%	33.6%
	40-49	25.1%	24.5%	24.8%
	50-59	14.4%	12.9%	13.5%
	60-69	4.8%	4.7%	6.6%
	70-79	0.3%	0.5%	1.4%
Race	Caucasian	78.9%	79.6%	75.2%
	African American	10.7%	9.0%	11.1%
	Hispanic	3.7%	4.8%	4.8%
	Asian	3.5%	4.0%	5.7%
	Other	2.9%	1.9%	1.5%
	Prefer not to answer	0.3%	0.7%	0.7%
Education	High school	4.9%	7.4%	8.9%
	Some college credit	12.3%	16.4%	27.2%
	Bachelor	58.7%	53.0%	38.5%
	Master	20.8%	19.2%	17.1%
	Doctor	1.8%	1.7%	3.7%
	Other	1.1%	2.1%	3.9%
	Prefer not to answer	0.3%	0.3%	0.6%
AI/ML experience	Not at all	16.1%	27.0%	46.3%
	Novice	19.2%	25.8%	37.1%
	Intermediate	28.0%	21.3%	12.6%
	Competent	27.3%	18.0%	3.8%
	Expert	9.4%	7.8%	0.1%

Table 4.1: Demographic information of the participants in the three experiments.

sensitivity in distinguishing fake news from real news but did not increase trust in it. Epstein et al. [134] replicated Seo et al.'s finding: trust did not increase with a warning with explanation, although explanations increased the effect of a warning. XAI is developed to increase the users' trust in the system, but warning studies have not demonstrated it so far. Therefore, our study tried to measure trust in the AI system after evaluating perceived accuracy ratings with warnings to see if our warning with explanations could have an impact on their trust.

4.3 Method

We conducted three online experiments to investigate the effect of explaining how an AI system debunks fake news can improve the effectiveness of fake news warnings in the context of social media platforms. In each experiment, participants evaluated twenty-four pieces of news by answering their perceived accuracy rating and confidence in the perceived accuracy decisions. As shown in Figure 4.1, we investigated the effect of a warning with a credibility explanation compared with a warning-only condition (**RQ.1**) in Experiment 1. In Experiment 2, we explored if the framing effect matters in a warning with explanation by comparing a positive framing (i.e., credibility) and a negative framing (i.e., falsity) (**RQ.1**). In Experiment 3, we examined the effect of the AI system's reliability on the proposed explanations (**RQ.2**). We also evaluated participants' trust in the AI systems across all experiments (**RQ.3**).

Participants. We recruited participants on Amazon Mechanical Turk (MTurk) through the Human Intelligent Task (HIT) for all experiments. The HITs included the task description, and workers were able to decide whether they would like to perform the task. In each experiment, we required the workers to be those who (1) are at least 18 years old, (2) live in the U.S., and (3) have finished more than 100 HITs with a HIT approval rate of at least 95%. MTurk workers were allowed to participate in our study once.

We recruited $1,246$ (EXP.1), $1,686$ (EXP.2), and $1,196$ (EXP.3) participants. We manually checked the responses and ensured that there was no duplicate participation across experiments. After removing responses (1) submitted out of the U.S; (2) with duplicate IP; (3) failed the comprehension test; (4) failed the attention check and (5) completion time shorter than 3 min (average median completion time: 15 mins); the number of participants we accepted was 710, 1014, and 968, respectively. The high exclusion rate to ensure our data quality,[3] which was necessary given the concerns on the MTurk platform [156]. Based on an hourly rate of $7.5, participants were paid $1.8 for completing a study. Participants' demographic information is shown in Table 4.1.

News Articles. We selected 25 news articles about COVID-19 released from September to November 2021. Twelve pieces of fake news were searched from *snopes.com* or *politifact.com*, both of which are representative fact-checking websites. Thirteen pieces of

[3]See Exclusion Details in the supplementary materials.

real news were selected from major news platforms such as *cnn.com* or *apnews.com*. A piece of real news was for an attention check [157].

Warning and Explanation Interfaces. In all experiments, we presented a piece of real or fake news in the form of a news headline with two fictional users' comments (see Figure 4.2 -4.5). The news part was composed of a title, a snippet of the article, and a source. For the source, real news had URLs from major news platforms where real news was excerpted. Fake news had social media URLs where the misinformation was posted. The users' comments for fake news had negation-style sentences debunking the misinformation, and the comments for real news had a neutral tone, not directly pointing out information veracity [158].

In our design, we assumed that each piece of fake news had been detected by an Artificial Intelligence (AI) system. A warning label was also shown for each piece of fake news, in which we made it clear that the fake news was disputed by an AI system (see Figure 4.2 -4.5). There was a baseline condition in each experiment in which we presented the warning label. Considering the robust effect of warning labels [122] and our main interest in the effect of AI explanations, we omitted a condition without warning and defined the baseline with a warning label as the control (CON).

An abstract way of presenting the factors that AI systems consider when debunking fake news has been developed, which serves as an explanation for AI decision-making. Seo et al. [80] presented that participants incorporated information provided by the summary index (i.e., bar chart) into their fake news evaluation. We adapted their design and created two types of explanations: positive framing (POS, see Figure 4.4) and negative framing (NEG, see Figure 4.5). We added a bar chart below the warning tag to illustrate the fake news credibility (POS) or falsity (NEG).

Positive Framing(POS). In the explanation interface, we present three factors that our hypothetical AI system considers, including content credibility based on the news title and news content, source credibility based on the news source, and social credibility based on users' comments. A filled blue bar graph accompanies each factor, and the length of each bar indicates the credibility score that the AI system derived for the evaluation of the factor. For the bar graphs, "More Credible," "Average," and "Less Credible" are marked on the top of the bar graph panel, and numbers 1 through 5 are marked on the bottom of the panel. For score 1, a small blue bar is displayed in the bar graph. For score 5, the bar is displayed at the most right edge of the bar graph. For the

bar graphs, an outline frame indicates the possible maximum score of 5 so that the range of the score was clear to the participants and easy visual comparison was enabled among the bars [159].

We made 12 bar charts to be added to the 12 pieces of fake news. We scored each factor using a 5-point score. We used either 1, 2, 4 or 1, 2, 5 for value combination avoiding 3, a neutral number. Each factor showed its high credibility (i.e., 4 or 5) four times among the 12 pieces of fake news. Moreover, we separated the 12 pieces of fake news into three sets and implemented a Latin-square design to counterbalance the credibility value combinations across the different sets. We focused on the credibility/falsity value combinations but did not control the value alignment for each factor. Our post-hoc analysis on the perceived accuracy rating of fake news showed no significant differences between the high and low source-credibility scores, suggesting limited impact.

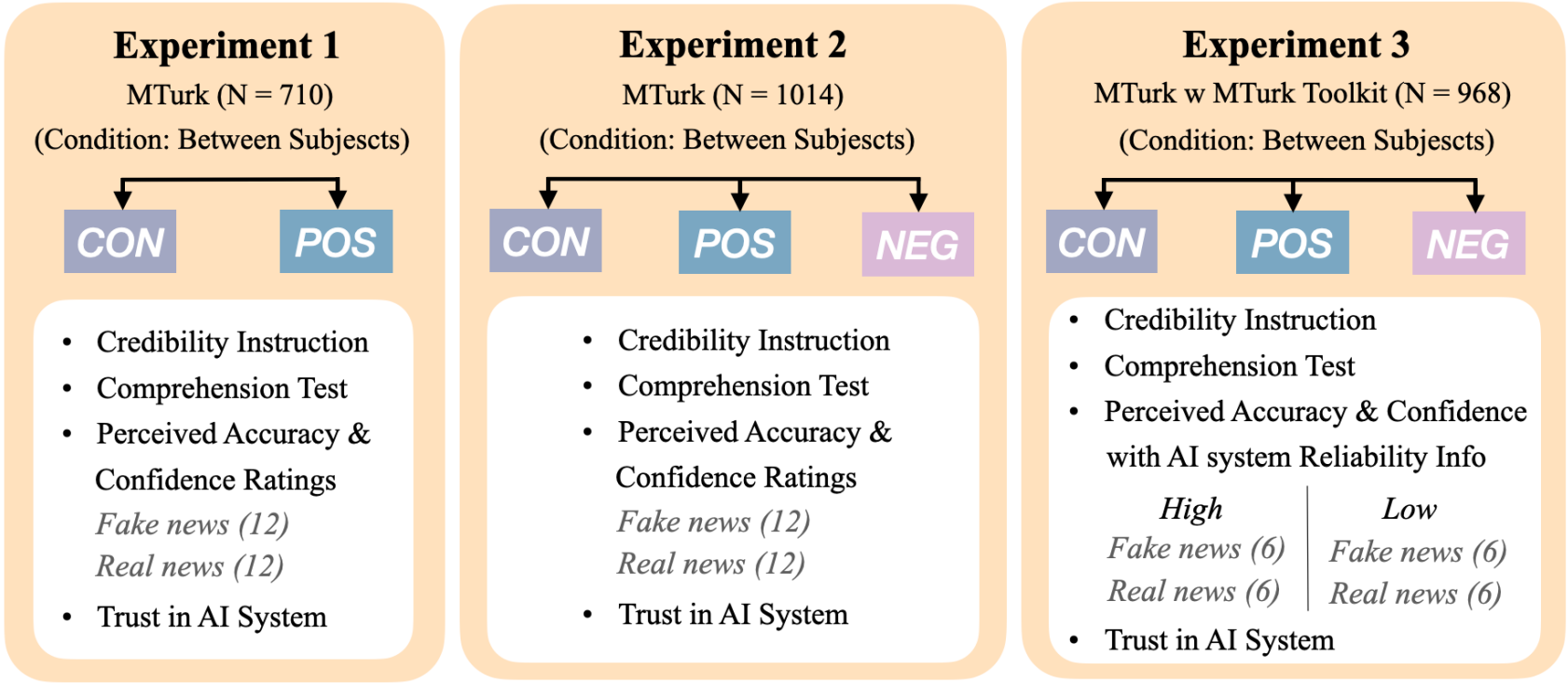

Figure 4.1: An overview of the experiment design. Experiments 1 and 2 focus on the framing effect. Experiment 3 focuses on reliability. *CON* means control, *POS* means positive framing and *NEG* means negative framing.

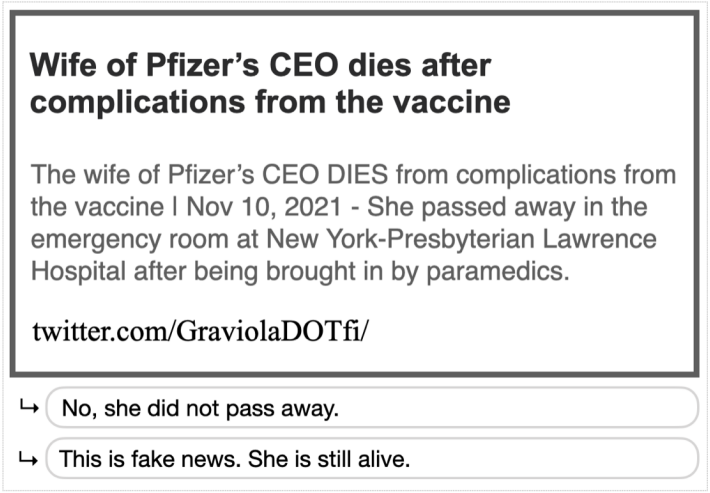

Figure 4.2: An example of fake news stimuli including COVID-19 fake news including the news title, a snippet of the news article, and a source followed by two comments.

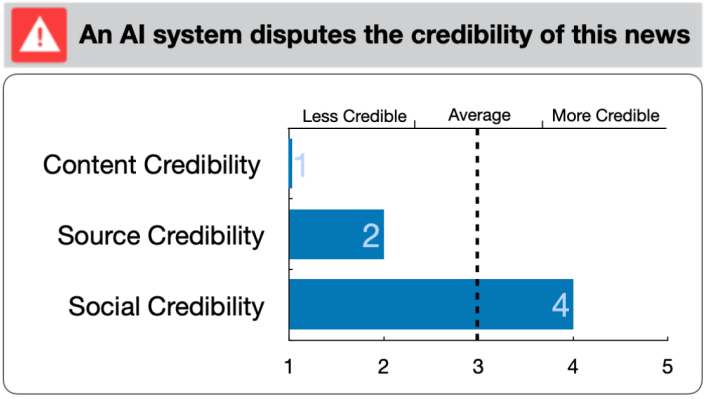

Figure 4.3: This is an example image of a warning with a positively framed explanation that was used in EXP.1.

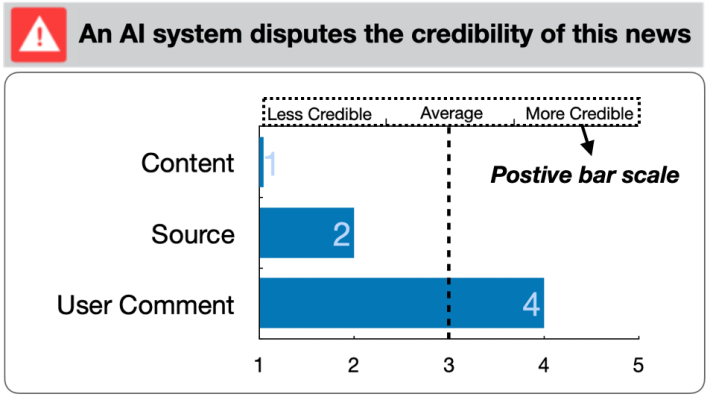

Figure 4.4: This is an example image of a warning with a positively framed explanation that was used in EXP.2 and EXP.3. We modified the y-axis' names to minimize confusion.

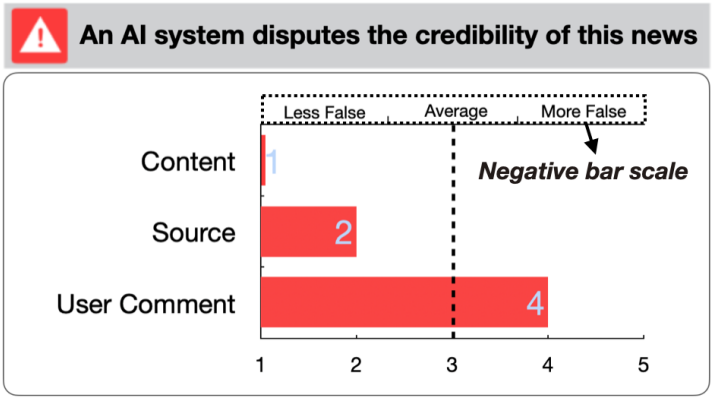

Figure 4.5: This is an example image of a warning with a negatively framed explanation. For fake news, a warning label was shown below the two comments. The bar chart for an explanation is shown below the warning label. For the warning-only condition (*CON*), only the warning message was shown.

Negative Framing (NEG). In addition to the positive framing using blue bars, we proposed a negative framing explanation [145]. The interface was the same as the credibility explanation except that we changed the wordings of the bar scale (e.g., "Less Credible" to "Less False") and the color of bar graphs from blue (Figure 4.4) to red (Figure 4.5). Instead of an equivalent framing, we applied the same score set to the falsity explanation. As for the credibility score, each factor had the highest value four times for the falsity score. Thus, compared to the credibility explanations with positive framing, fake news in the credibility explanations with negative framing was less false (See Figures 4.4 and 4.5). To make the interface comparable to that of *NEG* in Experiment 2, we removed "Credibility" in the factor description of *POS* in Experiment 2 because it duplicates the gauge on the top.

Procedure. Qualtrics was used for designing our online studies. After informed consent, participants were randomly assigned to one condition in each experiment. We first described our simulated warning system and asked participants questions to check their comprehension of our design. We asked two common questions for all conditions but added three more questions for *POS* and *NEG* to check participants' comprehension of each bar chart category. Then, the twenty-five pieces of stimuli were presented in a randomized order. Twelve of them included fake news, and thirteen of them included real news. One of the real news was for an attention check. We provided participants with specific instructions on how to answer the attention-check question [157]. For any participants who failed to follow the instructions, their survey was terminated immediately.

We paid those participants a base payment of $0.5.

We asked two questions to investigate participants' acceptance of the "claim" in the news article of each stimulus. First, participants rated their perceived accuracy rating of the news claim, "How accurate is the claim in the above news (1: very inaccurate, 7: very accurate)?". Then they answered a question about their confidence in their perceived accuracy rating, "How confident are you in answering the question above (1:not confident at all, 7:fully confident)?".

After answering questions for the 25 pieces of news, there was a post-session questionnaire. We asked participants four questions to measure their trust in the AI system disputing the fake news, including "I trust the AI System when making judgments about news veracity.", " The AI warning is informative when I make judgments about news veracity.", "The AI warning is helpful when I make judgments about news veracity.", "I would like to see the AI system implemented on social media." Participants rated their agreement for each question using a 7-point Likert scale, with "1" indicating "strongly disagree" and "7" indicating "strongly agree". In the end, participants filled in their demographic information, including age, gender, ethnicity, and education, and their experience in AI or machine learning.

4.4 Results

Our statistical analysis focused on three measures (*perceived accuracy rating*, *confidence in accuracy rating*, *trust in the AI system*). We manipulated two factors in Experiments 1 and 2, with *condition* between subjects and *news veracity* within subjects. In Experiment 3, we varied *reliability* as another within-subjects factor. We used SPSS version 29 for the data analysis.

4.4.1 Experiment 1

To quantify the effects, perceived accuracy and confidence results were entered into 2 (*news veracity*: real, fake) $\times$ 2 (*condition*: CON, POS) mixed analysis of variances (ANOVAs) with a significance level of .05. We chose ANOVAs since it is robust enough to yield the right answer even when distributional assumptions are violated [160]. Post hoc tests with Bonferroni correction were performed, testing all pairwise comparisons with corrected p values for possible inflation. The numbers of participants included for data analysis are 413 (CON) and 297 (POS).

Perceived Accuracy Rating. As shown in Figure 4.6, participants clearly distinguished the real news (5.48) from the fake news (3.65), $F_{(1,708)} = 560.03$, $p < .001$, $\eta_p^2 = .442$. Regardless of the news veracity, participants in the *POS* gave lower ratings (4.47) than those in the *CON* (4.67), in general, $F_{(1,708)} = 6.74$, $p < .010$, $\eta_p^2 = .009$. Follow-up tests on each veracity revealed that the main effect of the condition was significant for the real news, $F_{(1,709)} = 6.35$, $p < .012$, $\eta_p^2 = .012$. but not the fake news, $F_{(1,709)} = 3.07$, $p = .080$, $\eta_p^2 = .004$. Nevertheless, the two-way interaction of *news veracity* $\times$ *condition* was not significant, $F < 1.0$.

Confidence in Accuracy Rating. Participants were confident in their perceived accuracy in general (average rating above 4, see Figure 4.6). Neither the main effect of *news veracity*, $F_{(1,708)} = 2.45$, $p = .118$, $\eta_p^2 = .003$, nor the main effect of *condition*, $F_{(1,708)} = 1.52$, $p = .219$, $\eta_p^2 = .002$, were significant. The participants gave similar confidence ratings in their decisions of fake news (5.64) and real news (5.59). They also showed high confidence in both conditions (*CON*: 5.67 and *POS*: 5.58). The two-way interaction of *news veracity* $\times$ *condition* was not significant either, $F < 1.0$. Thus, the extra explanation in *POS* did not impact participants' confidence in their perceived accuracy ratings.

Trust in the AI System. We calculated the average ratings of the four questions asking about participants' trust in the AI system. Participants showed higher trust of the AI system in *CON* condition (5.51) than that in the *POS* condition (5.23), $F_{(1,709)} = 9.67$, $p = .002$, $\eta_p^2 = .013$.

Influential Credibility. To understand the relative weighting of the three factors in the *POS* condition, we analyzed the perceived accuracy rating results using ANOVA. The mean values of content credibility (3.59), source credibility (3.51), and social credibility (3.48) showed no statistical difference, $F_{(1,296)} = 2.82$, $p = .094$, $\eta_p^2 = .009$.

Summary. In Experiment 1, we examined the effect of a warning label with explanations (i.e., a summary index showing the factors that an AI system considers when debunking fake news) on participants' perceived accuracy rating of fake claims. We observed that the participants who were exposed to the extra explanations (*POS*) decreased their perceived accuracy ratings on both fake and real news, particularly on real news. However, the participants showed similar confidence in their accuracy ratings

across the conditions. Moreover, the participants showed more trust in the warning label (*CON*) than the label with extra explanations (*POS*). Thus, besides reducing participants' perceived accuracy of fake news, the extra explanations in *POS* have made participants more cautious overall.

