Profiling the Dynamics of Trust & Distrust in Social Media: A Survey Study

Yixuan Zhang*
William & Mary
Williamsburg, VA, USA
yzhang104@wm.edu

Yimeng Wang*
William & Mary
Williamsburg, VA, USA
ywang139@wm.edu

Nutchanon Yongsatianchot
Thammasat University
Thailand
ynutchan@engr.tu.ac.th

Joseph D Gaggiano
Georgia Institute of Technology
Atlanta, GA, USA
jgaggiano@gatech.edu

Nurul M Suhaimi
Universiti Malaysia Pahang
Pahang, Malaysia
nmsuhaimi@ump.edu.my

Anne Okrah
Northeastern University
Boston, MA, USA
okrah.a@northeastern.edu

Miso Kim
Northeastern University
Boston, MA, USA
m.kim@northeastern.edu

Jacqueline Griffin
Northeastern University
Boston, MA, USA
ja.griffin@northeastern.edu

Andrea G Parker
Georgia Institute of Technology
Atlanta, GA, USA
andrea@cc.gatech.edu

ABSTRACT

In the era of digital communication, misinformation on social media threatens the foundational trust in these platforms. While myriad measures have been implemented to counteract misinformation, the complex relationship between these interventions and the multifaceted dynamics of trust and distrust on social media remains underexplored. To bridge this gap, we surveyed 1,769 participants in the U.S. to gauge their trust and distrust in social media and examine their experiences with anti-misinformation features. Our research demonstrates how trust and distrust in social media are not simply two ends of a spectrum; but can also co-exist, enriching the theoretical understanding of these constructs. Furthermore, participants exhibited varying patterns of trust and distrust across demographic characteristics and platforms. Our results also show that current misinformation interventions helped heighten awareness of misinformation and bolstered trust in social media, but did not alleviate underlying distrust. We discuss theoretical and practical implications for future research.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

social media, trust, distrust, misinformation, survey

*Both authors contributed equally to this research.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642927

ACM Reference Format:
Yixuan Zhang, Yimeng Wang, Nutchanon Yongsatianchot, Joseph D Gaggiano, Nurul M Suhaimi, Anne Okrah, Miso Kim, Jacqueline Griffin, and Andrea G Parker. 2024. Profiling the Dynamics of Trust & Distrust in Social Media: A Survey Study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3613904.3642927

1 INTRODUCTION

In this digital age, where information spreads at an unprecedented speed and volume, misinformation poses a pervasive and challenging threat. As social media platforms have grown from mere social connectors to global influencers, they have also become major vehicles for spreading misinformation. This spread of misinformation can have profound implications ranging from impacting public behavior during health crises to shaping political landscapes [5, 20]. Governments, industry, and other stakeholders recognize the urgency of addressing issues of misinformation [29, 68] and have begun to implement solutions aimed at tackling these issues, such as fact-checking or flagging misleading content using algorithmic-centered approaches [3, 28, 64, 75].

However, misinformation isn't just a technical problem to be solved; it is a human-centric issue rooted in perception, cognition, and emotion [78]. As misinformation pervades social media, one important negative consequence of misinformation is the erosion of trust in social media platforms and information sources [3]. The erosion of trust not only challenges the credibility of platforms, but also risks turning them into echo chambers, limiting their role as vibrant, diverse, and informative spaces. Therefore, it is crucial to examine to what extent people trust social media. Without a situated understanding of trust, interventions might fall flat or even exacerbate the issue [77]. It is also important to acknowledge that not all misinformation interventions inherently warrant increased trust. To mitigate the impact of misinformation, we need to examine if and how trust can be reconstructed in the wake of its breach.