4.4.2 Experiment 2

Human decision-making in risky contexts is influenced by how a problem is framed (The framing effect [24]). Considerations of compatibility indicate that positive dimensions are weighted more when the task is to accept, whereas negative dimensions are weighted more when the task is to reject [161]. Prior work on app selection has shown that the safety framing of the information is more compatible with a selection task for best apps [162]. Considering the AI system in our study is mainly debunking fake news, we conjecture that the ineffectiveness of the extra explanations in *POS* could be due to the positive framing in the design. We gauged the credibility of each factor ("more" means "better") and presented the credibility score using a blue color.

In Experiment 2, we proposed a *NEG* condition to further understand the effect of warning labels with extra explanations. The overall experimental setting was similar to Experiment 1 except that we added *NEG* as another condition. The *POS* and *NEG* conditions were the same, except that the wordings for the bar gauge and the color of the bar chart were different (See Figure 4.1, Figure 4.4, and Figure 4.5 for the details). The number of participants included for data analysis is as follows: 390 (*CON*), 309 (*POS*), 315 (*NEG*). To quantify the effects, perceived accuracy rating and confidence rating were entered into 2 (*news veracity*: real, fake) × 3 (*condition*: *CON*, *POS*, *NEG*) mixed analysis of variances (ANOVAs). Post hoc comparisons were conducted in the same way as Experiment 1.

Perceived Accuracy Rating. The average results of the real and fake news for each condition are shown in Figure 4.6. Same as Experiment 1, the main effects of *news veracity*, $F_{(1,1011)} = 1353.97$, $p < .001$, $\eta_p^2 = .573$, and *condition*, $F_{(1,1011)} = 5.25$, $p = .005$, $\eta_p^2 = .010$, were significant. Specifically, participants gave higher accuracy ratings for the real news (5.40) than for the fake news (3.09). The average rating of *NEG* (4.11) was smaller than those of *CON* (4.32, $p = .01$) and *POS* (4.31, $p = .02$), respectively. Follow-up tests on each veracity revealed that the simple main effect of the condition was only significant for fake news (*CON*: 3.17, *POS*: 3.22, *NEG*: 2.89), $F_{(1,1011)} = 3.24$, $p = .039$, $\eta_p^2 = .006$, but not real news (*CON*: 5.48, *POS*: 5.41, *NEG*: 5.33), $F_{(1,1011)} = 2.73$,

$p = .066$, $\eta_p^2 = .005$. Such results were opposite to what we obtained in Experiment 1, indicating the framing effect. Nevertheless, the two-way interaction of news *veracity* × *condition* was not significant either, $F_{(1,1011)} = 1.15$, $p = .316$, $\eta_p^2 = .002$. As shown in Figure 4.6, participants in the *NEG* also showed a non-significant trend of reducing their perceived accuracy for real news.

Confidence in Accuracy Rating. The average confidence rating of each condition is shown in Figure 4.6. Different from Experiment 1, participants were more confident in their ratings of fake news (5.64) than real news (5.51), $F_{(1,1011)} = 21.70$, $p < .001$, $\eta_p^2 = .021$. Participants' confidence ratings were similar across three conditions (*CON*: 5.60; *POS*: 5.57; *NEG*: 5.56), $F < 1.0$. Although the two-way interaction of *veracity* × *conditions* was significant, $F_{(1,1011)} = 3.66$, p =.026, $\eta_p^2 = .007$, posthoc tests on the main effect of the condition were not significant in either veracity, $F_s < 1.77$. Thus, the two-way interaction was mainly revealed by the participants in the *NEG* gave a numerically highest rating on the fake news but the numerically lowest rating on the real news (see Figure 4.6).

Trust in the AI System. Participants' trust score varied across conditions, $F_{(1,1011)} = 13.54$, $p < .001$, $\eta_p^2 = .026$. Post-hoc pairwise comparisons revealed that the participants in the *CON* condition trust the AI system the most (5.49), which was significantly higher than those of *POS* (5.18, p=.006) and *NEG* (4.98, $p < .001$). However, participants' trust in the two explanation conditions was similar (p=.163).

Influential Credibility. We analyzed the mean values of the perceived accuracy rating among the three factors in *POS* and *NEG* conditions using mixed ANOVAs with 3 (*factor*: content, source, comments) as a within-subject factor and 2 (*condition*: *POS*, *NEG*) as a between-subject factor. The main effects of *factor*, $F_{(1,622)} = 4, 25$, $p = .04$, $\eta_p^2 = .007$, *condition*, $F_{(1,622)} = 6.02$, $p = .01$, $\eta_p^2 = .01$, and their two-way interaction, $F_{(1,622)} = 26.76$, $p < .001$, $\eta_p^2 = .04$, were all significant. Post hoc pairwise comparisons revealed significant differences among each pair for the *NEG* condition ($p_s \leq .027$). However, only the difference between content credibility (3.31) and comments credibility (3.13) was significant ($p = .009$) in the *POS* condition. Thus, among the three factors, participants mainly relied on the content for the veracity evaluation.

Summary. In Experiment 2, we further examined the effects of warning with explanations using negative framing (*NEG*) and positive framing (*POS*). The framing

effect was revealed in the perceived accuracy rating of fake news and confidence ratings. Participants in the *NEG* tended to give lower accuracy ratings for fake news and tended to be more confident in their ratings. Same as Experiment 1, participants were suspicious about the veracity of real claims, especially in the *NEG*.

4.4.3 Experiment 3

Across Experiments 1 and 2, we obtained that the participants did not always depend upon the warning or warning with explanations for their accuracy ratings. The participants were especially suspicious when the AI system did not tag a piece of real news (i.e., a system error of miss). Such results suggest that *local* explanations of specific decisions are not sufficient. Participants have concerns about the AI system's performance at a *global* level. Thus, in Experiment 3, we varied the reliability of the AI systems to detect fake news on two levels (high vs. low) and examined its impacts on the three dependent measures.

The experimental design was the same as Experiment 2 except as noted. We varied the reliability within subjects but counterbalanced the order of the two reliabilities between subjects. At the beginning of each phase, the reliability information of the AI system was presented. We adapted the instructions of [149]. In the low-reliability phase, we presented, "In this phase, the AI system to detect fake news could be pretty unreliable, so it probably will make a lot of mistakes." In the high-reliability phase, "In this phase, the AI system to detect fake news would be pretty reliable, so it probably will NOT make a lot of mistakes." was shown. We used the same news stimuli as Experiments 1 and 2 but split it into two sets (See Figure 4.1). In each reliability phase, six pieces of real news and six pieces of fake news were shown, respectively. The two sets were chosen to have a similar distribution based on the perceived accuracy ratings of each piece of news in Experiment 2.

Perceived Accuracy Rating. Results of the average perceived accuracy ratings are shown in Figure 4.6. We ran mixed ANOVAs with 3 (*conditions*: *CON*, *POS*, *NEG*) $\times$ 2 (*news veracity*: real, fake) $\times$ 2 (*reliability*: low, high). Same as the prior two experiments, participants clearly distinguished real news (5.51) from fake news (2.15), $F_{(1,965)} = 4021.36$, $p < .001$, $\eta_p^2 = .81$. However, the perceived accuracy rating of fake news in Experiment 3 was much lower than those in the prior two experiments (see Figure 4.6), indicating a floor effect [163].

The main effect of *reliability* was also significant, $F_{(1,965)} = 56.92$, $p < .001$, $\eta_p^2 = .056$, as well as the two-way interaction of *news veracity* $\times$ *reliability*, $F_{(1,965)} = 67.57$, $p < .001$,

$\eta_p^2 = .065$. Compared to the low-reliability condition, participants in the high-reliability condition increased their accuracy ratings for the real news (low: 5.35 vs high: 5.68, $p < .001$), but their accuracy ratings for fake news showed no significant differences (low: 2.16 vs high: 2.14, $p = .656$). The three-way interaction of *news veracity* × *reliability* × *condition* was also significant, $F_{(1,965)} = 5.72$, $p = .003$, $\eta_p^2 = .012$. For the real news, the increased accuracy ratings due to increased system reliability were similar across conditions ($p_s < .001$). Such results indicate that the system's reliability *did* address the participants' suspicions on the real news evaluation. However, for the fake news, the participants in the *CON* gave lower perceived accuracy ratings when the AI system's reliability became higher ($p = .008$) but not those in the other two conditions ($p_s < .204$).

Confidence in Accuracy Rating. Same as Experiment 2, the main effect of *news veracity*, $F_{(1,965)} = 71.66$, $p < .001$, $\eta_p^2 = .069$, and the two-way interaction of *news veracity* × *condition* was significant, $F_{(1,965)} = 4.91$, $p = .008$, $\eta_p^2 = .010$. The participants showed higher confidence in their ratings of the fake news (5.76) than the real news (5.53). Such gap was more evident in the *CON* ($p < .001$) and *NEG* ($p < .001$) conditions than in the *POS* ($p = .013$) condition.

The main effect of *reliability*, $F_{(1,965)} = 5.72$, $p = .003$, $\eta_p^2 = .012$, the two-way interactions of *reliability* × *condition*, $F_{(1,965)} = 10.87$, $p < .001$, $\eta_p^2 = .022$, and *reliability* × *veracity*, $F_{(1,965)} = 7.21$, $p = .007$, $\eta_p^2 = .007$, were also significant. Participants were more confident in their accuracy ratings when the AI system's reliability became higher. For AI systems with high reliability, participants' confidence ratings varied across the conditions. Specifically, participants' rating in the *CON* (5.90) was higher than that of *NEG* (5.68, p=.007). The confidence rating in the *POS* (5.73) showed no significant differences compared to the other two conditions ($p_s > .051$). For AI systems with low reliability, participants' confidence ratings were similar across the conditions ($p_s > .797$). The confidence rating gap between real and fake news was larger when the system reliability was low (0.27) than when it was high (0.18).

Trust in the AI System. Participants' trust score varied across conditions, $F_{(1,965)} = 8.41$, $p < .001$, $\eta_p^2 = .017$. Post-hoc pairwise comparison showed that the participants trust the *CON* condition the most (5.10), followed by *POS* (5.07, $p = .003$) and *NEG* (4.77, $p < .001$). The main effect of *reliability* was also significant, $F_{(1,965)} = 331.51$, $p < .001$, $\eta_p^2 = .256$. Participants gave a higher trust score for the system of high reliability (5.21) than that of low reliability (4.52). However, the two-way interaction of

reliability × *condition* was not significant, $F < 1.0$.

Summary. We confirmed the system's reliability is critical to address the participants' suspicions about the accuracy rating of real news. When the AI system's reliability became higher, participants reduced their perceived accuracy ratings of fake news in the *CON*, but not the other two framing conditions. Such results indicate the impact of other factors, which we discuss in the next section.

4.5 General Discussion

Across three experiments, we evaluated the effect of explaining how an AI system debunks fake news on humans' detection of misinformation with a warning. We proposed credibility explanations in both positive framing (i.e., *POS*) and negative framing (i.e., *NEG*) and examined the framing effect in Experiment 2. In Experiment 3, we further varied the AI system's reliability (i.e., whether or not the AI system will make a lot of mistakes).

We obtained the evidence of the framing effect: participants who were exposed to the credibility explanation under negative framing tended to give lower accuracy ratings for fake news and tended to be more confident in their accuracy decisions. Yet, the participants did not always depend on the warning or warning with explanations for detecting misinformation. In particular, the participants became suspicious when the AI system did not tag a piece of real news (i.e., concern about the system error due to *miss*). Moreover, we confirmed that the system's reliability is critical to address such suspicion of the participants.

4.5.1 The Framing Effect on Explaining Fake News Debunking Decision

We proposed positively- and negatively-framed bar charts (i.e., a warning with explanations) showing three credibility factors that an AI system considers in debunking fake news. We examined the effect of those explanations on participants' perceived accuracy rating of fake claims. When we presented the positively-framed explanation (i.e., *POS*) in Experiment 1, we did not obtain any significant decrease in participants' accuracy rating on fake news. Instead, participants showed more doubts about real news and reduced their perceived accuracy ratings. In Experiment 2, we presented both positively- and negatively-framed (i.e., *POS*, *NEG*) explanations. Compared with the warning-only

condition (i.e., *CON*), we found that their accuracy rating of fake news was significantly reduced with the negatively-framed explanations.

One possible reason for such results is that the explanations under negative framing (i.e., "more false" and red color) are more compatible with the AI system's decision to "dispute" the news claim than the explanations under positive framing (i.e., "more credible" and blue color). Consequently, those explanations might have been more intuitive for the participants to interpret. Those results are similar to prior research that shows that users are often likely to rely more on negative information than positive information to reject apps [145, 162].

To the best of our knowledge, this is the first work to investigate the framing effect in explaining an AI system's decision to debunk fake news. We suggest future work to explore further the framing effect of debunking messages/explanations on the associated decisions and the compatibility between them.

4.5.2 The Impacts of System Reliability

With the implied truth effect [83], it is expected that participants should have little doubt in judging real news when fake news warnings are absent. Opposite to the prediction, our studies showed that participants did not credit real news cases by default but had concerns about *miss* errors (i.e., false negatives) of the AI system. We varied the AI system's reliability to detect fake news in Experiment 3 and confirmed that system reliability is critical to address the participants' suspicions about the accuracy of real news.

One possibility is that participants in the *POS* and *NEG* conditions might have paid more attention to news claims after viewing the bar charts, especially the three credibility factors. Since we did not explain how the AI system evaluated each factor and derived the score, participants seemed to have increased their bias to judge news claims as fake when the warning label and explanation were not presented for the real news. Thus, regardless of the framing, the extra explanations seemed to have made participants more conservative in detecting fake news when the reliability of the AI system is unsure. Such results are consistent with prior work, which shows that when automation systems make miss errors, users reduce their reliance on the system [164, 165]. Reliance refers to the status in which users refrain from a response when the system is silent or indicating normal operation [149]. However, when participants were informed of the system reliability in Experiment 3, their criterion of judging a piece of news as fake was adjusted. Thus, no significant differences were obtained across the three conditions

for the real news in each reliability level.

Moreover, participants in the *CON* reduced their perceived accuracy rating of fake news when the AI system became more reliable. One possible explanation is that participants might have been able to detect the fake news without any warning or explanation since the news set we implemented was collected in late 2021 while the experiment was conducted in 2022. We also obtained a somewhat floor effect on the fake news accuracy rating in Experiment 3 compared to Experiments 1 and 2. Moreover, we arbitrarily assigned score values to each factor in the bar chart. Participants might have questions about the quality of the AI explanations across the different pieces of fake news (e.g., a score of "1" for Twitter in one trial and a score of "5" in another trial). Future work could better control the accountability of the AI explanations and further investigate the interaction between the reliability and the framing effect.

4.5.3 Trust in the Warning

We investigated participants' trust in the AI system after finishing the main tasks about perceived accuracy rating and confidence rating. Across all experiments, *CON* consistently exhibited a higher trust score compared to *POS* and *NEG*, respectively. Such results are consistent with previous studies [80, 134]. For example, Seo et al. found that a *Fact-checking* warning was the most trusted condition across two experiments regardless of the most effective warning type. In their second experiment, although the *Machine-Learning-Graph* warning (which inspired our *POS* warning) was effective, the *Fact-checking* warning was still trusted the most by the participants.

These results can be understood by the effect of *familiarity* on trust. Literature in different fields has shown that familiarity contributes to building trust [166–169]. While the proposed warning explanations in our study were novel to the participants, they could be familiar with the warning label, which originated from Facebook. Moreover, the warning icon and red color have been widely used in our daily life to indicate risks or hazards [170]. Therefore, even though participants' accuracy decisions could be influenced by the extra explanations, they still showed more trust in the warning itself.

4.5.4 Higher Confidence in Fake News

Participants were confident in their perceived accuracy rating in general (average ratings above 5 points out of 7). Between real and fake news, participants showed more confidence in their decisions about fake news than those of real news, particularly in Experiments

2 and 3. Those results revealed that the participants responded to the warning or the warning with explanations when a piece of fake news was labeled. Such high *compliance* (i.e., users respond when a signal is issued) suggests that participants did not worry about any false alarm (i.e., false positives) of the AI system.

It is noteworthy that AI systems are not always reliable, showing errors of false alarm (i.e., false positives) or miss (i.e., false negatives). Our results suggest that users tend to think about false negatives rather than false positives when a warning or a warning with explanations is presented. These findings highlight the importance of informing users of the possible error types of AI systems [150].

4.5.5 Limitations

There are several limitations in the current study. *First*, we chose to recruit MTurk workers for a large sample. Although MTurk workers' demographics are more diverse compared to college students' [171], they do not represent the whole population [172]. Future studies could consider more comprehensive recruiting methods.

Second, we observed a larger perceived accuracy gap between real news and fake news in Experiment 3 compared to Experiments 1 and 2. One possible reason might be due to using MTurk toolkit provided by CloudResearch in Experiment 3, which could effectively exclude inattentive workers and enhance data quality [173]. Another possible reason could be the time gap between experiments. EXP.3 was launched about 8 months after EXP.2 due to a natural delay in the research process. Consequently, participants might have been aware of the news veracity before the study.

Third, it can be difficult for platforms to disclose the reliability of their misinformation-detecting systems. However, social media platforms such as Facebook and Twitter have been actively responding to mitigate fake news and are well aware of the issue of information transparency.[4] Thus, if our findings could be continuously verified through follow-up studies, there will be a good chance that those platforms will take the initiative to introduce warnings with explanations and provide reliability information to online users.

Lastly, our warning with explanation could be unfamiliar or not intuitive for some participants. Therefore, if some designers consider creating a warning with explanations, then they may want to consider users' graph literacy and highlight the contents' credibility information, which was considered the most for participants' perceived accuracy rating.

[4]https://transparency.fb.com/

4.5.6 Conclusion

In order to verify the effect of a *warning with explanations*, we conducted three experiments. We found that *the effect of a warning with explanations on participants' perceived accuracy rating depends on the reliability of the AI system.* If the reliability information is unknown, the negatively framed warning with explanations is more influential to participants, as they did not trust the system. When the reliability of the system was known to be high, a warning with explanations was not in effect. Rather, only a warning message was effective. Accordingly, if the level of reliability of the fake news detection system cannot be revealed, providing a warning with negatively framed explanations showing how the AI system evaluated news veracity can assist participants in avoiding fake news.

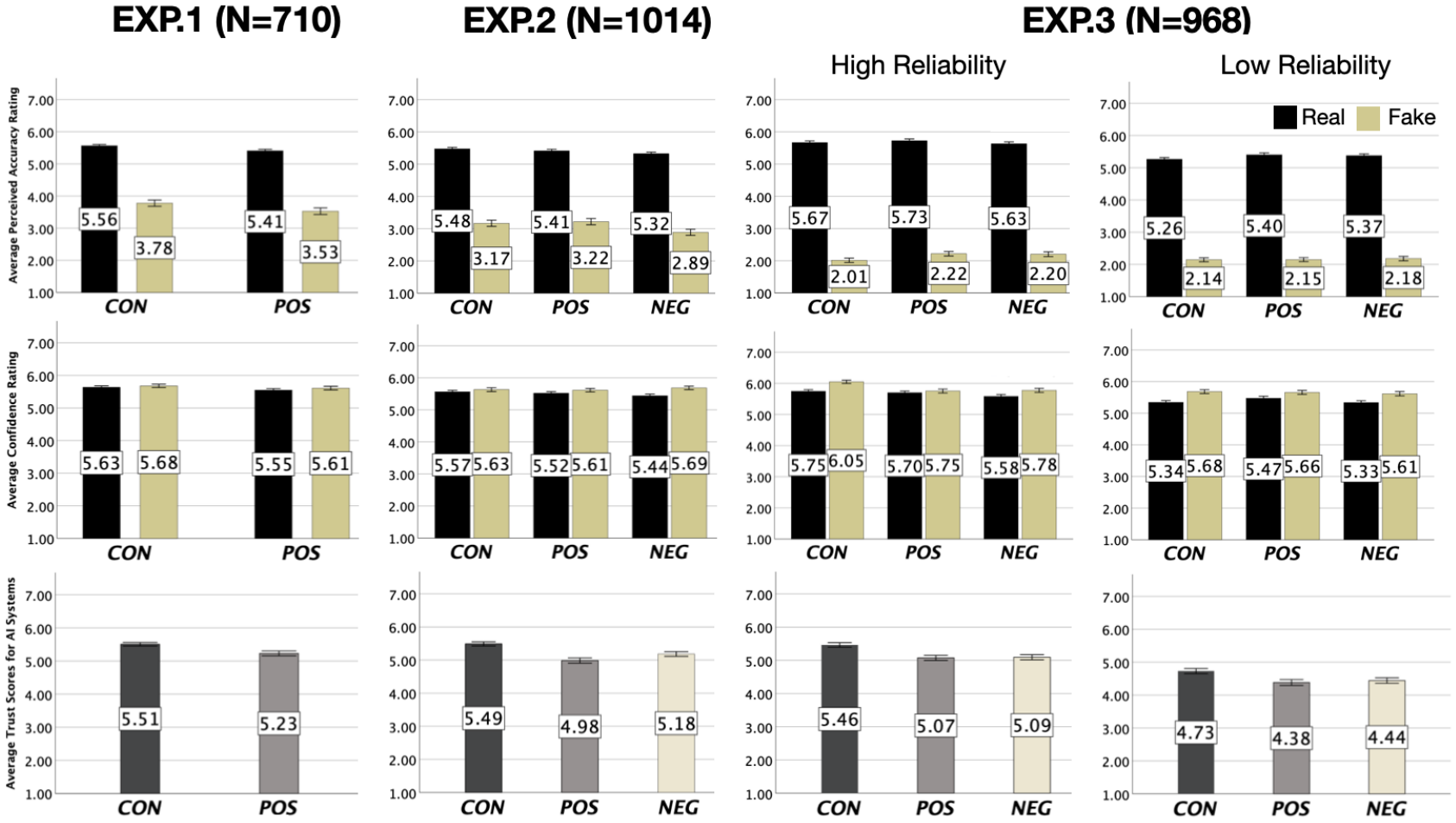

Figure 4.6: An overview of the experiment design. Experiments 1 and 2 focus on the framing effect. Experiment 3 focuses on reliability. *CON* means control, *POS* means positive framing and *NEG* means negative framing.