Trust, the cornerstone of any relationship or interaction, is especially pertinent in digital ecosystems since it involves security, authenticity, and reliability in a world where connections are formed virtually [26]. In the context of social media information dissemination and communication, trust influences user behavior, from engagement with content to decisions based on information received from these platforms [30]. More crucially, the concept of "*distrust*," a related term of trust, further complicates our understanding of the dynamics of trust. Many existing works have either considered trust and distrust to be two extremes of the same dimension or did not explicitly examine distrust [77]. Simultaneously, some scholars contend whether trust and distrust are indeed two poles of a single continuum or if they stand as distinct, independent concepts [9, 38]. These contrasting perspectives highlight the need for conceptual clarity, particularly within the realm of social media, as these platforms are powerful vehicles for information dissemination. Thus, a nuanced understanding of not only trust but also distrust in the context of social media is essential.

Furthermore, different populations often bring with them unique historical, cultural, and socio-economic experiences that shape their interactions and perceptions of social media [33]. These differences may influence how various groups view and respond to misinformation interventions, and consequently, how they trust or distrust digital entities. In parallel, each social media platform has its own culture [72]. For example, with its concise messaging format and rapid news sharing, Twitter [1] may be perceived and trusted differently than Facebook's community-centric feeds. And yet, little work has explored both trust and distrust concerning diverse demographic groups across various social media platforms. Understanding these demographic and platform-specific nuances is crucial, as it allows for deeper insights into misinformation interventions and how they align with the perceptions of diverse audiences. Our work addresses these research gaps.

This paper investigates the complexities of *trust and distrust* in this heightened age of misinformation on social media. Specifically, the following research questions (RQs) guided our research:

RQ1. Are trust and distrust in social media inherently linked, such that an increase in one means a decrease in the other? Or can they coexist independently?

RQ2a. How do trust and distrust in social media differ across platforms?

RQ2b. How do trust and distrust in social media vary across different demographic groups?

RQ3. How do people's experiences with misinformation interventions associate with their trust and distrust in social media?

To answer these questions, we conducted a survey study with a nationally representative sample in the U.S. (1,769 participants) in March 2023. Our results offer empirical evidence supporting the idea that trust and distrust can be viewed as distinct concepts rather than merely opposite sides of a singular notion. This dual trust-distrust perspective enriches our comprehension of the complex dynamics of online trust. Our analysis further suggests that individuals can be grouped into different categories based on their trust and distrust levels. Our results also reveal that the levels of trust and distrust vary across platforms and show variations in how demographic factors relate to these levels on different social media platforms. Furthermore, our findings suggest that implementing misinformation interventions in social media has the potential to amplify individuals' awareness of misinformation while concurrently strengthening their trust in various social media platforms. However, these misinformation intervention features do not necessarily reduce underlying distrust in social media. Recognizing these nuances is essential, as it paves the way for addressing issues of trust and distrust and designing future misinformation interventions.

In this work, we contribute: **(1)** a comprehensive empirical study that investigates the relationship between trust and distrust in social media, along with an in-depth analysis of the variances in these dynamics across diverse platforms and among various demographic groups, collectively contributing an enhanced theoretical comprehension of trust and distrust dynamics; **(2)** new scales for measuring trust and distrust in social media that we validated in our study that benefit future researchers; **(3)** an understanding of people's use of, perceptions about, and trust in misinformation interventions on social media; and **(4)** theoretical and practical implications for future work.

Before further discussions, we first establish our operational definitions for trust and distrust. In this paper, we define *trust* as an individual's belief in the competence, benevolence, integrity, and reliance of social media [77]. We conceptualize *distrust* as a cognitive and emotional state stemming from perceived dishonesty, skepticism towards intentions or outcomes, fear of potential harm or deceit, and concerns of malevolence from another entity. We will discuss the measurements in detail in section 3 Methods.