Chapter 5
User-Initiated: Effects of Correction Comments on COVID-19 Misinformation

In this chapter, I present a study on user-initiated correction, complementing the platform-driven correction discussed in the previous chapter. Our research focuses on evaluating the effectiveness of user-initiated correction comments, with a specific emphasis on user comments as a corrective measure in the context of social media. We paid attention to the elements that make up correcting comments, with a particular interest in who is responsible for making the correction.

We categorized users into two primary types: organizational users and individual users. Organizational users are further subcategorized into health organizations and fact-checking websites. Over the course of three experiments, we compared the effectiveness of four types of correction comments to a control condition, aiming to identify the most effective form of correction comment.

5.1 Introduction

The explosion of fake news is one of the problems that social media has triggered [30]. On social media platforms, people can easily get and share news even before checking its veracity. Thus, platforms such as Twitter and Facebook have attempted to contrive ways to curtail misinformation, such as computationally detecting fake news [174] or correcting misinformation [175]. The focus of our work is on the latter.

Despite considerable research on correcting misinformation [26, 69, 79, 80, 158], it is still premature to conclude the most effective way to correct misinformation. In

addition to the methods to correct misinformation by platforms (e.g., the popping-up warning message for questionable contents), research on *user-initiated* correction has started recently [26, 93]. It is critical to investigate effective user-initiated methods to correct misinformation as users are the main characters sharing information on social media. Moreover, prior studies demonstrated that users have actively participated in the corrections on social media [25].

During the COVID-19 pandemic, a great deal of misinformation about the virus and treatments has been pouring out on social media, threatening personal and public health worldwide. In reality, there were many cases where people died from wrong treatments for COVID-19 due to fake news.[1] Since people with high health anxiety are more associated with seeking online health information [176, 177], they can be more vulnerable to COVID-19 misinformation and more resistant to the correction compared to people with low health anxiety. Therefore, it is imperative to examine the effective correction on COVID-19 misinformation, and understand how people's health anxiety level impacts the correction effects.

Focusing on the correction on COVID-19 misinformation, we investigated the following research questions (RQs) in current work.

- **RQ 1.** Will the correction from an individual user or an organization user (e.g., a health organization or a fact-checking website) reduce participants' susceptibility to fake news relative to a control condition in which there is no correction?
- **RQ 2.** Will more frequent correction reduce participants' susceptibility to fake news more?
- **RQ 3.** Will people's health anxiety level have an impact on their susceptibility to misinformation and the effect of misinformation correction?

We conducted three online experiments ($N = 2,841$) on Amazon Mechanical Turk, examining correction from three types of users (e.g., health organizations, fact-checking websites, and individual users) (**RQ1**). We verified that correction from all three types of users could reduce participants' perceived accuracy rating on the COVID-19 fake news. Critically, we unearthed that participants counted on the reliability of the sources (e.g., social media users or URLs in the correction). We did not obtain the frequency effect (**RQ2)**. However, we discovered that participants having high health anxiety were more likely to believe fake news than low anxiety participants. We also obtained

[1] https://www.bbc.com/news/world-53755067

evidence showing the correction effect only for participants with low health anxiety. Those results imply that people with high health anxiety are more susceptible to the COVID-19 misinformation and more resistant to the misinformation correction (**RQ3**).

To the best of our knowledge, the current paper is the first to experimentally investigate the effective user-initiated correction on COVID-19 misinformation. Our experiments improve the understanding of user-initiated correction on misinformation through the following contributions.

- Via systematically-designed experiments, we obtained correction effect on fake news from three different types of users: individual users, health organizations, and fact-checking websites.
- We unearthed that people weigh the reliable sources in correction the most when deciding perceived accuracy rating.
- We found that individuals with high health anxiety are more susceptible to health-related misinformation than those with low health anxiety and correction can work better for individuals with low health anxiety.

5.2 Related Work

5.2.1 User-Initiated Correction

Since users are the main actors of information sharing on social media [89, 90], the other approach concentrates upon user-initiated correction using user comments or posts [26, 91]. Vraga and Bode investigated the effect of user-initiated correction on Zika virus misinformation using users' comments on social media [26,69,93]. They manipulated one post within a simulated Twitter or Facebook feed. In the control condition, there was no misinformation. In the treatment conditions, a piece of fake news about the Zika outbreak in the U.S. was presented, with the image and the headline claiming that the release of GMO mosquitoes caused the outbreak. Following the misinformation, a debunking sentence with a reference link from the Centers for Disease Control and Prevention (CDC) was presented. Across conditions, the source of correcting comments varied from an organization (e.g., CDC), an individual, or both.

They recruited undergraduate students and measured participants' misperceptions about the Zika virus before and after the correction of misinformation. Across studies, they demonstrated that social correction could work if it includes sufficient source

information, including organization logos (such as *snopes.com* and CDC) and reference links from the organizations. Their results also revealed that platforms (Facebook or Twitter) did not matter [93] and social and algorithmic corrections were equally effective in mitigating misperceptions [69]. One of their studies found the correction effect from a reputable organization (e.g., CDC) but not from an individual user [26].

Moreover, recent work showed that people experience both misinformation and its correction on social media. Bode and Vraga (2021) conducted an online survey about people's correction experience regarding COVID-19 misinformation on social media. Out of 1,094 participants, 34% witnessed that others' wrong beliefs were corrected and 22% corrected others' misinformation, demonstrating users' active participation in correction. Given the impacts of user-initiative comments, it is critical to examine the effective correction in the setting of COVID-19. In our current work, we evaluated a single correction comment initiated by different social media users (e.g., an organization user or an individual user) with a more structured experimental design using COVID-19 news.

5.2.2 Health Anxiety

Besides an innumerable amount of COVID-19 fake news being spread online [178], many cases related to health anxiety (i.e., hypochondria) and mental health issues have been reported during COVID-19 pandemic [179]. Prior studies [180, 181] explain abnormally increased fear and anxiety for health as one reason for information seeking about COVID-19, which is also called "Cyberchondria" [177, 182]. Considering the health news topic, we also measured participants' health anxiety [183] and analyzed its impact on the effect of correction for COVID-19 misinformation.

5.3 Method

We investigated a single user-initiated correction comments using the format of the tweet in three online experiments. In contrast to prior works [26, 27], we recruited Amazon Mechanical Turk (MTurk) workers who are more demographically diverse [171, 184, 185] than college students. In each experiment, participants evaluated eight to twelve pieces of fake and real news about COVID-19 instead of evaluating a single piece of misinformation. We examined the effect of correction comments on people's acceptance of fake news about COVID-19 compared to a control in which comments without correction were

presented, rather than examining the correction effect by comparing participants' pre- and post-perception of misinformation [26].

We examined COVID-19 misinformation, which is most timely. Moreover, we ran our experiments at three different time points over six months during the COVID-19 pandemic. We updated the evaluated news set from Experiment 2 to reflect the rapidly pouring news. We proposed and evaluated different types of users on social media (e.g., individual users, health organizations, and fact-checking websites).

Experiment 1 examined the effect of a single correction comment about COVID-19 fake news by individual users or organization users. Experiment 2 was conducted to replicate the findings of Experiment 1 using a new set of COVID-19 real and fake news. Experiment 3 increased the types of correction to be investigated for a more systematic understanding of the correction effect.

For Experiments 1 and 2, using a between-subject design, we investigated the effect of correction comments on COVID-19 misinformation across three conditions: no correction (*CON*), correction by an individual user (*hIND*), and correction by a health organization (*hORG*). Specifically, fake news stimuli of *hIND* and *hORG* had the same correcting comment, including a reference link from a health organization. "*h*" indicates the identical link from health organizations (CDC or WHO). Moreover, to investigate the frequency effect of correction, we composed Phases 1 and 2. Half of the fake and real news of Phase 1 was presented again in Phase 2, each of which was with a similar correcting comment but from a different user and a different organizational source.

Besides health organizations, fact-checking websites investigated false claims about COVID-19 from the beginning of the pandemic [186]. Thus, we ran Experiment 3 not only to verify again the correction effects of existing conditions (*hORG*, *hIND*) but also to examine the correction effects with two new conditions that are relevant to fact-checking websites (*fcORG*, *fcIND*, where "*fc*" indicates fact-checking websites). All experiment designs and news contents were the same as Experiment 2 except as noted. *fcORG* condition included correction from a fact-checking website with a reference link from the site. *fcIND* condition included the identical reference link as the *fcORG* condition, but the correction was from an individual user. Regarding frequency effect, for both the *fcORG* and *fcIND* conditions, the link of *snopes.com* was used for the correction at Phase 1, and the link of *politifact.com* was used for the correction at Phase 2.

Participants In this and the following experiments, we recruited participants by posting the Human Intelligence Task (HIT) on MTurk. We restricted the workers to

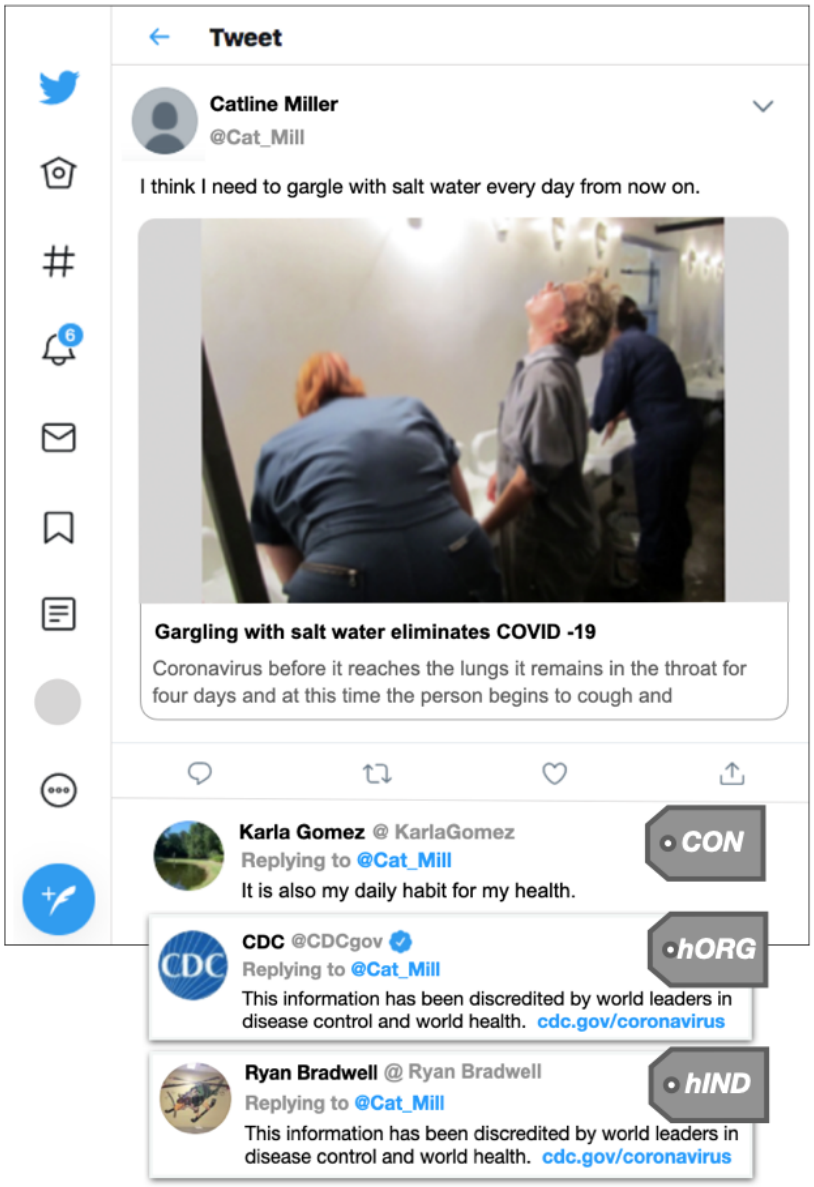

Figure 5.1: The average values of perceived accuracy ratings as a function of frequency × condition for real news (left panel) and fake news (right panel) with one standard error.

those who (1) were at least 18 years old; (2) were located in the U.S.; and (3) completed more than 100 HITs with a HIT approval rate of at least 95%. Qualtrics was used to program our online studies. Our study was approved by the institutional review board (IRB) office at our institution.

Materials For Experiment 1, we selected eight news articles about COVID-19 released between March and May 2020 from *snopes.com* or *politifact.com*, both of which are well-

regarded fact-checking websites. Four pieces of the news were fake and the other four pieces were real. Also, we used another piece of real news for an attention check [157].

As shown in Figure 5.3, we created a simulated Twitter interface in which each piece of news was embedded within a tweet message. For each stimulus, a tweet message from a fictional user was presented above the COVID-19 news. The tweet message was a short sentence related to the news without any correcting message. The embedded news was composed of an image, a headline, and a snippet of the news article. Following the news article, a comment from another user was presented.

For the fake news, each comment in the *hIND* and *hORG* conditions included a sentence pointing out the falsity of the fake news articles with a reference link from an authoritative organization (CDC for Phase 1, WHO for Phase2). The correction messages and the reference links were the same between *hORG* and *hIND*.

In contrast, each comment of the fake news in *CON* did not contain a correcting sentence or reference link. Instead, the comment included a commenter's plausible but non-correcting messages varied according to the contents of each piece of news. Likewise, the real news comments had non-correcting messages, which were constructed in the same way as the fake news in *CON*. We used the same set of real news and its comments across the three conditions [2].

When we presented half of the stimuli again at Phase 2, we varied the user of the comment for both fake and real news regardless of conditions. Furthermore, we presented a different reference link for fake news of *hORG* and *hIND* at Phase 2 (e.g., the comment from CDC and a reference link, *cdc.gov/coronavirus*, in Phase 1, the comment from World Health Organization (WHO) and a reference link, *who.int/coronavirus*, in Phase 2). The expressions of comments at both phases conveyed the same content but with some wording changes. All authors reached a consensus on the contents of all messages we used for experiments.

To replicate the findings of Experiment 1, we conducted Experiments 2 and 3 with up-to-date COVID-19 news articles released from May to July 2020. The experimental setting was the same as Experiment 1 except as noted. We created twelve stimuli, half about fake news and the other half about real news, on the simulated Twitter interface of Experiment 1. The attention check was between Phases 1 and 2.

Procedure Figure 5.2 illustrates the flow chart of Experiment 1. Participants were randomly assigned to one of three conditions. After participants provided informed

[2]https://osf.io/dxs9c/

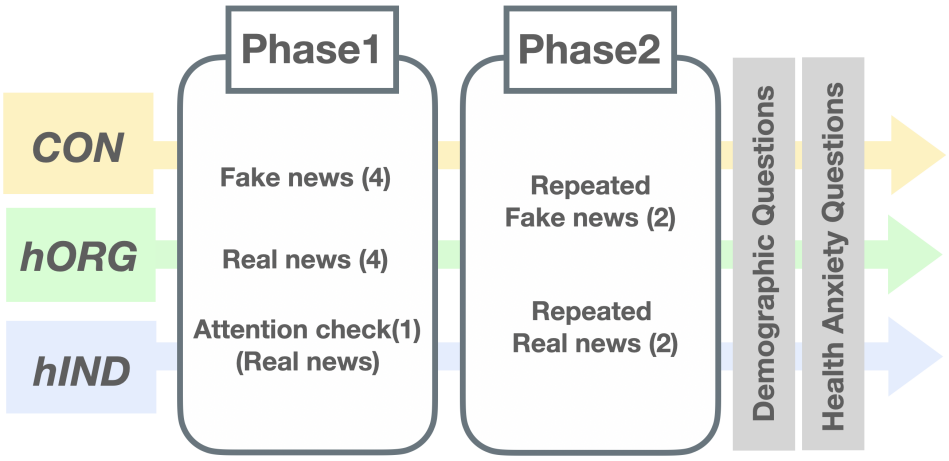

Figure 5.2: A flow chart of Experiment 1. *CON*, *hORG*, and *hIND* refer to the three between-subject conditions. In Phase 1, four pieces of fake news stimuli and four pieces of real news stimuli were shown in a randomized order for participants. One piece of real news stimulus was presented for an attention check. In Phase 2, half of the fake news stimuli and half of the real news stimuli from Phase 1 were shown again. We used a semi-Latin-square design for a better-balanced assignment of news shown in Phase 2. All stimuli in Phase 2 were randomly presented as well. After Phase 2, questions of demographic information and health anxiety were asked as post-session questions in Experiment 1.

consent, Phase 1 started, in which the eight pieces of stimuli were presented in a randomized order. Half of them included fake news, and the other half included real news. We asked two questions to examine participant's susceptibility to the "claim" embedded in the news article of each stimulus. First, participants answered two questions. The first question is "How accurate is the claim in the above news?" on a 7-point scale with "1" meaning "Very inaccurate" and "7" meaning "Very accurate." Then they rated their willingness to share the news by answering, "Would you consider sharing this news online (for example, through Facebook or Twitter)?" using another 7-point scale with "1" meaning "Never" and "7" meaning "Always."

Following Phase 1, two pieces of real news stimuli and two pieces of fake news stimuli were presented once more in a randomized order in Phase 2 to investigate the effect of correction frequency. We used a semi-Latin-square design for a better-balanced assignment of news shown in Phase 2. Participants answered the same two questions for each piece of stimuli as Phase 1.

After Phase 2, there was a post-session questionnaire. Participants first filled in their demographic information, including age, gender, ethnicity, and education. Then we measured participants' health anxiety level using four representative questions [183].

A 5-point scale with "1" meaning "None at all" and "5" meaning "A great deal" was used for the first two questions: 1) "How much do you usually worry about your health?" and 2) "How much are you ever worried that you may get a serious illness in the future?" Another 5-point scale with "1" meaning "Rarely" and "5" meaning "Usually" was used for the latter two questions: 3) "How often do you tend to read up about illness and disease to see if you may be suffering from one?" and 4) "How often do your bodily symptoms stop you from concentrating on what you are doing?" Participants were allowed to choose "Prefer not to answer" for all post-session questionnaires except for age.

An extra piece of real news was included in Phase 1 to exclude inattentive participants. We presented specific instructions about how to answer the attention-check question to the participants. For any participants who failed to follow the instructions, their survey was terminated immediately.

Meanwhile, in Experiments 2 and 3, we added two political-stance-related questions in the post-session questions due to the impact of political ideology on people's susceptibility to COVID-19 misinformation [187]. Moreover, we included follow-up questions to identify which factors among the given stimuli influenced participants' perceived accuracy rating.

5.4 Results

5.4.1 Experiment 1

We recruited $1,275$ MTurk workers in July 2020. We accepted 907 participants' answers after removing five responses submitted out of the U.S., two responses submitted within two minutes (median completion time was about six minutes), 103 responses that failed an attention check, and 258 duplicated submissions. The numbers of participants of the three conditions included in the data analysis are as follows: 295 (CON), 305 ($hORG$), and 307 ($hIND$). We paid \$0.75 for participants who completed the task based on an hourly payment of \$7.5. Participants' demographic information is shown in Table 4.1.