2 BACKGROUND & RELATED WORK

2.1 Trust and Distrust in Social Media

Trust is the foundational component that cements stable relationships, whether between individuals, organizations, or the intersection of information and technology where these connections are vital [22, 25, 35]. While a significant body of work has examined trust within the realm of social media, there remains a gap in understanding distrust in the same context [36]. The scholarly debate on this issue presents two dominant perspectives: one positing trust and distrust as opposite ends of a singular continuum [39], and the other conceiving them as separate, distinct entities [38]. For example, prior work has found that while trust may correlate with the frequency of Facebook usage, distrust doesn't necessarily mirror this trend [8]. To this end, scholars [66] have argued that failing to discern their interrelationship could yield incomplete insights and suggested that future work should further examine the dynamics between trust and distrust. As such, this ongoing debate highlights the need to clarify the discourse on trust and distrust with empirical findings.

Furthermore, trust and distrust are highly contextualized and are cultivated or eroded in specific tasks within particular situations [7, 24, 54]. For example, prior work has highlighted factors influencing trust, including service quality and the usability of a platform [60]. In another context, different factors were highlighted when examining trust in AI [31], such as the degree of automation in the AI system and its performance capabilities. These studies

[1] In April 2023, Twitter was renamed X. For the purpose of this paper, we continue to refer to it as Twitter.

underscore the idea that trust is deeply rooted in its context. Therefore, trust within the context of social media deserves its distinct analysis and attention [77]. Disentangling the complex relationship promises to deepen our insight and pave the way for creating more trustworthy social media spaces.

2.2 Demographic & Platform Differences between Trust and Distrust in Social Media

A substantial body of research has explored demographic differences in the establishment of trust across a diverse array of contexts, such as online commerce [32], health information websites [14], AI-supported tools [56], and social media [65]. These disparities in trust can span over a range of factors, from psychological considerations to demographic attributes. For instance, some studies highlighted the association between demographic variables, such as gender and age, and trust levels [11, 15, 59]. Their findings revealed that women and older adults tend to approach online information with greater caution and trust that information less than their male and younger counterparts, respectively [15, 41].

However, despite the considerable research on demographic differences in trust [44, 51, 69], research examining the interplay between demographic factors and distrust dynamics within social media is sparse. Addressing these multifaceted inquiries necessitates an integrated approach transcending isolated examination of demographic variables. Given the heightened emphasis on understanding trust, it is important to extend this focus to distrust. Furthermore, the rapid evolution of the social media environment, characterized by new platforms and features within existing platforms, poses a moving target. This dynamic landscape underscores the need for ongoing research to maintain the timeliness and relevance of our understanding of trust dynamics. Therefore, our work seeks to bridge this research gap; to comprehensively explore the demographic factors that underpin not only trust but also distrust in social media.

Likewise, each social media platform, inherently designed with unique features and user experiences, fosters its own distinct culture in the digital ecosystem [72]. Prior work comparing trust across these platforms indicates that user trust varies considerably [10, 16, 48, 74], suggesting that there is no monolithic "trust" sentiment when it comes to social media; rather, user trust is fragmented, nuanced, and platform-specific. Furthermore, little research has examined platform differences in distrust in social media. However, the lack of understanding of how distrust varies across platforms may lead to misguided interventions and policy implementations, ultimately failing to address core issues. Our research seeks to address these research gaps.

2.3 Trust, Distrust, and Misinformation Interventions

Scholars have argued that trust should be understood and measured as a fluid state that evolves based on various situational factors [78]. From this perspective, both trust and distrust arise from specific interactions and contextual circumstances. Within the scope of our study, we focus on trust in the context of misinformation on social media.

Prior work has shown that repeated exposure to information leads people to perceive that information as more likely to be accurate, illustrating the persuasive influence of repeated misinformation [71]. Additionally, when confronted with information they previously believed to be false, individuals tend to develop negative emotions toward social media platforms, and they tend to instead gravitate towards other platforms that elicit positive emotions [46]. Consequently, platforms face the challenging yet vital task of combating misinformation to retain users. In recent years, social media platforms have implemented various strategies to counter misinformation, anticipating that it would enhance trust and diminish distrust, such as fact-checking, warning labels, and content removal [50].