For data analysis, we used three levels of news frequency: $ONCE$ (Phase 1 results of news shown at Phase 1 only), TWO_{1st} (Phase 1 results of news shown at both phases), TWO_{2nd} (Phase 2 results of news shown at both phases).

Perceived accuracy rating and willingness-to-share measures were entered into 3 (condition: CON, $hORG$, $hIND$) $\times$ 2 (veracity: *fake*, *real*) $\times$ 2 (frequency: $ONCE$, TWO_{2nd}) mixed analysis of variances (ANOVAs) [188] with a significance level of .05, respectively. Post-hoc tests with Bonferroni correction were performed. We report the

Item	Options	Exp.1	Exp.2	Exp.3
Gender	Female	49.4%	51.8%	55.7%
	Male	49.6%	47.8%	44.2%
	Prefer not to answer	1.0%	0.4%	0.2%
Age	18-27	20.6%	19.0%	16.4%
	28-37	33.8%	33.6%	39.5%
	38-47	21.1%	20.8%	23.3%
	48-57	13.3%	13.9%	11.9%
	58-67	8.2%	9.6%	6.9%
	Over 67	3.0%	3.0%	2.1%
Ethnicity	Asian	8.9%	7.3%	5.9%
	African American	12.4%	9.8%	11.8%
	Hispanic/Latino	4.5%	5.5%	6.6%
	Caucasian	70.3%	74.4%	73.5%
	Other	2.8%	2.3%	1.6%
	Prefer not to answer	1.1%	0.8%	0.5%
Education	High school	6.3%	7.9%	7.0%
	Bachelor's degree	48.5%	44.9%	47.4%
	Master's degree	20.3%	19.0%	19.2%
	Doctorate degree	2.9%	2.1%	3.3%
	Other	21.8%	25.5%	22.9%
	Prefer not to answer	0.2%	0.5%	0.3%

Table 5.1: Demographic information of the participants in the three experiments.

effect size using η_p^2 reported by SPSS [26, 189].[3]

To understand the frequency effect, we also analyzed the results between *ONCE* and TWO_{1st}, and between TWO_{1st} and TWO_{2nd}. Across the three experiments, the analysis results did not show any significant difference between *ONCE* and TWO_{1st}. Also, the results of the analysis between TWO_{1st} and TWO_{2nd} showed similar patterns and had only marginal differences compared to the results between *ONCE* and TWO_{2nd}. Thus, we evaluated the main analysis between *ONCE* and TWO_{2nd} in this and the following experiments.

Perceived Accuracy Rating. Results of the average perceived accuracy rating are shown in Figure 5.3. Participants clearly distinguished real news (5.15) from fake news (2.55), ${F_{(1,904)}}^4 = 1914.98$, $p < .001$, $\eta_p^2 = .679$. The two-way interaction of news veracity

[3]We are aware of η_G^2, which was recommended to report [190, 191] considering factors manipulated and measured between subjects [192]. To make the results comparable to the literature [26], we report the η_p^2 to show the effect size based on SPSS analysis.

[4]F value equals to variance estimate based on variability among group means divided by variance estimate based on variability within groups. Hence, the larger F value indicates that the variability in

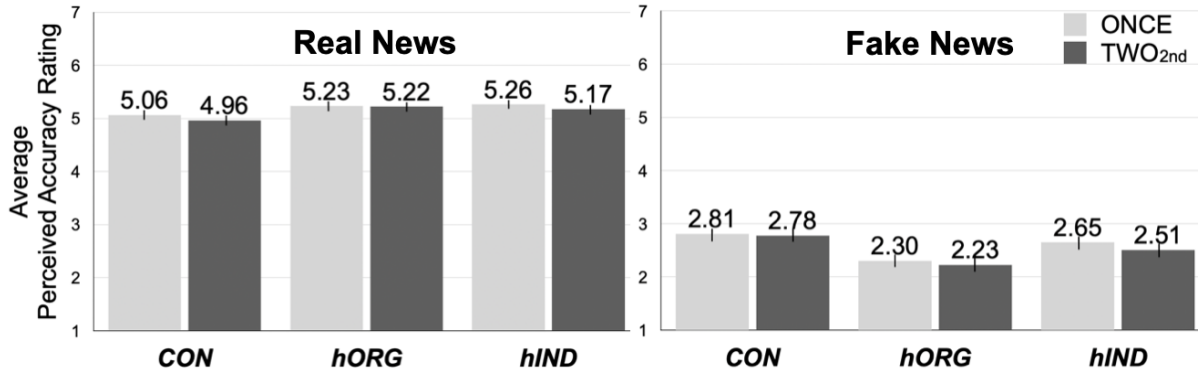

Figure 5.3: The average values of perceived accuracy ratings in Experiment 1 as a function of frequency × condition for real news (left panel) and fake news (right panel) with one standard error.

× condition was also significant, $F_{(2,904)} = 12.85$, $p < .001$, $\eta_p^2 = .028$. Post-hoc tests revealed that the effect of the condition was significant for both fake news, $F_{(2,904)} = 6.65$, $p = .001$, $\eta_p^2 = .014$, and real news, $F_{(2,904)} = 4.29$, $p = .014$, $\eta_p^2 = .009$. Nevertheless, the effect of condition revealed different patterns. For fake news, only participants in the *hORG* condition (2.27) gave lower accuracy ratings relative to the *CON* condition (2.79), $p_{adj.} = .001$. The other two pairwise comparisons (i.e., *CON* vs. *hIND* (2.58), $p_{adj.} = .427$, and *hORG* vs. *hIND*, $p_{adj.} = .090$) were not significant. For real news, relative to *CON* (5.01), participants gave higher accuracy ratings for both *hORG* (5.22), $p_{adj.} = .029$, and *hIND* (5.22), $p_{adj.} = .039$. However, the perceived accuracy ratings between *hORG* and *hIND* conditions were not significantly different, $p_{adj.} > .999$.

Thus, we obtained the correction effect of a comment from an organization user for COVID-19 fake news, which is in agreement with the prior work about Zika virus misinformation [26]. Also, we found a positive impact of the news correction on real news.

Sharing Decisions. Results of the willingness-to-share measure are presented in Figure 5.4. Participants' willingness to share real news (3.68) was higher than that of fake news (2.35), $F_{(1,904)} = 708.13$, $p < .001$, $\eta_p^2 = .439$. The interaction between veracity × condition only showed a trend to be significant, $F_{(2,904)} = 2.87$, $p = .057$, $\eta_p^2 = .006$. Moreover, participants' overall rating for the willingness-to-share measure (*real*: 3.68, *fake*: 2.35) was lower than that of the perceived accuracy rate (*real*: 5.15, *fake*: 2.55), indicating that they tended to be conservative in sharing decisions than perceived accuracy evaluation.

the measurements is mostly determined by the group differences [188].

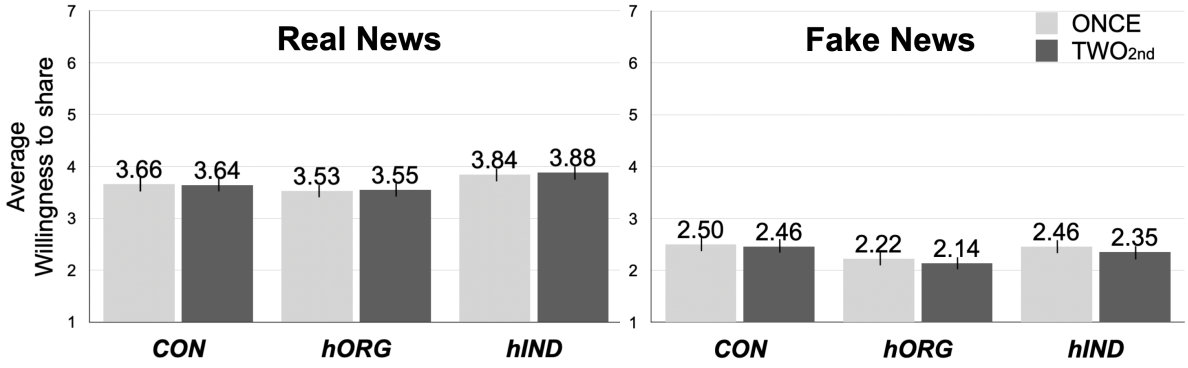

Figure 5.4: The average values of willingness-to-share in Experiment 1 as a function of frequency × condition for real news (left panel) and fake news (right panel) with one standard error.

Health Anxiety. We calculated a mean score of the four questions about health anxiety after removing the results of two participants who refused to answer all of the questions (*CON*:294, *hORG*:305, *hIND*:306). We then classified the results into two groups: *low health anxiety* (scores from 1 to 2) and *high health anxiety* (scores from 3 to 5). For the statistical tests, we added *health anxiety* as an additional between-subject factor into the main analysis. For perceived accuracy rating, participants with *high health anxiety* (4.30) gave a higher accuracy rating than those with *low health anxiety* (3.58), $F_{(1,899)} = 93.24$, $p < .001$, $\eta_p^2 = .094$, and its interaction with veracity, $F_{(1,899)} = 71.67$, $p < .001$, $\eta_p^2 = .074$, were significant. The effect of health anxiety was more evident for the fake news (*high health anxiety*: 3.28; *low health anxiety*: 2.09), $F_{(1,903)} = 104.84$, $p < .001$, $\eta_p^2 = .104$ than for the real news (*high health anxiety*: 5.29; *low health anxiety*: 5.06), $F_{(1,903)} = 10.99$, $p < .001$, $\eta_p^2 = .012$.

Likewise, participants with *high health anxiety* (3.78) showed more willingness to share news than those with *low health anxiety* (2.57), $F_{(1,899)} = 125.79$, $p < .001$, $\eta_p^2 = .123$. The two-way interaction of veracity × health anxiety was also significant, $F_{(1,899)} = 9.76$, $p = .002$, $\eta_p^2 = .011$. The effect of health anxiety was also more evident for the fake news (*high health anxiety*: 3.19; *low health anxiety*: 1.84), $F_{(1,903)} = 124.00$, $p < .001$, $\eta_p^2 = .121$, than for the real news (*high health anxiety*: 4.33; *low health anxiety*: 3.29), $F_{(1,903)} = 78.45$, $p < .001$, $\eta_p^2 = .080$. We also obtained a four-way interaction of veracity × frequency × conditions × health, $F_{(2,899)} = 3.13$, $p = .044$, $\eta_p^2 = .007$. The post-hoc test presented that it was mainly due to a non-significant trend of correction effect on fake news for the *low health anxiety* group, $F_{(2,557)} = 2.81$, $p = .061$, $\eta_p^2 = .010$, suggesting the

misinformation susceptibility of people with high health anxiety was somewhat difficult to mitigate.

Summary. In Experiment 1, we examined the effects of correcting comments from health organizations (*hORG*) and individual users (*hIND*) (**RQ1**), as well as the effect of correction frequency (**RQ2**) in helping users mitigate fake news. Consequently, we verified the effect of correction from a health organization (*hORG*) for reducing perceived accuracy rating on fake news as in the prior study [26] but did not find a frequency effect. Furthermore, we discovered perceived accuracy rating of real news was higher given correction compared to *CON*, which indicates correction increased participants' confidence in real news through learning effects from the correction on fake news.

In addition, we found that participants with high health anxiety were more susceptible to COVID-19 misinformation than those with low health anxiety (**RQ3**): People with high health anxiety gave higher perceived accuracy ratings and were more willing to share health-related news than those with low health anxiety, and such a pattern was more evident for the fake news than for the real news.

5.4.2 Experiment 2

We recruited 1,255 MTurk workers in December 2020. We accepted 768 participants' answers after removing two responses submitted out of the U.S., 291 who failed an attention check, 186 duplicated submissions, and eight responses submitted in less than three minutes (median completion time is about ten minutes). The numbers of participants included for data analysis are as follows: 253 (*CON*), 261 (*hORG*), and 254 (*hIND*). The base payment was \$0.50. There was a bonus of \$0.75 for participants who passed the attention check and completed the task. The payment rate (\$7.5/hr) is the same as Experiment 1. Participants' demographic information is shown in Table 4.1. We analyzed the data in the same way as Experiment 1.

Perceived Accuracy Rating. Average results of the real and fake news for each condition are shown in Figure 5.5. The main effects of news veracity, $F_{(1,765)} = 1411.78$, $p < .001$, $\eta_p = .649$, condition, $F_{(2,765)} = 4.15$, $p = .016$, $\eta_p^2 = .011$, as well as the two-way interaction of news veracity $\times$ condition, $F_{(2,765)} = 17.53$, $p < .001$, $\eta_p^2 = .044$, were significant. Same as in Experiment 1, participants can distinguish real news (4.91) from fake news (2.71). Post-hoc analysis revealed that the perceived accuracy ratings across the conditions were similar for real news but different for fake news. Compared to *CON*

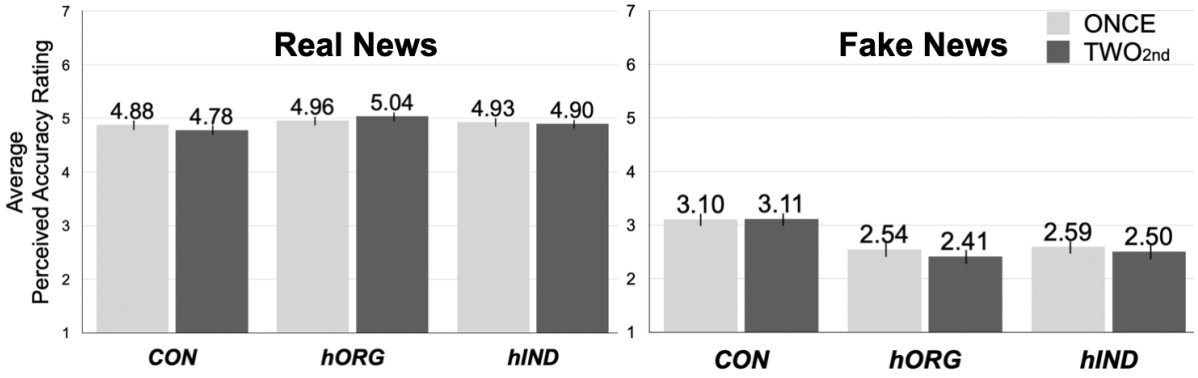

Figure 5.5: The average values of perceived accuracy ratings in Experiment 2 as a function of frequency × condition for real news (left panel) and fake news (right panel) with one standard error.

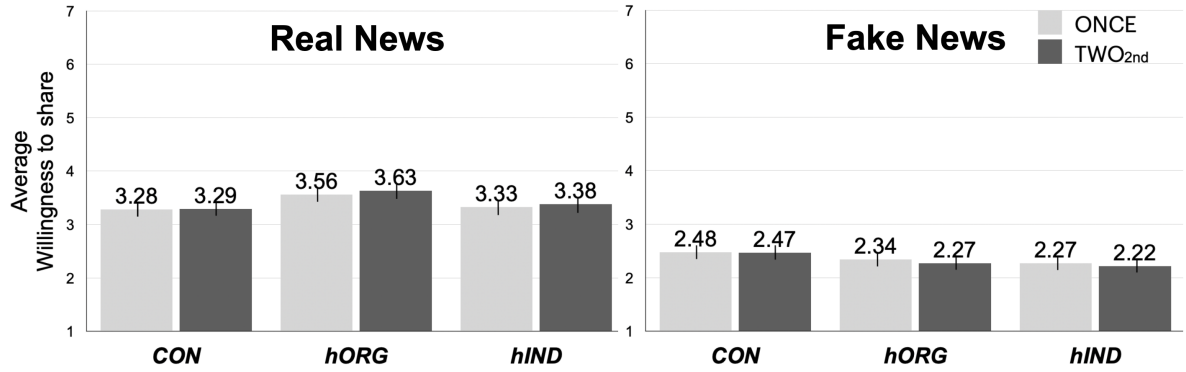

Figure 5.6: The average values of willingness-to-share in Experiment 2 as a function of frequency × condition for real news (left panel) and fake news (right panel) with one standard error.

(3.11), a lower accuracy rating was evident for fake news at *hORG* (2.48), $p_{adj.} < .001$, and *hIND* (2.55), $p_{adj.} < .001$, respectively. Thus, the correction effect was evident for both *hORG* and *hIND* conditions. In addition, there was a three-way interaction of veracity × frequency × conditions, $F_{(2,765)} = 3.45$, $p = .032$, $\eta_p^2 = .009$, showing decreased perceived accuracy rating for fake news from *ONCE* to TWO_{2nd} in *hORG* (2.54 → 2.41) and *hIND* (2.59 → 2.50) but not in *CON* (3.10 → 3.11).

Sharing Decisions. Results of willingness-to-share measure are presented in Figure 5.6. Participants showed more willingness to share real news (3.41) than fake news (2.34), $F_{(1,765)} = 480.27$, $p < .001$, $\eta_p^2 = .386$. The interaction of veracity × condition was significant, $F_{(2,765)} = 8.25$, $p < .001$, $\eta_p^2 = .021$. The main effect of the condition was not significant at each veracity level. However, participants' willingness-to-share for the *hORG* and *hIND* conditions showed a trend to be larger than that of *CON* for the real

news, while an opposite pattern was revealed for fake news.

Moreover, the two-way interaction of veracity × frequency approached significance, $F_{(1,765)} = 3.76$, $p = .053$, $\eta_p^2 = .005$, suggesting an increase of willingness to share for real news but a decreasing trend for fake news from *ONCE* to TWO_{2nd}. As in Experiment 1, the average willingness-to-share measure was lower than that of the perceived accuracy rating, indicating that participants tended to be conservative in sharing news regardless of news accuracy.

Health Anxiety. After removing the results of three participants who did not complete the questions, we analyzed 765 (*CON*:251, *hORG*:261, *hIND*:253) participants' results by adding *health anxiety* in the main analyses. For perceived accuracy rating, participants with *high health anxiety* (4.09) gave a higher accuracy rating than those with *low health anxiety* (3.63), $F_{(1,759)} = 37.62$, $p < .001$, $\eta_p^2 = .047$. And its interaction with veracity, $F_{(1,759)} = 7.41$, $p = .007$, $\eta_p^2 = .010$ was also significant. As in Experiment 1, the effect of health anxiety was more evident for the fake news (*high health anxiety*: 3.09; *low health anxiety*: 2.46), $F_{(1,763)} = 28.07$, $p < .001$, $\eta_p^2 = .035$, than for the real news (*high health anxiety*: 5.10; *low health anxiety*: 4.80), $F_{(1,763)} = 18.73$, $p < .001$, $\eta_p^2 = .024$. Moreover, the three-way interaction of veracity× condition × health anxiety was significant, $F_{(1,759)} = 3.38$, $p = .034$, $\eta_p^2 = .009$. Post-hoc comparison showed that the correction on fake news turned out to be effective for participants with *low health anxiety* in the *hORG* (2.09), $p_{adj.} < .001$, and the *hIND* (2.27), $p_{adj.} < .001$, than those in the *CON* (3.03), respectively.

Participants with *high health anxiety* (3.44) gave higher willingness-to-share score than those with *low health anxiety* (2.51), $F_{(1,759)} = 67.99$, $p < .001$, $\eta_p^2 = .082$. Thus, people who are highly anxious about their health tend to share more health-related news regardless of news veracity.

Political Stance. At the post-session questions, we also measured participants' political stance with a 5-point scale ("1" meaning "very liberal," "5" meaning "very conservative"). Participants who gave a rating of "1" or "2" were categorized as liberals (336), and those who gave a rating of "4" or "5" were categorized as conservatives (228). We excluded moderates (204), i.e., who gave a rating of "3" from the data analysis. We added political stance (*liberals, conservatives*) as another factor into ANOVAs of perceived accuracy rating and willingness-to-share measure, respectively.

For both perceived accuracy rating and willingness-to-share measure, the main effect

of political stance, $Fs_{(1,558)} = 52.89$ and 30.50, $ps < .001$, $\eta^2_{ps} = .087$ and .052, and its interaction with veracity, $Fs_{(1,558)} = 135.63$ and 65.09, $ps < .001$, $\eta^2_{ps} = .196$ and .104, were significant. Specifically, for perceived accuracy rating, the effect of political stance was only significant for the fake news (*liberals*: 2.18, *conservatives*: 3.58), $F_{(1,562)} = 108.85$, $p < .001$, $\eta^2_p = .162$, but not for the real news (*liberals*: 5.0, *conservatives*: 4.91), $F_{(1,562)} = 1.306$, $p = .254$, $\eta^2_p = .002$. For the willingness-to-share measure, the effect of political stance was more evident for the fake news (*liberals*:1.90, *conservatives*: 3.12), $F_{(1,562)} = 66.94$, $p < .001$,$\eta^2_p = .106$, than for the real news (*liberals*:3.32, *conservatives*: 3.65), $F_{(1,562)} = 4.84$, $p = .028$, $\eta^2_p = .009$.

Yet, we did not obtain the three-way interaction of political stance × veracity × condition for perceived accuracy rating or willingness-to-share measure, $Fs_{(1,558)} = 2.18$ and 1.82, $ps = .114$ and .162, $\eta^2_{ps} = .008$ and .006, indicating minimal impacts of correction on addressing conservatives' higher susceptibility to COVID-19 misinformation.

Influential Factors. In the post-session question, we also asked participants to specify factors that impacted their perceived accuracy rating, including "user tweet (text)," "users tweet (image)," "comments," and "others." For participants who selected "comments," we further asked them to select the parts of comments that affected their decision the most among the options "who wrote the comment," "how persuasively the comment was written," "whether the comment included a reference URL," and "other."

As shown in Figure 5.7 top panel, the majority of the participants chose "other's comments" as the most influential factors, and the results were similar between *hORG* (36%) and *hIND* (33.9%). For participants who chose "comment," those in the *hORG* condition chose "who wrote" the most (67.6%) while those in the *hIND* condition chose "URL" the most (49.6%), $\chi^2_{(3)} = 103.54$, $p < .001$, revealing the influence of reliable sources.

Summary. In Experiment 2, we not only replicated the effects of a correcting comment from health organizations (*hORG*) but also verified the correction effect from individual users (*hIND*). Both *hORG* and *hIND* reduced perceived accuracy rating on fake news (**RQ1**). Moreover, we obtained that participants relied on reliable sources of correcting comments. Specifically, *hORG* valued "who wrote the comment" the most while *hIND* valued "URL" the most. We also obtained evidence of the frequency effect showing decreased perceived accuracy rating for the fake news at Phase 2 (**RQ2**). Meanwhile, we found both *hORG* and *hIND* were effective for the *low anxiety* group to reduce their perceived accuracy rating on fake news (**RQ3**).

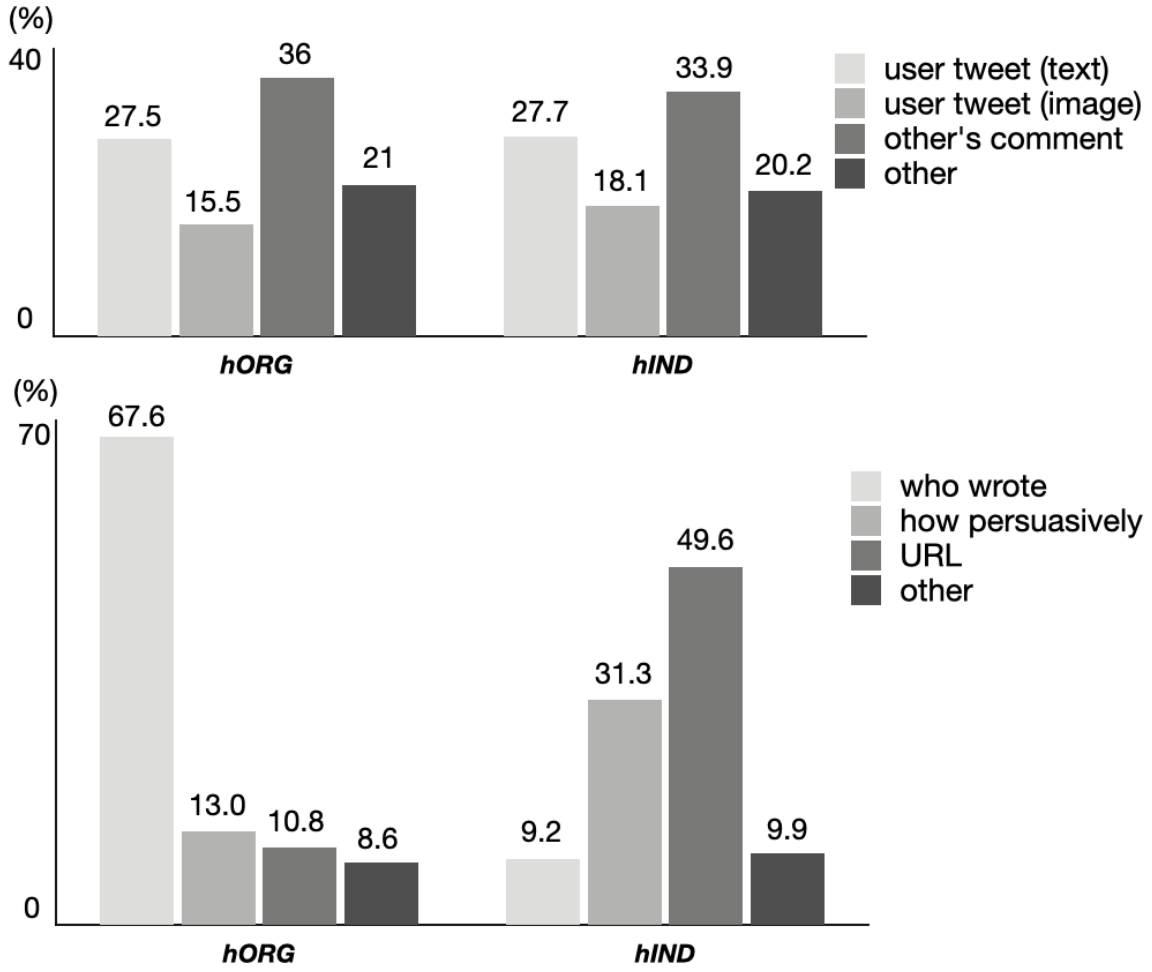

Figure 5.7: The top panel shows the response rate of the follow-up question in Experiment 2, asking the most influential factors in participants' perceived accuracy rating, and the bottom panel shows that of the most influential factors in the comment.

5.4.3 Experiment 3

We recruited 2,060 MTurk workers from November to December 2020. We accepted 1,166 participants' answers after removing one incomplete submission, four responses submitted out of the U.S., 406 who failed an attention check, 473 duplicated submissions, and ten responses submitted in less than three minutes (median completion time is about 10 minutes). The number of participants for each condition is as follows: 250 (*CON*), 236 (*hORG*), 227 (*hIND*), 214 (*fcORG*), and 239 (*fcIND*). The payment was the same as in Experiment 2. Participants' demographic information is as Table 5.1.

Perceived accuracy rating and willingness-to-share measure were entered into 5 (condition: *CON*, *hORG*, *hIND*, *fcORG*, *fcIND*) × 2 (veracity: *fake*, *real*) × 2 (frequency: *ONCE*,

=

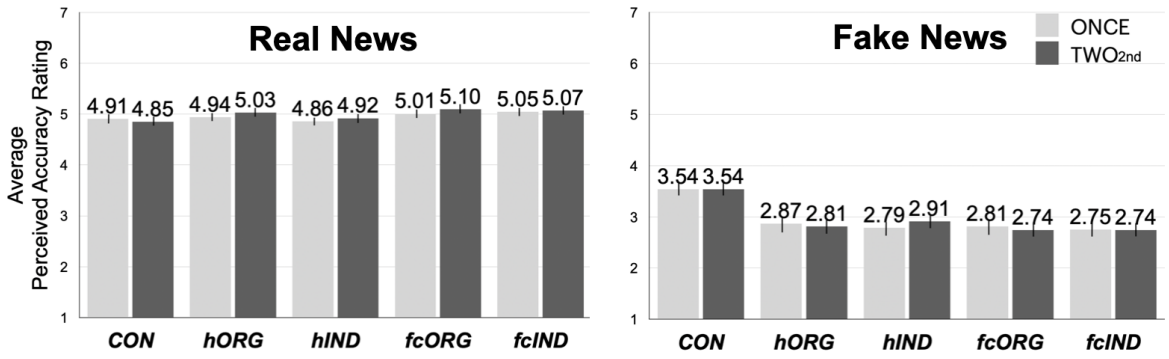

Figure 5.8: The average values of perceived accuracy ratings in Experiment 3 as a function of frequency × condition for real news (left panel) and fake news (right panel) with one standard error.

TWO_{2nd}) mixed ANOVAs with a significance level of .05, respectively. Post-hoc tests with Bonferroni correction were performed.

Perceived Accuracy Rating. Results of the perceived accuracy rating are shown in Figure 5.8. Same as the prior two experiments, participants can distinguish real news (4.97) from fake news (2.95), $F_{(1,1161)} = 1577.27$, $p < .001$, $\eta^2 = .576$. The main effect of condition was also significant, $F_{(4,1161)} = 3.76$, $p = .005$, $\eta_p^2 = .013$, and the difference among conditions was qualified by the effect of news veracity, $F_{(4,1161)} = 12.59$, $p < .001$, $\eta_p^2 = .042$. Same as Experiment 2, post-hoc tests revealed that the effect of each treatment condition was significant for the fake news compared to *CON*, $p_{adjs.} < .001$.

Sharing Decisions. Results of the willingness-to-share measure are presented in Figure 5.9. Participants showed more willingness to share real news (3.71) than fake news (2.66), $F_{(1,1161)} = 605.38$, $p < .001$, $\eta_p^2 = .343$. Also, the interaction of veracity × condition was significant, $F_{(4,1161)} = 4.32$, $p = .002$, $\eta_p^2 = .015$. In the post-hoc comparisons, only the gap between *fcIND* and *CON* for fake news was significant, $p_{adj.} = .033$.

Health Anxiety. We analyzed 1166 participants' results as in previous experiments. For perceived accuracy rating, participants with *high health anxiety* (4.32) gave a higher accuracy rating than those with *low health anxiety* (3.72) $F_{(1,1156)} = 84.42$, $p < .001$, $\eta_p^2 = .068$, and such pattern was more evident for fake news (*high health anxiety*: 3.45; *low health anxiety*: 2.63), $F_{(1,1164)} = 67.58$, $p < .001$, $\eta_p^2 = .051$ than for real news (*high health anxiety*: 5.20; *low health anxiety*: 4.82), $F_{(1,1164)} = 41.63$, $p < .001$, $\eta_p^2 = .035$.

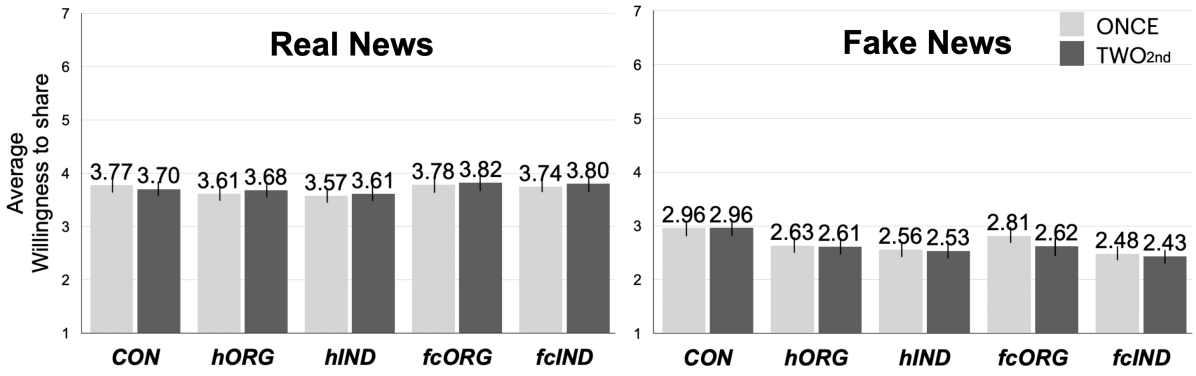

Figure 5.9: The average values of willingness-to-share in Experiment 3 as a function of frequency × condition for real news (left panel) and fake news (right panel) with one standard error.

For willingness-to-share, participants with *high health anxiety* (3.83) gave higher willingness-to-share score than those with *low health anxiety* (2.75), $F_{(1,1156)} = 124.33$, $p < .001$, $\eta_p^2 = .097$, similar to Experiments 1 and 2. Thus, highly anxious people about their health tended to share more health-related news regardless of news veracity.

Political Stance. We analyzed the effect of political stance as in Experiment 2 with 464 of liberals and 377 of conservatives after removing 325 moderates. The main effect of political stance, $Fs_{(1,831)} = 24.53$ and 17.95, $ps < .001$, $\eta_{ps}^2 = .029$ and .021, and its interaction with veracity, $Fs_{(1,831)} = 161.99$ and 80.38, $ps < .001$, $\eta_{ps}^2 = .163$ and .088, were significant for both perceived accuracy rating and willingness-to-share measure. Specifically, for perceived accuracy rating, the effect of political stance was more evident for the fake news (*liberals*: 2.58, *conservatives*: 3.73), $F_{(1,839)} = 86.07$, $p < .001$, $\eta_p^2 = .093$, than for the real news (*liberals*: 5.16, *conservatives*: 4.83), $F_{(1,839)} = 22.61$, $p < .001$, $\eta_p^2 = .026$. For the willingness-to-share measure, the effect of political stance was only significant for the fake news (*liberals*:2.37, *conservatives*: 3.33), $F_{(1,831)} = 48.54$, $p < .001$, $\eta_p^2 = .055$, but not the real news (*liberals*:3.75, *conservatives*: 3.82).

As in Experiment 2, conservatives were more susceptible to COVID-19 misinformation than liberals [193]. However, we again did not obtain the three-way interaction of political stance × veracity × condition for neither measures, $F_s < 1.0$, showing limited impacts of correction on mitigating conservatives' higher susceptibility to misinformation as Experiment 2.

Influential Factors. As in Experiment 2, we asked participants which factors were influential for their decision-making and which parts of comments influenced the most.

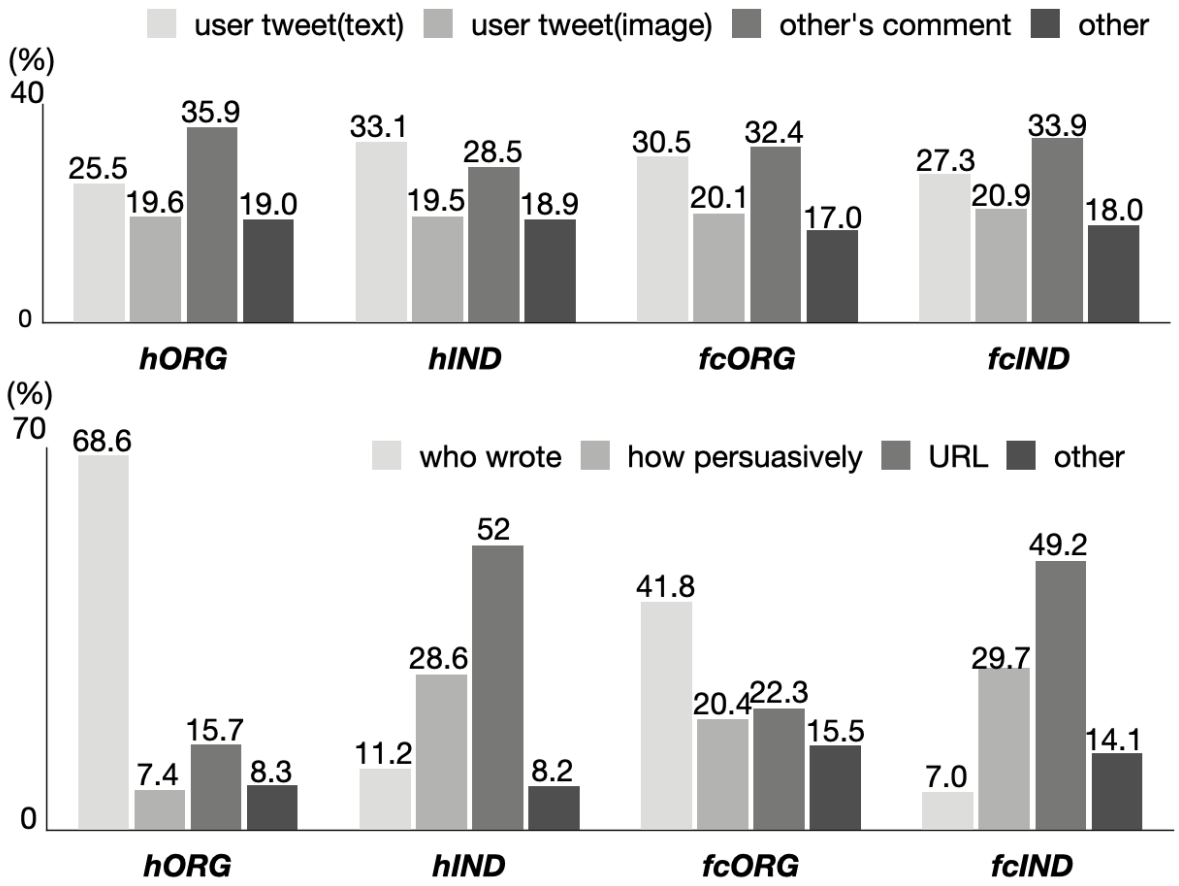

Figure 5.10: The top panel shows the response rate of the follow-up question in Experiment 3, asking the most influential factors in participants' perceived accuracy rating, and the bottom panel shows the most influential factors in the comment.

Across all treatment conditions, participants chose "other's comments" the most, except for those in the *hIND* condition (see Figure 5.10 top panel). For the following question asking the most influential part in the comments, participants in both organization conditions chose "who wrote" the most while those in both individual conditions chose "URL" the most, $\chi^2_{(9)} = 145.80$, $p < .001$ (see Figure 5.10 bottom panel). Overall, the obtained results were consistent with those found in Experiment 2.

Summary. The findings of Experiment 3 were consistent with the previous two experiments. All types of corrections were effective in reducing participants' perceived accuracy rating of fake news. We verified the effects of a correcting comment (**RQ1**) from fact-checking websites (*fcORG*) as well as the one from health organizations (*hORG*). Also, we found the effects of individual users' correction, which has a reference link from either fact-checking websites (*fcIND*) or health organizations (*hIND*). Moreover,

we discovered that participants counted on the reliable source of a correcting comment. Specifically, participants in the *hORG* and *fcORG* weighed "who wrote the comment" the most, while those in the *hIND* and *fcIND* counted on "URL" the most. As in Experiment 1, we did not obtain the frequency effect (**RQ2**). Meanwhile, we found the minimal impacts of health anxiety or political stance on the correction effect (**RQ3**).

5.5 General Discussion

In the current study, we investigated if the correction from organization users or individual users can reduce participants' susceptibility to COVID-19 fake news (**RQ1**). We also examined whether more frequent correction can further reduce the susceptibility (**RQ2**), and whether individuals' health anxiety level has an impact on the effect of correction (**RQ3**). Across the three online experiments with $2,841$ participants, we examined the correction effects of three types of users on social media. We verified the effect of user-initiated correction in general, with the fact that participants counted on the reliability of correction. We also found that participants with high health anxiety were more susceptible to COVID-19 fake news than those with low health anxiety in all experiments.

5.5.1 Effect of Correction from a Single User

Previous work obtained the effect of user-initiated correction on social media by conducting a survey [25] or analyzing Twitter data [194]. Also, other studies [26, 93] showed the effect of correcting comments in social media contexts by conducting experiments. Our study extended those works by demonstrating the consistent effects of user-initiated correction with reliable sources in reducing perceived accuracy rating on COVID-19 fake news in a social media context.

We corroborated that people's perceived accuracy rating on fake news could be reduced by a single correction comment by health organizations, fact-checking websites, or individual users. In particular, the correction effect was similar across different types of users. Critically, we unearthed that participants depended on the reliability of sources in the correction to decide their perceived accuracy rating. Participants in *hORG* and *fcORG* chose "who wrote the comment" the most, while those in *hIND* and *fcIND* chose "whether the comment included a reference URL" the most. The only difference between *ORG* and *IND* was whether the correction was directly delivered by reliable users or

indirectly delivered through reliable URLs. Thus, our findings on the indirect effect of a reliable source contribute to the literature about the source effect [26, 80].

Although the *hIND* was effective on correction in Experiments 2 and 3, it did not show a significant difference compared to the no correction condition in Experiment 1. One possible explanation for the difference might be related to users' increased knowledge about COVID-19 news [25], and consequently increased reliance on other individual users beyond health organizations to gain more information. The above reason may also explain the non-significant results of Vraga and Bode's work (2017) since they implemented a relatively unfamiliar topic to the participants in their experiment. Moreover, across the three experiments, we found that participants' perceived accuracy rating on real news in the treatment conditions was increased or similar to that in the control conditions, indicating limited side effects of user-initiated correction compared to platform-driven correction (e.g., fact-checking warnings) [122].

In all experiments, participants were more conservative in sharing news than the perceived accuracy rating, which may contribute to the minimal correction effect on misinformation sharing. Various motivations such as information-seeking, socializing, status-seeking, or prior social media sharing experience [195] could drive people's intention of news sharing on social media regardless of correction. Future studies could investigate effective correction to prevent misinformation sharing, considering such motivations of sharing behaviors.

5.5.2 Frequency Effect on Correction

Throughout the experiments, the frequency effect on correction was only evident in Experiment 2. Results of Experiments 1 and 3 were consistent with the correction effect but did not show statistical significance: perceived accuracy rating and willingness-to-share measures were numerically smaller for the second correction than the first correction. Those results may be due to the use of the same comment messages across phases since people typically expect varied comments from different social media users. Future work could consider varying correction messages to understand further the frequency effect of correction.

5.5.3 Correction Effect Depending on Health Anxiety

We discovered that participants with high health anxiety tended to believe and share more news regardless of news veracity than those with low health anxiety, indicating

that the highly anxious people might seek reassurance through health information [176]. In particular, we verified that *hORG* and *hIND* were only effective for the low health anxiety group to reduce their perceived accuracy rating on fake news in Experiment 2. To the best of our knowledge, our study was the first to deal with the impact of health anxiety on correction for COVID-19-related fake news. Future studies should contrive correction methods, especially for people with high health anxiety, to mitigate their susceptibility to COVID-19 misinformation in particular and fake health news in general.

5.5.4 Correction Effect Depending on Political Stance

Experiments 2 and 3 revealed that conservatives were more susceptible to COVID-19 fake news than liberals. Such results are in agreement with a recent study showing stronger beliefs in COVID-19 fake news by conservatives [193]. In both of our experiments, corrections showed minimal impacts on helping conservatives. Considering the impacts of political stances on various misinformation literature [196–198], we believe that further investigation on effective correction methods for more vulnerable populations (e.g., conservatives) is essential.

5.5.5 Limitations and Future Work

We discuss a few limitations that could be addressed in future studies. First, we chose MTurk for recruitment to gain a reasonably large sample size as previous misinformation studies did [79, 122]. MTurk workers are more demographically diverse than the college students [171, 185]. However, the MTurk population, in general, cannot fully represent the whole population. For instance, most MTurk workers tend to be in their 30s [172]. Therefore, a more comprehensive recruiting method could be used to generalize our findings to other sample in future studies. In addition, in terms of materials, we used a more recent news set in Experiments 2 and 3 since COVID-19 news has been quickly updated and diversified. This change seemed to lead to different average gaps of perceived accuracy ratings between real and fake news among experiments: Exp.1 (2.59), Exp.2 (2.20), Exp3 (2.02). Furthermore, it should be recognized that the effects found in our experiments may not appear in practice due to exclusion of other factors on social media (e.g., multiple replies and social relationships among users, etc.). Moreover, we are aware that participants could not pay attention to the correction in reality because of many distracting factors on social media, such as other postings and interactions with other users in real-time. Also, we did not measure participants' prior beliefs in the fake news.

Therefore, we can not rule out the possibility of having negative effects from correction (e.g., backfire effects) [31, 70]. Future works could develop experimental designs with more ecological validity and evaluate the generalizability of our findings.

5.6 Conclusion

In this work, we carried out three online experiments with a more systematic design to comprehend the impact of a single correction comment on mitigating users' fake news susceptibility on social media. In total, three types of users were investigated across the experiments. We verified the correction effects on reducing the user's perceived accuracy ratings from individual users, health organizations, and fact-checking websites. Moreover, our study revealed that participants counted on the reliability of correction sources for their decision-making. We also found that high health anxiety people could be more susceptible to COVID-19 misinformation. Additionally, our results showed that conservatives are more susceptible to COVID-19 fake news than liberals. In conclusion, our findings highlight 1) the importance of encouraging social media users to leave correcting comments on fake news, as long as they have reliable sources, and 2) the necessity to develop effective correction methods considering individual differences (e.g., health anxiety level and political stance).

Chapter 6
Conclusion

6.1 Summary

The rampant spread of misinformation on social media platforms often increases uncertainty for users to distinguish trustworthy information. This problem becomes especially noticeable during major events like presidential elections and pandemics, where the amount of information on social media is overwhelming. For the average user, rapidly ascertaining the veracity of this abundance of information is a daunting task.

Numerous algorithmic detection methods have been developed to combat false information, and social media platforms have invested in deploying these techniques to identify fake news. However, the most crucial step is empowering users to effectively differentiate and subsequently avoid the proliferation of this detected misinformation. Misinformation correction involves strategically presenting information that highlights inaccuracies in the content, thereby alerting users to its falsehood.

Our research endeavors were dedicated to uncovering potent misinformation correction methods, with a particular emphasis on the domain of social media platforms. These platforms have emerged as pivotal players in the misinformation arena, notably in the context of proliferating fake news. In this context, we conducted experimental research to find effective methods for correcting misinformation considering the social media context. Our primary goal was to unveil and appraise the efficacy of diverse misinformation correction methods, each designed to enhance users' awareness of counterfeit news. The experiments were conducted from two perspectives based on the agent of correction. One is platform-driven correction, and the other is user-initiated correction.

Through three research projects, we aimed to address the following issues. For platform-driven correction, considering the increasing development and utilization of numerous fake news detectors, we took the initiative to move beyond the conventional

human fact-checking warnings. We introduced machine-learning warnings and examined whether machine-learning warnings can effectively perform misinformation correction. Following the verification of the potential of machine learning warnings with explanations. In our follow-up study, we conducted a more in-depth investigation into the effectiveness of warnings with explanations, examining how the effectiveness varies based on the manner of explanation. For the user-initiated correction study, we focused on user comments as the most direct means of responding to fake news postings. We evaluated the effectiveness of misinformation correction in correcting comments written by various users.

We conducted online experiments using participants from a crowdsourcing platform like Amazon Mechanical Turk. These experiments yielded noteworthy results, indicating the effectiveness of both platform-driven and user-initiated correction methods in improving users' capacity to discern and identify misinformation. The findings underscore the potential of these strategies in bolstering users' abilities to navigate the intricate landscape of online information and make informed judgments.

In the initial study [80] of platform-driven correction, we devised three types of machine-learning warnings and undertook a comprehensive investigation into their impact on enhancing users' ability to detect misinformation across two experiments. Our findings indicated the potential of machine-learning warnings accompanied by explanations. This warning significantly improved participants' capacity to identify fake news, especially in cases where the news source was not disclosed, irrespective of their level of trust in the warning.

Building upon this, our subsequent study [199] delved further into the realm of machine learning warnings with explanations. Notably, we discovered that a negatively framed explanation proved to be more effective in rectifying misinformation when the reliability of an AI system was uncertain. Conversely, warning only without explanations emerged as a more effective correction method when users know the reliability information of an AI system. These outcomes emphasize the pivotal role of transparency within machine-learning warnings as well as the importance of providing reliability information about AI systems.

In the domain of user-initiated correction, our research [200] journey encompassed a comprehensive series of three experiments, each meticulously designed to assess the effectiveness of distinct types of user-initiated correction. Our experiments examined various facets of user correction within a simulated Twitter environment. The experiments involved manipulating profile types and news source links across different conditions.

The ensuing findings substantiated a noteworthy conclusion: the mere presence of a single correction comment had a discernible and positive impact on participants' ability to identify fake news. Importantly, this effect held true regardless of the commenter's identity or the specific source link utilized if a user can see a reliable source.

6.2 Contributions

In this dissertation, I have contributed to the informatics field by investigating misinformation correction. The following paragraphs outline this research's key contributions, including novel insights, methodological advancements, and practical implications.

A systematic approach to agent-centered misinformation correction. In pursuing effective misinformation correction methods, I initially sought to systematically structure approaches to misinformation correction by categorizing them based on the correction agent. Considering the dynamics of social media, I took into account both platform and user perspectives, resulting in the classification of platform-driven correction and user-initiated correction.

Classifying correction agents allowed us to navigate the complex landscape of social media misinformation more effectively. By distinguishing between platform-driven correction, initiated and executed by the platform itself, and user-initiated correction, carried out by the active participants, I endeavored to find a nuanced understanding of the multifaceted challenge we faced. This distinction, in essence, served as a compass to guide our research.

Ultimately, this research yielded a valuable outcome in discovering effective methods for misinformation correction in both platform-driven and user-initiated correction scenarios. By understanding the distinct roles played by these correction agents, I could unveil a blueprint for combating misinformation on social media.

An extension of platform-driven warnings that reflect the contemporary context. For platform-driven correction, I introduced machine-learning warning messages designed to reflect the detection capabilities of machine learning or AI systems. These messages distinguished themselves from conventional human fact-checker-centered warnings.

Furthermore, providing explanations regarding the judgments made by AI systems in fake news warnings was pioneering at the time of this study. Within the relatively

limited research landscape on misinformation warnings, the focus was predominantly on the established warning framing used by platforms like Facebook. In contrast, my study expanded the realm of warning research by incorporating the perspective of machine learning algorithms responsible for fake news detection. Our aim was to enrich the field of warning research by providing logical insights into the detection process, aligning with the broader trend towards explainable AI. Our research affected several other studies [129, 134, 201, 202] which have explored the effectiveness of explanatory mechanisms within AI system warnings, further attesting to the increasing significance of explanation in a warning.

Moreover, in what I consider pioneering efforts, we explored the impact of framing in explanations provided by machine learning warnings on the rationale behind detection judgments. Ultimately, in situations where there is uncertainty about the reliability of AI system detection, our research confirmed the effectiveness of using a negative framing in the explanations provided by machine learning warnings.

Emphasizing the Expected Role of Social Media Users in Misinformation Correction. Experimental studies about the effectiveness of user-initiated correction have been relatively scarce compared to platform-driven correction. Recognizing that users are central to shaping information activities on social media, I have researched misinformation correction by users. I differentiated users into organizational users and individual users and explored correction from each user type. For the case of organizational users, I explored two types of organizational user profiles: health organization users and fact-checking website users. Consequently, I could compare four types of users' correcting comments.

Ultimately, my research revealed that incorporating information from a 'reliable source' consistently resulted in a substantial decrease in the perceived accuracy rating of fake news, regardless of who corrected the misinformation. This discovery serves as a fundamental incentive for encouraging users to engage proactively in misinformation correction, highlighting the pivotal role of their active participation.

Methodological Efforts to Strengthen Experimental Robustness. I primarily centered my research on developing effective misinformation correction methods within the extensive landscape of social media, achieved via a comprehensive series of experimental studies in which I aimed to enhance methodological robustness by rigorously collecting, analyzing, and interpreting data.

First, the stimuli employed in this research were thoughtfully designed to mirror authentic scenarios in which misinformation prominently circulates across social media platforms. This approach enabled the practical assessment and applicability of correction strategies in addressing real-world challenges. Each experimental study incorporated a diverse array of stimuli, including political and COVID-19 news.

We recruited a significant number of participants for online experiments (Study 1: $1,176$, Study 2: $2,841$, Study 3: $2,692$) to strengthen the credibility of our findings, providing us with a robust and diverse sample and several distinct advantages. This approach bolstered the statistical power of our studies, allowing us to detect both subtle and substantial effects, while the larger sample size led to more precise estimates with reduced margins of error and enhanced generalizability of our conclusions. Our commitment to utilizing data from a substantial number of participants across all studies reinforced reliability and inspired confidence in the strength and accuracy of our findings.

Moreover, our research embraced a series of deliberate iterations and variations in the experimental design, a rigorous endeavor undertaken to achieve increasingly refined and precise results. This dedication to refinement not only strengthened the scientific rigor of our investigations but also ensured that our findings were robust, reliable, and applicable across a range of scenarios. Through these methodological refinements and adjustments, we enhanced the depth of our insights and the validity of our conclusions, ultimately contributing to the overall quality of our research.

We aimed to minimize the influence of news content on users by presenting them with a diverse selection of news content, reducing potential variations in user responses linked to different content. This approach offers several advantages. It reinforces the robustness of our results, ensuring that observed effects persist across different stimuli by reducing potential bias associated with certain stimuli. Additionally, employing multiple stimuli provides a more comprehensive understanding of the topic and uncovers deeper insights into underlying mechanisms. Also, it minimizes the impact of confounding factors, increasing the validity and reliability of our research.

Furthermore, we employed randomization as a fundamental practice in stimulus sequencing to minimize potential biases in the outcomes that may arise from the order in which stimuli are presented. In addition to this, we further heightened the control over our experimental design by creating multiple subsets based on the Latin square design. This methodical approach guaranteed that our research outcomes were not unduly affected by the sequence in which stimuli were encountered, strengthening the internal validity of our study.

6.3 Limitations

In the pursuit of exploring effective methods for platform-driven correction and user-initiated correction on social media, with a specific emphasis on the agent of correction, this research conducted a total of 11 experiments across three studies. While stringent measures were in place to ensure the experimental rigor and validity of these endeavors, it is important to acknowledge several limitations that pave the way for future research improvements.

In our research, participants for all experiments were recruited through Amazon Mechanical Turk, a common approach in many online studies to secure a large and diverse participant pool. It is important to note that controlling data quality from Amazon Mechanical Turk workers can be challenging despite diligent data cleaning efforts. This challenge stems from the possibility of including data from participants with malicious intentions, such as the use of bots [203]. While replicating the same scale of participants in laboratory experiments may not be feasible, conducting experiments with a reasonable number of participants in a controlled lab environment offers the advantage of a more focused setting. This controlled environment can allow for a more accurate assessment of potential differences in data quality. Additionally, it may be worthwhile to explore recent, high-quality crowd-sourcing platforms (e.g., Prolific) [204] as an alternative option.

Secondly, it is important to acknowledge that, by design, our experiments excluded several real-world factors. The stimuli used in our experiments primarily consisted of news content paired with correction messages. In the platform-driven correction study, we focused on news content and warning messages, while the user-initiated correction study included a single correcting comment alongside the news content. These stimuli were carefully crafted to simulate correction scenarios within the context of social media platforms, such as Facebook and Twitter. However, due to the fixed image format in which the stimuli were presented, we had to omit the nuanced and interactive dynamics that are characteristic of real-world social media interactions.

Furthermore, the absence of reaction buttons, such as ‘like’ or ‘heart,’ within the stimuli limited our ability to explore essential elements for assessing content credibility and the dynamics of correction in a real social media context. As a result, the study did not fully capture the diverse spectrum of content interaction that is characteristic

of social media platforms. To address this limitation, future research could consider incrementally introducing these elements as separate factors, potentially leading to the identification of more realistic and effective correction methods.

In the meantime, our primary objective was to identify distinct correction effects related to different correction types through comparative analysis. As a result, the examination of effects based on participants' characteristics was primarily conducted in post-session, using a limited set of follow-up questions. These characteristics included inquiries about political orientation, social media usage habits, health anxiety, and more. Among these, it was suggested that political orientation and health anxiety lead to different susceptibilities to political and health news corrections. In addition to these characteristics, individuals may inherently differ in their attitudes and receptiveness to warnings. Multiple factors, including prior experiences with accepting warnings, existing knowledge about specific news topics, mental models, literacy [205, 206], and more, could be explored to determine their significant impact on how information is received and corrected. In future research, considering a wider range of demographic characteristics as variables in the analysis may lead to obtaining more diverse insights.

Additionally, I acknowledge that I used a restricted set of news content within a limited timeframe. I focused on political news and COVID-19 news, which were of utmost societal importance during the respective study periods. It is crucial to recognize that correction effectiveness may vary depending on the type and characteristics of the topics, and this variability should not be overlooked. While addressing urgent issues is necessary, future research aimed at establishing the generalizability of correction effects could benefit from conducting experiments covering a more extensive range of news topics.

In addition, a between-subject design may introduce some variability related to participant characteristics. I consistently employed a between-subject design in all experiments to assess the effectiveness of various misinformation correction types. This approach was selected due to several advantages, such as reducing time demands on participants, facilitating comparisons across multiple conditions, and simplifying the interpretation of diverse within-subject conditions. In future research, as part of supplementary investigations, the utilization of a within-subject design could be considered to compare different conditions within the same individuals more effectively. This alternative approach could provide additional insights and complement the findings obtained from the between-subject design.

My research has primarily concentrated on investigating practical misinformation

correction strategies within well-controlled experimental settings. However, it is crucial to stress the importance of further validation to evaluate their real-world effectiveness. This need for validation stems from an awareness that the real-world context introduces numerous uncontrolled variables and complexities that can potentially impact the outcomes.

6.4 Implications

Having presented and analyzed the results of this study, it is now essential to consider the broader implications of our findings and their potential applications in various contexts. In this section, I will discuss how the insights gained from our research can inform and impact the field of misinformation correction studies.

First, I have introduced specific methods for effective misinformation correction tailored to real-world social media scenarios. In the context of platform-driven correction, social media platforms like Facebook and X (previously known as "Twitter") could contemplate the adoption of machine learning warnings. Although these platforms continually advance their fake news detection models, they have yet to incorporate warnings generated by machine learning or AI systems. This reluctance may stem from concerns about unfamiliar terminologies from a layperson's perspective. It could also be attributed to the platforms' need to establish the widespread reliability of their detection sources.

However, if subsequent research corroborates the efficacy of machine-learning warnings, social media platforms might consider introducing them as a complementary approach alongside fact-checking warnings. Specifically, our study's proposed bar chart-style explanations could serve as a visual means to elucidate the rationale behind machine learning warnings, potentially enhancing their credibility. The principles underpinning such explanations may also inspire future developments in user-initiated correction methods beyond platform-driven correction. Furthermore, the disclosure of detection model reliability scores has the potential to incentivize users to embrace warnings proactively.

In the user-initiated correction study, we highlighted the effectiveness of incorporating a reliable source in a correcting comment, regardless of the initiator of the correction. That is, by adding the URL of the news containing the correcting information, even regular users can play an active role in the misinformation correction process. Consequently, it becomes imperative for fact-checking non-governmental organizations (NGOs) and social

media platforms to incentivize individual users to partake in misinformation correction by including reliable sources within their correction comments.

Given the nature of social media, where individual influence often carries more weight in shaping the acceptance of information within their immediate circles of family and friends [207], it is evident that a top-down, platform-driven approach to misinformation correction alone may not suffice. Instead, each individual can emerge as a central node, fostering a culture of misinformation correction within their personal networks. This collaborative effort between fact-checking entities and individual users can significantly contribute to more informed and reliable social media.

Next, the insights derived from my research can extend beyond the realm of misinformation correction on social media platforms. For instance, our correction insights may prove essential when utilizing recently developed, widely recognized services based on large language models, such as ChatGPT. Notwithstanding their remarkable capabilities, their models may occasionally yield biased or incomplete responses due to their training data [208, 209]. For instance, when models might inadvertently incorporate inaccurate information, offering reliability scores for their responses or more transparent explanations for their decision-making processes could empower users to make more informed choices when engaging with the provided information. This can ultimately contribute to a more reliable and user-centric experience in using such a large language model-based service.

Besides, the rapid advancement of deepfake technology has spurred the widespread dissemination of manipulated images, often harnessed to propagate false information effectively. Deepfake images are frequently employed to impersonate experts across diverse domains, deceitfully garnering people's trust for deceptive purposes [210]. This underscores the growing significance of correcting deepfake content to mitigate the risk of users falling victim to false information. Finding specialized misinformation correction techniques for deepfake content could contribute to mitigating the evolving forms of misinformation. For instance, the visualization of the rationale behind identifying deepfakes or the inclusion of reliable sources are avenues that merit consideration. Such enhancements could substantially contribute to the ongoing battle against the spread of deceptive information facilitated by deepfake technologies.

6.5 Future Directions

As I wrap up this dissertation, I want to look ahead and identify potential avenues for future research. Several intriguing directions emerge from our works that can further

strengthen our understanding and strategies for misinformation correction in the digital landscape.

Misinformation Correction from an AI System with Different Reliability. Our study [199] unearthed intriguing insights into how participants interact with fake news warning messages and utilize reliable information provided by AI systems. In particular, upon exposure to fake news warning messages with explanations, participants displayed a notable inclination to lower their perceived accuracy rating for real news if they could not be sure about the reliability of the AI system. This suggests a heightened awareness of the potential presence of misinformation (i.e., fear of miss errors). In practical terms, it is unfeasible for an AI system to comprehensively censor and label all news as either fake or real, making it challenging to alleviate the concerns about miss errors. Moreover, in real-world scenarios, besides the presence of miss errors (false negatives), false alarms (false positives) add to the complexity.

Hence, a natural avenue for research arises: if advance notice is given about the potential occurrence of such instances and information about the reliability of the AI system is provided, could this mitigate users' vague suspicions about the system's trustworthiness? Upon the research question, I propose a more in-depth exploration of warning effects within AI systems that vary in reliability. While prior research has diligently examined the influence of warning signs, there exists a notable gap in understanding the limitations or inadvertent consequences of warning systems that mimic real-world scenarios. I aim to explore the effects of warnings in AI systems characterized by different reliability levels, leading to four distinct warning-labeling scenarios: true positives, true negatives, false positives, and false negatives.

This endeavor serves to uncover whether participants are susceptible to both false positives and false negatives and, subsequently, to consider implementing a more nuanced and adaptable warning system or boosting detection algorithms. Through this research, I aspire to offer a comprehensive understanding of the implications of AI systems in the realm of fake news detection, thereby contributing to a more effective and reliable approach to countering misinformation.

User-initiated correction using machine detectors. We did not specifically investigate user-initiated correction using machine detectors during our research endeavors. This omission was rooted in the prevailing assumption that typical users primarily rely on well-established human fact-checking websites as their primary sources for corrections.

However, the landscape is evolving, and there is a growing interest among the public in the field of machine learning, coupled with significant advancements in machine-learning fake news detectors.

Consequently, if users can access information about lists of fake news identified by machine detectors, there is a possibility that users may gradually consider these sources for correction purposes. If users are inclined to utilize results from machine detectors, delving into experimental studies to pinpoint the most effective methods for users to communicate the machine detector's findings would be a worthwhile avenue to explore. This research aims to uncover fresh perspectives and insights not explored in traditional misinformation correction studies. By actively investigating user-centered misinformation correction that embraces the evolving results of machine detectors, I anticipate revealing novel viewpoints and discoveries in this field.

The integrated comparative study of platform-driven correction and user-initiated correction methods. My research has been characterized by a comprehensive examination of platform-driven correction and user-initiated correction methods, ultimately leading to the delineation of effective strategies for combating misinformation within each of these correction approaches. The ideal scenario envisions both platforms and users actively participating in the correction process to counter the dissemination of misinformation. However, whose correction methods can be more effective has not been revealed. Future research endeavors could benefit from an integrated comparative study to further inform the development of misinformation correction strategies and gain a more subtle understanding of the dynamics. I aim to assess the effectiveness of platform-driven correction versus user-initiated correction across various scenarios. This exploration will shed light on how these correction methods can synergistically collaborate in different contexts.

Bibliography

[1] Allcott, H. and M. Gentzkow (2017) "Social media and fake news in the 2016 election," *J. of Economic Perspectives*, **31**(2), pp. 211–36.

[2] Armitage, R. and C. Vaccari (2021) "Misinformation and disinformation," in *The Routledge companion to media disinformation and populism*, Routledge, pp. 38–48.

[3] Lazer, D. M., M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. (2018) "The science of fake news," *Science*, **359**(6380), pp. 1094–1096.

[4] Posetti, J. and A. Matthews (2018) "A short guide to the history of 'fake news' and disinformation," *International Center for Journalists*, **7**(2018), pp. 2018–07.

[5] Lewandowsky, S., U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook (2012) "Misinformation and its correction: Continued influence and successful debiasing," *Psychological Science in the Public Interest*, **13**(3), pp. 106–131.

[6] Mayfield, A. (2008) "What is social media," *Joint Force Quarterly*, **60**(1).

[7] Kaplan, A. M. and M. Haenlein (2010) "Users of the world, unite! The challenges and opportunities of Social Media," *Business horizons*, **53**(1), pp. 59–68.

[8] Li, J. and X. Chang (2023) "Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media," *Information systems frontiers*, **25**(4), pp. 1479–1493.

[9] Whitehead, H. S., C. E. French, D. M. Caldwell, L. Letley, and S. Mounier-Jack (2023) "A systematic review of communication interventions for countering vaccine misinformation," *Vaccine*.

[10] Chan, M.-p. S. and D. Albarracín (2023) "A meta-analysis of correction effects in science-relevant misinformation," *Nature Human Behaviour*, pp. 1–12.

[11] Broniatowski, D. A., J. R. Simons, J. Gu, A. M. Jamison, and L. C. Abroms (2023) "The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic," *Science Advances*, **9**(37), p. eadh2132.

[12] Koch, T. K., L. Frischlich, and E. Lermer (2023) "Effects of fact-checking warning labels and social endorsement cues on climate change fake news credibility and engagement on social media," *Journal of Applied Social Psychology*.

[13] Conroy, N. J., V. L. Rubin, and Y. Chen (2015) "Automatic deception detection: Methods for finding fake news," in *78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, vol. 52, pp. 1–4.

[14] Tacchini, E., G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro (2017) "Some like it hoax: Automated fake news detection in social networks," *arXiv preprint arXiv:1704.07506*.

[15] Shu, K., S. Wang, and H. Liu (2018) "Understanding user profiles on social media for fake news detection," in *IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR)*, pp. 430–435.

[16] Walter, N. and S. T. Murphy (2018) "How to unring the bell: A meta-analytic approach to correction of misinformation," *Communication Monographs*, **85**(3), pp. 423–441.

[17] Blank, H. and C. Launay (2014) "How to protect eyewitness memory against the misinformation effect: A meta-analysis of post-warning studies," *Journal of Applied Research in Memory and Cognition*, **3**(2), pp. 77–88.

[18] Ecker, U. K., S. Lewandowsky, and D. T. Tang (2010) "Explicit warnings reduce but do not eliminate the continued influence of misinformation," *Memory & cognition*, **38**, pp. 1087–1100.

[19] Taprial, V. and P. Kanwar (2012) *Understanding social media*, Bookboon.

[20] Pennycook, G. and D. G. Rand (2018) "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*.

[21] Clayton, K., S. Blair, J. A. Busam, and et al. (2019) "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*, pp. 1–23.

[22] Adadi, A. and M. Berrada (2018) "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, **6**, pp. 52138–52160.

[23] Gunning, D. and D. Aha (2019) "DARPA's explainable artificial intelligence (XAI) program," *AI Magazine*, **40**(2), pp. 44–58.

[24] Tversky, A. and D. Kahneman (1981) "The framing of decisions and the psychology of choice," *Science*, **211**(4481), pp. 453–458.

[25] Bode, L. and E. K. Vraga (2021) "Correction Experiences on Social Media During COVID-19," *Social Media + Society*, **7**(2), https://doi.org/10.1177/20563051211008829.

[26] Vraga, E. K. and L. Bode (2017) "Using expert sources to correct health misinformation in social media," *Science Communication*, **39**(5), pp. 621–645.

[27] Bode, L. and E. 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), pp. 619–638.

[28] Shu, K., A. Sliva, S. Wang, J. Tang, and H. Liu (2017) "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, **19**(1), pp. 22–36.

[29] Cook, J., U. Ecker, and S. Lewandowsky (2015) "Misinformation and how to correct it," in *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource*, Wiley Online Library, pp. 1–17.

[30] Ha, L., L. Andreu Perez, and R. Ray (2021) "Mapping recent development in scholarship on fake news and misinformation, 2008 to 2017: Disciplinary contribution, topics, and impact," *American Behavioral Scientist*, **65**(2), pp. 290–315.

[31] Nyhan, B. and J. Reifler (2010) "When corrections fail: The persistence of political misperceptions," *Political Behavior*, **32**(2), pp. 303–330.

[32] Wu, L., F. Morstatter, K. M. Carley, and H. Liu (2019) "Misinformation in social media: definition, manipulation, and detection," *ACM SIGKDD explorations newsletter*, **21**(2), pp. 80–90.

[33] Karlova, N. A. and K. E. Fisher (2013) "A social diffusion model of misinformation and disinformation for understanding human information behaviour," **18**(1).

[34] Van der Linden, S., A. Leiserowitz, S. Rosenthal, and E. Maibach (2017) "Inoculating the public against misinformation about climate change," *Global Challenges*, **1**(2).

[35] Treen, K. M. d., H. T. Williams, and S. J. O'Neill (2020) "Online misinformation about climate change," *Wiley Interdisciplinary Reviews: Climate Change*, **11**(5), p. e665.

[36] Altay, S., M. Berriche, H. Heuer, J. Farkas, and S. Rathje (2023) "A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field," *Harvard Kennedy School Misinformation Review*, **4**(4), pp. 1–34.

[37] Molina, M. D., S. S. Sundar, T. Le, and D. Lee (2021) ""Fake news" is not simply false information: A concept explication and taxonomy of online content," *American Behavioral Scientist*, **65**(2), pp. 180–212.

[38] Zannettou, S., M. Sirivianos, J. Blackburn, and N. Kourtellis (2019) "The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans," *Journal of Data and Information Quality (JDIQ)*, **11**(3), pp. 1–37.

[39] Prooijen, J.-W. (2018) *The psychology of conspiracy theories*, Routledge.

[40] DiFonzo, N. and P. Bordia (2007) *Rumor psychology: Social and organizational approaches.*, American Psychological Association.

[41] Chen, Y., N. J. Conroy, and V. L. Rubin (2015) "Misleading online content: recognizing clickbait as" false news"," in *Proceedings of the 2015 ACM on workshop on multimodal deception detection*, pp. 15–19.

[42] Westerlund, M. (2019) "The emergence of deepfake technology: A review," *Technology innovation management review*, **9**(11).

[43] Lewandowsky, S. and S. Van Der Linden (2021) "Countering misinformation and fake news through inoculation and prebunking," *European Review of Social Psychology*, **32**(2), pp. 348–384.

[44] Scheufele, D. A. and N. M. Krause (2019) "Science audiences, misinformation, and fake news," *Proceedings of the National Academy of Sciences*, **116**(16), pp. 7662–7669.

[45] Tandoc Jr, E. C., Z. W. Lim, and R. Ling (2018) "Defining "fake news" A typology of scholarly definitions," *Digital journalism*, **6**(2), pp. 137–153.

[46] Ha, L., L. Andreu Perez, and R. Ray (2019) "Mapping recent development in scholarship on fake news and misinformation, 2008 to 2017: Disciplinary contribution, topics, and impact," *American Behavioral Scientist*, p. 0002764219869402.

[47] Quandt, T., L. Frischlich, S. Boberg, and T. Schatto-Eckrodt (2019) "Fake news," *The International Encyclopedia of Journalism Studies*, pp. 1–6.

[48] Zhang, X. and A. A. Ghorbani (2020) "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, **57**(2), p. 102025.

[49] Silverman, C. (2016), "This Analysis Shows How Fake Election News Stories Outperformed Real News On Facebook." .
URL `https://bit.ly/2KYA8Wq`

[50] Potthast, M., J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein (2017) "A stylometric inquiry into hyperpartisan and fake news," *arXiv preprint arXiv:1702.05638*.

[51] Oldenbourg, A. (2022) "Digital freedom and corporate power in social media," *Critical Review of International Social and Political Philosophy*, pp. 1–22.

[52] Bermes, A. (2021) "Information overload and fake news sharing: A transactional stress perspective exploring the mitigating role of consumers' resilience during COVID-19," *Journal of Retailing and Consumer Services*, **61**, p. 102555.

[53] Cinelli, M., G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini (2021) "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, **118**(9), p. e2023301118.

[54] Seargeant, P. and C. Tagg (2019) "Social media and the future of open debate: A user-oriented approach to Facebook's filter bubble conundrum," *Discourse, Context & Media*, **27**, pp. 41–48.

[55] Zimmer, F., K. Scheibe, M. Stock, and W. G. Stock (2019) "Echo chambers and filter bubbles of fake news in social media. Man-made or produced by algorithms," in *8th annual arts, humanities, social sciences & education conference*, pp. 1–22.

[56] Dou, Y., K. Shu, C. Xia, P. S. Yu, and L. Sun (2021) "User preference-aware fake news detection," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2051–2055.

[57] Wardle, C. and H. Derakhshan (2017) *Information disorder: Toward an interdisciplinary framework for research and policymaking*, vol. 27, Council of Europe Strasbourg.

[58] Islam, M. S., T. Sarkar, S. H. Khan, A.-H. M. Kamal, S. M. Hasan, A. Kabir, D. Yeasmin, M. A. Islam, K. I. A. Chowdhury, K. S. Anwar, et al. (2020) "COVID-19–related infodemic and its impact on public health: A global social media analysis," *The American journal of tropical medicine and hygiene*, **103**(4), p. 1621.

[59] De Coninck, D., T. Frissen, K. Matthijs, L. d'Haenens, G. Lits, O. Champagne-Poirier, M.-E. Carignan, M. D. David, N. Pignard-Cheynel, S. Salerno, et al. (2021) "Beliefs in conspiracy theories and misinformation about COVID-19: Comparative perspectives on the role of anxiety, depression and exposure to and trust in information sources," *Frontiers in psychology*, **12**, p. 646394.

[60] Freiling, I., N. M. Krause, D. A. Scheufele, and D. Brossard (2023) "Believing and sharing misinformation, fact-checks, and accurate information on

social media: The role of anxiety during COVID-19," *New Media & Society*, **25**(1), pp. 141–162.

[61] Brandtzaeg, P. B. and A. Følstad (2017) "Trust and distrust in online fact-checking services," *Communications of the ACM*, **60**(9), pp. 65–71.

[62] Zhou, X. and R. Zafarani (2020) "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, **53**(5), pp. 1–40.

[63] Rashkin, H., E. Choi, J. Y. Jang, S. Volkova, and Y. Choi (2017) "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2931–2937.

[64] Bhatt, G., A. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal (2018) "Combining neural, statistical and external features for fake news stance identification," in *The Web Conf. (WWW)*, pp. 1353–1357.

[65] Ruchansky, N., S. Seo, and Y. Liu (2017) "Csi: A hybrid deep model for fake news detection," in *ACM Conf. on Information and Knowledge Management (CIKM)*, ACM, pp. 797–806.

[66] Jin, Z., J. Cao, H. Guo, Y. Zhang, and J. Luo (2017) "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *ACM Multimedia Conf.*, pp. 795–816.

[67] Wang, Y., F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao (2018) "Eann: Event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ACM, pp. 849–857.

[68] Parikh, S. B. and P. K. Atrey (2018) "Media-rich fake news detection: A survey," in *IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, pp. 436–441.

[69] Bode, L. and E. K. Vraga (2018) "See something, say something: Correction of global health misinformation on social media," *Health Communication*, **33**(9), pp. 1131–1140.

[70] Mosleh, M., C. Martel, D. Eckles, and D. 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*, pp. 1–13.

[71] Nyhan, B. and J. Reifler (2015) "Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information," *Vaccine*, **33**(3), pp. 459–464.

[72] Jiang, S. and C. Wilson (2018) "Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media," *Proceedings of the ACM on Human-Computer Interaction*, **2**(CSCW), p. 82.

[73] Ecker, U. K., S. Lewandowsky, B. Swire, and D. Chang (2011) "Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction," *Psychonomic Bulletin & Review*, **18**(3), pp. 570–578.

[74] Loftus, E. F. (2005) "Planting misinformation in the human mind: A 30-year investigation of the malleability of memory," *Learning & Memory*, **12**(4), pp. 361–366.

[75] Seifert, C. M. (2002) "The continued influence of misinformation in memory: What makes a correction effective?" in *Psychology of Learning and Motivation*, vol. 41, Elsevier, pp. 265–292.

[76] Johnson, H. M. and C. M. Seifert (1994) "Sources of the continued influence effect: When misinformation in memory affects later inferences." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**(6), p. 1420.

[77] Thorson, E. (2016) "Belief echoes: The persistent effects of corrected misinformation," *Political Communication*, **33**(3), pp. 460–480.

[78] Rocha, Y. M., G. A. de Moura, G. A. Desidério, C. H. de Oliveira, F. D. Lourencco, and L. D. 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*, pp. 1–10.

[79] Pennycook, G., T. Cannon, and D. G. Rand (2018) "Prior exposure increases perceived accuracy of fake news," *J. of Experimental Psychology: General*, **147**(12), pp. 1865–1880.

[80] Seo, H., A. Xiong, and D. Lee (2019) "Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation," in *Proceedings of the 10th ACM Conference on Web Science*, pp. 265–274.

[81] Smith, C. N. and H. H. Seitz (2019) "Correcting misinformation about neuroscience via social media," *Science Communication*, **41**(6), pp. 790–819.

[82] Ross, B., A. Jung, J. Heisel, and S. Stieglitz (2018) "Fake News on Social Media: The (In)Effectiveness of Warning Messages," in *Proceedings of the 39th International Conference on Information Systems (ICIS 2018)*, Association for Information Systems, p. 16.

[83] Pennycook, G., A. Bear, E. T. Collins, and D. G. Rand (2020) "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management Science*, **66**(11), pp. 4944–4957.

[84] Van Lent, M., W. Fisher, and M. Mancuso (2004) "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the National Conference on Artificial Intelligence*, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, pp. 900–907.

[85] Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. (2020) "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, **58**, pp. 82–115.

[86] Lee, M. K. (2018) "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data & Society*, **5**(1), p. 2053951718756684.

[87] Confalonieri, R., L. Coba, B. Wagner, and T. R. Besold (2021) "A historical perspective of explainable Artificial Intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **11**(1), p. e1391.

[88] Shin, D. (2021) "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *International Journal of Human-Computer Studies*, **146**, p. 102551.

[89] Boyd, D. M. and N. B. Ellison (2007) "Social network sites: Definition, history, and scholarship," *Journal of Computer-mediated Communication*, **13**(1), pp. 210–230.

[90] Bechmann, A. and S. Lomborg (2013) "Mapping actor roles in social media: Different perspectives on value creation in theories of user participation," *New Media & Society*, **15**(5), pp. 765–781.

[91] Gesser-Edelsburg, A., A. Diamant, R. Hijazi, and G. S. Mesch (2018) "Correcting misinformation by health organizations during measles outbreaks: A controlled experiment," *PLoS One*, **13**(12), p. e0209505, `https://doi.org/10.1371/journal.pone.0209505`.

[92] Vo, N. and K. Lee (2019) "Learning from fact-checkers: Analysis and generation of fact-checking language," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–344.

[93] Vraga, E. K. and L. Bode (2018) "I do not believe you: how providing a source corrects health misperceptions across social media platforms," *Information, Communication & Society*, **21**(10), pp. 1337–1353.

[94] Bawden, D. and L. Robinson (2009) "The dark side of information: Overload, anxiety and other paradoxes and pathologies," *J. of Information Science*, **35**(2), pp. 180–191.

[95] Cavanagh, M. (2018), "Climate change: 'Fake news', real fallout." Accessed: 2019-01-10.
URL `https://goo.gl/tCbwYq`

[96] Shane, S. (2017), "From headline to photograph, a fake news masterpiece," .
URL `https://goo.gl/tmiw7s`

[97] et al., C. S. (2016), "Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate." .
URL `https://goo.gl/6pWtTT`

[98] Cadwalladr, C. (2017), "The great British Brexit robbery: how our democracy was hijacked," `https://tinyurl.com/lkhgkdk`, accessed: 2019-01-10.

[99] Datta, A., M. C. Tschantz, and A. Datta (2015) "Automated experiments on ad privacy settings," in *Privacy Enhancing Technologies*, 1, pp. 92–112.

[100] Pasquale, F. (2015) *The black box society: The secret algorithms that control money and information*, Harvard University Press, Cambridge, MA.

[101] Ribeiro, M. T., S. Singh, and C. Guestrin (2016) "Why should i trust you?: Explaining the predictions of any classifier," in *ACM SIGKDD int'l conf. on knowledge discovery and data mining (KDD)*, ACM, pp. 1135–1144.

[102] Macmillan, N. A. and D. C. Creelman (2004) *Detection theory: A user's guide*, Lawrence Erlbaum, Mahwah, NJ.

[103] Swets, J. A. (1964) *Signal detection and recognition in human observers: Contemporary readings*, Wiley, New York, NY.

[104] Burrell, J. (2016) "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society*, **3**(1), pp. 1–12.

[105] Green, D. M. and J. A. Swets (1966) *Signal detection theory and psychophysics*, Wiley, New York, NY.

[106] Canfield, C. I., B. Fischhoff, and A. Davis (2016) "Quantifying phishing susceptibility for detection and behavior decisions," *Human Factors*, **58**(8), pp. 1158–1172.

[107] Xiong, A., R. W. Proctor, W. Yang, and N. Li (2017) "Is domain highlighting actually helpful in identifying phishing web pages?" *Human Factors*, **59**(4), pp. 640–660.

[108] Bai, H. (2018), "Evidence that a large amount of low quality responses on MTurk can be detected with repeated GPS coordinates," `https://goo.gl/19KCHG`.

[109] Hautus, M. J. (1995) "Corrections for extreme proportions and their biasing effects on estimated values of d'," *Behavior Research Methods, Instruments, & Computers*, **27**(1), pp. 46–51.

[110] Herzberg, A. and A. Gbara (2004) *Trustbar: Protecting (even naive) web users from spoofing and phishing attacks*, *Tech. rep.*, Cryptology ePrint Archive, Report 2004/155. http://eprint. iacr. org/2004/155.

[111] Lin, E., S. Greenberg, E. Trotter, D. Ma, and J. Aycock (2011) "Does domain highlighting help people identify phishing sites?" in *ACM CHI*, ACM, pp. 2075–2084.

[112] Egelman, S., L. F. Cranor, and J. Hong (2008) "You've been warned: An empirical study of the effectiveness of web browser phishing warnings," in *ACM CHI*, ACM, pp. 1065–1074.

[113] Felt, A. P., A. Ainslie, R. W. Reeder, S. Consolvo, S. Thyagaraja, A. Bettes, H. Harris, and J. Grimes (2015) "Improving SSL warnings: Comprehension and adherence," in *ACM CHI*, ACM, pp. 2893–2902.

[114] Wu, M., R. C. Miller, and S. L. Garfinkel (2006) "Do security toolbars actually prevent phishing attacks?" in *ACM CHI*, ACM, pp. 601–610.

[115] Proctor, R. W. and J. Chen (2015) "The role of human factors/ergonomics in the science of security: decision making and action selection in cyberspace," *Human Factors*, **57**(5), pp. 721–727.

[116] Xiong, A., R. W. Proctor, W. Yang, and N. Li (2018) "Embedding training within warnings improves skills of identifying phishing webpages," *Human Factors*.

[117] Kelley, C. M. and L. L. Jacoby (2000) "Recollection and familiarity: Process-dissociation," in *The Oxford handbook of memory* (E. E. Tulving and F. I. M. Craik, eds.), Oxford University Press, New York, pp. 215–228.

[118] Guess, A., J. Nagler, and J. Tucker (2019) "Less than you think: Prevalence and predictors of fake news dissemination on Facebook," *Science Advances*, **5**(1), p. eaau4586.

[119] Anderson, J. R. and R. Milson (1989) "Human memory: An adaptive perspective." *Psychological Review*, **96**(4), pp. 703–719.

[120] Roediger III, H. L. and K. B. McDermott (2000) "Tricks of memory," *Current Directions in Psychological Science*, **9**(4), pp. 123–127.

[121] Gao, M., Z. Xiao, K. Karahalios, and W.-T. Fu (2018) "To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles," *ACM CHI*, **2**(CSCW), p. 55.

[122] Clayton, K., S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. 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), pp. 1073–1095.

[123] Reis, J. C., A. Correia, F. Murai, A. Veloso, and F. Benevenuto (2019) "Supervised learning for fake news detection," *IEEE Intelligent Systems*, **34**(2), pp. 76–81.

[124] Mosallanezhad, A., M. Karami, K. Shu, M. V. Mancenido, and H. Liu (2022) "Domain Adaptive Fake News Detection via Reinforcement Learning," in *Proceedings of the ACM Web Conference 2022*, pp. 3632–3640.

[125] Shu, K., L. Cui, S. Wang, D. Lee, and H. Liu (2019) "defend: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 395–405.

[126] Mohseni, S., F. Yang, S. K. Pentyala, M. Du, Y. Liu, N. Lupfer, X. Hu, S. Ji, and E. D. Ragan (2021) "Machine Learning Explanations to Prevent Overtrust in Fake News Detection." in *ICWSM*, pp. 421–431.

[127] Nguyen, A. T., A. Kharosekar, S. Krishnan, S. Krishnan, E. Tate, B. C. Wallace, and M. Lease (2018) "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*, pp. 189–199.

[128] Horne, B. D., D. Nevo, J. O'Donovan, J.-H. Cho, and S. Adali (2019) "Rating reliability and bias in news articles: Does AI assistance help everyone?" in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 247–256.

[129] Lu, Z., P. Li, W. Wang, and M. Yin (2022) "The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence," *Proceedings of the ACM on Human-Computer Interaction*, **6**(CSCW2), pp. 1–27.

[130] Kim, T. and H. Song (2020) "The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence's Suggestion," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–8.

[131] Calisto, F. M., N. Nunes, and J. C. Nascimento (2022) "Modeling adoption of intelligent agents in medical imaging," *International Journal of Human-Computer Studies*, **168**, p. 102922.

[132] Chong, L., G. Zhang, K. Goucher-Lambert, K. Kotovsky, and J. Cagan (2022) "Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice," *Computers in Human Behavior*, **127**, p. 107018.

[133] Wintersberger, P., N. van Berkel, N. Fereydooni, B. Tag, E. L. Glassman, D. Buschek, A. Blandford, and F. Michahelles (2022) "Designing for Continuous Interaction with Artificial Intelligence Systems," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–4.

[134] Epstein, Z., N. Foppiani, S. Hilgard, S. Sharma, E. Glassman, and D. Rand (2022) "Do explanations increase the effectiveness of AI-crowd generated fake news warnings?" in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, pp. 183–193.

[135] Savolainen, R. (2021) "Assessing the credibility of COVID-19 vaccine mis/disinformation in online discussion," *Journal of Information Science*, p. 01655515211040653.

[136] Flanagin, A. J. and M. J. Metzger (2000) "Perceptions of Internet information credibility," *Journalism & Mass Communication Quarterly*, **77**(3), pp. 515–540.

[137] Gaziano, C. and K. McGrath (1986) "Measuring the concept of credibility," *Journalism Quarterly*, **63**(3), pp. 451–462.

[138] Kang, M. (2010) "Measuring social media credibility: A study on a measure of blog credibility," *Institute for Public Relations*, **4**(4), pp. 59–68.

[139] Kim, S. (2010) "Questioners' credibility judgments of answers in a social question and answer site," *Information Research*, **15**(2), pp. 15–2.

[140] Westerman, D., P. R. Spence, and B. Van Der Heide (2014) "Social media as information source: Recency of updates and credibility of information," *Journal of Computer-mediated Communication*, **19**(2), pp. 171–183.

[141] Lin, X., P. R. Spence, and K. A. Lachlan (2016) "Social media and credibility indicators: The effect of influence cues," *Computers in Human Behavior*, **63**, pp. 264–271.

[142] Pieters, W. (2011) "Explanation and trust: what to tell the user in security and AI?" *Ethics and Information Technology*, **13**(1), pp. 53–64.

[143] Cheng, H.-F., R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu (2019) "Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders," in *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–12.

[144] Wang, X. and M. Yin (2021) "Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making," in *26th International Conference on Intelligent User Interfaces*, pp. 318–328.

[145] Choe, E. K., J. Jung, B. Lee, and K. Fisher (2013) "Nudging people away from privacy-invasive mobile apps through visual framing," in *IFIP Conference on Human-Computer Interaction*, Springer, pp. 74–91.

[146] Rosenblatt, D. H., S. Bode, H. Dixon, C. Murawski, P. Summerell, A. Ng, and M. Wakefield (2018) "Health warnings promote healthier dietary decision making: Effects of positive versus negative message framing and graphic versus text-based warnings," *Appetite*, **127**, pp. 280–288.

[147] Greene, C. M. and G. Murphy (2021) "Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation." *Journal of Experimental Psychology: Applied*, **27**(4), p. 773.

[148] Dzindolet, M. T., S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck (2003) "The role of trust in automation reliance," *International Journal of Human-computer Studies*, **58**(6), pp. 697–718.

[149] Chancey, E. T., J. P. Bliss, Y. Yamani, and H. A. Handley (2017) "Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence," *Human Factors*, **59**(3), pp. 333–345.

[150] Kocielnik, R., S. Amershi, and P. N. Bennett (2019) "Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

[151] Mayer, R. C., J. H. Davis, and F. D. Schoorman (1995) "An integrative model of organizational trust," *Academy of Management Review*, **20**(3), pp. 709–734.

[152] Hoff, K. A. and M. Bashir (2015) "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, **57**(3), pp. 407–434.

[153] Lee, J. D. and K. A. See (2004) "Trust in automation: Designing for appropriate reliance," *Human Factors*, **46**(1), pp. 50–80.

[154] Lipton, Z. C. (2018) "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, **16**(3), pp. 31–57.

[155] Toreini, E., M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. Van Moorsel (2020) "The relationship between trust in AI and trustworthy machine learning technologies," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 272–283.

[156] Peer, E., D. Rothschild, A. Gordon, Z. Evernden, and E. Damer (2022) "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, **54**(4), pp. 1643–1662.

[157] Hauser, D. J. and N. Schwarz (2016) "Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants," *Behavior Research Methods*, **48**(1), pp. 400–407.

[158] Seo, H., A. Xiong, S. Lee, and D. Lee (2021) "(In)effectiveness of Accumulated Correction on COVID-19 Misinformation," in *TMS Proceedings 2021*, https://tmb.apaopen.org/pub/ss8t2ayg.

[159] Cleveland, W. S. (1985) *The elements of graphing data*, Wadsworth Publ. Co.

[160] Norman, G. (2010) "Likert scales, levels of measurement and the "laws" of statistics," *Advances in Health Sciences Education*, **15**(5), pp. 625–632.

[161] Shafir, E. (1993) "Choosing versus rejecting: Why some options are both better and worse than others," *Memory & Cognition*, **21**(4), pp. 546–556.

[162] Chen, J., C. S. Gates, N. Li, and R. W. Proctor (2015) "Influence of risk/safety information framing on android app-installation decisions," *Journal of Cognitive Engineering and Decision Making*, **9**(2), pp. 149–168.

[163] Cavanagh, M. (2017), "Floor Effect / Basement Effect: Definition." Accessed: 2023-01-10.
URL `https://www.statisticshowto.com/floor-effect/`

[164] Dixon, S. R., C. D. Wickens, and J. S. McCarley (2007) "On the independence of compliance and reliance: Are automation false alarms worse than misses?" *Human Factors*, **49**(4), pp. 564–572.

[165] Rice, S. (2009) "Examining single-and multiple-process theories of trust in automation," *The Journal of General Pychology*, **136**(3), pp. 303–322.

[166] Gulati, R. (1995) "Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances," *Academy of Management Journal*, **38**(1), pp. 85–112.

[167] Barr, A. (1999) *Familiarity and trust: An experimental investigation*, University of Oxford.

[168] Zhang, J., A. A. Ghorbani, et al. (2004) "Familiarity and Trust: Measuring Familiarity with a Web Site." in *PST*, Citeseer, pp. 23–28.

[169] Gulati, R. and M. Sytch (2008) "Does familiarity breed trust? Revisiting the antecedents of trust," *Managerial and Decision Economics*, **29**(2-3), pp. 165–190.

[170] Wogalter, M. S., D. DeJoy, and K. R. Laughery (1999) *Warnings and risk communication*, CRC Press.

[171] Weigold, A. and I. K. Weigold (2021) "Traditional and Modern Convenience Samples: An Investigation of College Student, Mechanical Turk, and Mechanical Turk College Student Samples," *Social Science Computer Review.*, p. https://doi.org/10.1177/08944393211006847.

[172] Burnham, M. J., Y. K. Le, and R. L. Piedmont (2018) "Who is Mturk? Personal characteristics and sample consistency of these online workers," *Mental Health, Religion & Culture*, **21**(9-10), pp. 934–944.

[173] Hauser, D. J., A. J. Moss, C. Rosenzweig, S. N. Jaffe, J. Robinson, and L. Litman (2022) "Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk," *Behavior Research Methods*, pp. 1–12.

[174] Facebook (2020), "Here's how we're using AI to help detect misinformation," `https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/`.

[175] Roth, Y. and N. Pickles (2021), "Updating our approach to misleading information," `https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information`.

[176] Starcevic, V. and D. Berle (2013) "Cyberchondria: towards a better understanding of excessive health-related Internet use," *Expert Review of Neurotherapeutics*, **13**(2), pp. 205–213.

[177] McMullan, R. D., D. Berle, S. Arnáez, and V. Starcevic (2019) "The relationships between health anxiety, online health information seeking, and cyberchondria: Systematic review and meta-analysis," *Journal of Affective Disorders*, **245**, pp. 270–278.

[178] Tasnim, S., M. M. Hossain, and H. Mazumder (2020) "Impact of rumors and misinformation on COVID-19 in social media," *Journal of Preventive Medicine and Public Health*, **53**(3), pp. 171–174.

[179] Jungmann, S. M. and M. Witthöft (2020) "Health anxiety, cyberchondria, and coping in the current COVID-19 pandemic: Which factors are related to coronavirus anxiety?" *Journal of Anxiety Disorders*, **35**, p. 102239.

[180] Asmundson, G. J. and S. Taylor (2020) "How health anxiety influences responses to viral outbreaks like COVID-19: What all decision-makers, health authorities, and health care professionals need to know," *Journal of Anxiety Disorders*, **71**, p. 102211, `https://doi.org/10.1016/j.janxdis.2020.102211`.

[181] Banerjee, D., T. S. Rao, et al. (2020) "Psychology of misinformation and the media: Insights from the COVID-19 pandemic," *Indian Journal of Social Psychiatry*, **36**(5), pp. 131–137.

[182] Laato, S., A. Islam, M. N. Islam, and E. Whelan (2020) "Why do people share misinformation during the Covid-19 pandemic?" *arXiv preprint arXiv:2004.09600*.

[183] Lucock, M. P. and S. Morley (1996) "The health anxiety questionnaire," *British Journal of Health Psychology*, **1**(2), pp. 137–150.

[184] Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012) "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk," *Political Analysis*, **20**(3), pp. 351–368.

[185] Briones, E. M. and G. Benham (2017) "An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples," *Behavior Research Methods*, **49**(1), pp. 320–334.

[186] Brennen, J. S., F. Simon, P. N. Howard, and R. K. Nielsen (2020) "Types, sources, and claims of COVID-19 misinformation," *Reuters Institute*, **7**(3), pp. 1–13.

[187] Calvillo, D. P., B. J. Ross, R. J. Garcia, T. J. Smelter, and A. M. Rutchick (2020) "Political ideology predicts perceptions of the threat of covid-19 (and susceptibility to fake news about it)," *Social Psychological and Personality Science*, **11**(8), pp. 1119–1128.

[188] Herzog, M. H., G. Francis, and A. Clarke (2019) "ANOVA," in *Understanding Statistics and Experimental Design*, Springer, pp. 67–82.

[189] Ecker, U. K., S. Lewandowsky, and J. Apai (2011) "Terrorists brought down the plane!—No, actually it was a technical fault: Processing corrections of emotive information," *Quarterly Journal of Experimental Psychology*, **64**(2), pp. 283–310.

[190] Bakeman, R. (2005) "Recommended effect size statistics for repeated measures designs," *Behavior Research Methods*, **37**(3), pp. 379–384.

[191] Lakens, D. (2013) "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs," *Frontiers in Psychology*, **4**, p. 863.

[192] Olejnik, S. and J. Algina (2003) "Generalized eta and omega squared statistics: measures of effect size for some common research designs." *Psychological Methods*, **8**(4), pp. 434–447.

[193] Uscinski, J. E., A. M. Enders, C. Klofstad, M. Seelig, J. Funchion, C. Everett, S. Wuchty, K. Premaratne, and M. Murthi (2020) "Why do people believe COVID-19 conspiracy theories?" *Harvard Kennedy School Misinformation Review*, **1**, p. 3.

[194] Jiang, S., M. Metzger, A. Flanagin, and C. Wilson (2020) "Modeling and measuring expressed (dis) belief in (mis) information," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 315–326.

[195] Lee, C. S. and L. Ma (2012) "News sharing in social media: The effect of gratifications and prior experience," *Computers in Human Behavior*, **28**(2), pp. 331–339.

[196] Frenda, S. J., E. D. Knowles, W. Saletan, and E. F. Loftus (2013) "False memories of fabricated political events," *Journal of Experimental Social Psychology*, **49**(2), pp. 280–286.

[197] Benegal, S. D. and L. A. Scruggs (2018) "Correcting misinformation about climate change: The impact of partisanship in an experimental setting," *Climatic Change*, **148**(1), pp. 61–80.

[198] Pennycook, G. and D. G. Rand (2019) "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*, **188**, pp. 39–50.

[199] Seo, H., S. Lee, D. Lee, and A. Xiong (2024) "Reliability Matters: Exploring the Effect of AI Explanations on Misinformation Detection With a Warning," .

[200] Seo, H., A. Xiong, S. Lee, and D. Lee (2022) "If You Have a Reliable Source, Say Something: Effects of Correction Comments on COVID-19 Misinformation," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, pp. 896–907.

[201] Kaiser, B., J. Wei, E. Lucherini, K. Lee, J. N. Matias, and J. Mayer (2020) "Adapting security warnings to counter online disinformation," *arXiv preprint arXiv:2008.10772*.

[202] Vereschak, O., G. Bailly, and B. Caramiaux (2021) "How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies," *Proceedings of the ACM on Human-Computer Interaction*, **5**(CSCW2), pp. 1–39.

[203] Moss, A. and L. Litman (2018) "After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it," *Retrieved February*, **4**, p. 2019.

[204] Eyal, P., R. David, G. Andrew, E. Zak, and D. Ekaterina (2021) "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, pp. 1–20.

[205] Ecker, U. K., S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, and M. A. Amazeen (2022) "The psychological drivers of misinformation belief and its resistance to correction," *Nature Reviews Psychology*, **1**(1), pp. 13–29.

[206] Greifeneder, R., M. Jaffe, E. Newman, and N. Schwarz (2021) *The psychology of fake news: Accepting, sharing, and correcting misinformation*.

[207] Anspach, N. M. (2017) "The new personal influence: How our Facebook friends influence the news we read," *Political communication*, **34**(4), pp. 590–606.

[208] McGee, R. W. (2023) "Is chat gpt biased against conservatives? an empirical study," *An Empirical Study (February 15, 2023)*.

[209] Hartmann, J., J. Schwenzow, and M. Witte (2023) "The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation," *arXiv preprint arXiv:2301.01768*.

[210] Mirsky, Y. and W. Lee (2021) "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, **54**(1), pp. 1–41.

Vita

Haeseung Seo

EDUCATION

The Pennsylvania State University 08/2017 - present

Ph.D. Candidate in Informatics
Advisor: Dr.Dongwon Lee, Dr.Aiping Xiong

Seoul National University

M.S. in Engineering 09/2012 - 02/2015
B.A. in Information and Culture Technology 03/2005 - 08/2012
B.A. in Political Science
B.A. in Korean History

SELECTED PUBLICATIONS

- **Haeseung Seo**, Sian Lee, Dongwon Lee, Aiping Xiong. "Reliability Matters: Exploring the Effect of AI Explanations on Misinformation Detection With a Warning." *AAAI ICWSM*. 2024. (In press)
- Sian Lee, Aiping Xiong, **Haeseung Seo**, Dongwon Lee. "Data-driven Approach to FactChecking the Fact-checkers." *HKS Misinformation Review*. 2023. (In press)
- Sian Lee, **Haeseung Seo**, Dongwon Lee, Aiping Xiong. "Associative Inference Can Increase People's Susceptibility to Misinformation." *AAAI ICWSM*. 2023.
- **Haeseung Seo**, Aiping Xiong, Sian Lee, Dongwon Lee. "If You See a Reliable Source, Say Something: Effects of Correction Comments on COVID-19 Misinformation." *AAAI ICWSM*. 2022.
- Aiping Xiong, Sian Lee, **Haeseung Seo**, Dongwon Lee. " Effects of Associative Inference on Individuals' Susceptibility to Partisan Misinformation." *Journal of Experimental Psychology: Applied*. 2022.
- **Haeseung Seo**, Aiping Xiong, Sian Lee, Dongwon Lee. "(In)effectiveness of Accumulated Correction on COVID-19 Misinformation." *Technology, Mind and Society*. 2021.
- Limeng Cui, **Haeseung Seo**, Maryam Tabar, Fenglong Ma, Suhang Wang, Dongwon Lee. "DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation." *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2020.
- **Haeseung Seo**, Aiping Xiong, Dongwon Lee. "Trust It or Not: Effects of Machine Learning Warning in Helping Individuals Mitigate Misinformation." *ACM Web Science*. 2019.