As misinformation interventions have continued to evolve, a body of research has investigated their effectiveness from various perspectives. For instance, some studies have used machine learning models to identify and flag misinformation surrounding crisis events [64, 73]. Furthermore, recent work has investigated the effectiveness of misinformation interventions, such as fact-checking, time-lagged approaches, account banning, and combined approaches [4]. This study found that existing interventions were unlikely to be effective if implemented individually. The success of an integrated approach is contingent upon the characteristics of each intervention, their interplay, the pattern of misinformation propagation, the length of the event, patterns of user engagement, the number of followers users have, and the evolution of these elements during a disinformation campaign [4]. Another area of research examines how specific features of misinformation interventions influence people's attitudes. For example, Mena's experimental study highlights the significant impact of flagging misinformation to reduce intentions of sharing false news [40]. However, very little research has explored users' perceptions, trust, and distrust regarding misinformation interventions employed on social media platforms. Therefore, our work aims to understand the ways in which these interventions may be related to people's trust and distrust in the context of the misinformation age on social media.

3 METHODS

In this section, we describe our study procedure, including an analysis of the changes implemented in social media platforms to address the spread of misinformation (see subsection 3.1) and our survey study (see subsection 3.2). Figure 1 shows an overview of our study flow. This research was approved by the Institutional Review Board at our institution.

3.1 Characterizing Changes in Social Media Platforms to Combat Misinformation

Prior work has proposed a variety of approaches to examine the changes made to social media platforms, such as analyzing industry blog posts and screenshots using the Internet Archive Wayback Machine and examining the changes logged to the social media repositories [13, 19, 70]. Inspired by existing work, we first collected and exported all available blog posts from four major social media companies, including Facebook, Twitter, YouTube, and TikTok, about their platforms. These platforms were selected as they are

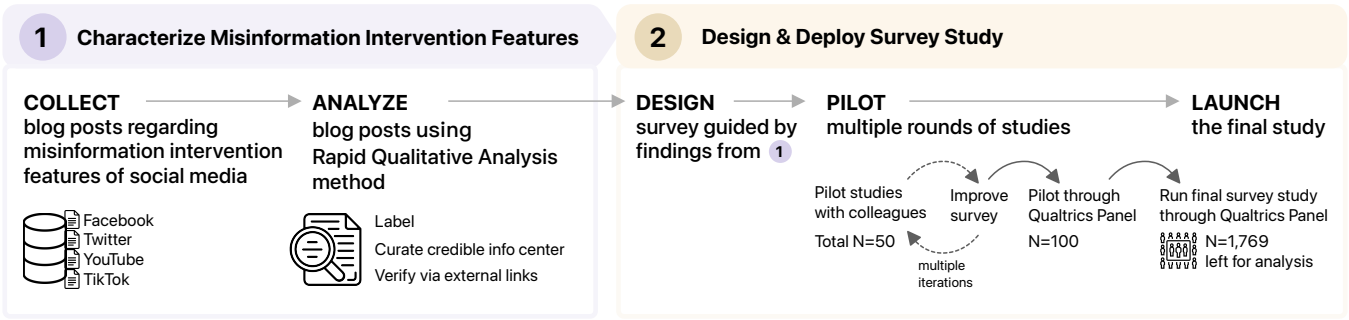

Figure 1: Overview of our study flow: Step 1 includes collecting blog posts regarding misinformation intervention features on social media and analyzing these blogs using the Rapid Qualitative Analysis method [45]. Step 2 includes designing and deploying the survey study (guided by the findings from 1), followed by multiple rounds of pilot studies and the final launch of the study.

among the most commonly-used social media platforms for gathering news and information at the time of our study [34]. In this step, we focused on *"visible"* misinformation intervention features that can be seen by users as they serve as more direct, tangible touchpoints between the platform and its users than the backend or algorithmic interventions.

We specifically targeted the period from January 2017 to January 2023 in our data collection process among social media platforms' blogs, aligning with the rise of widespread misinformation campaigns [42]. The blog posts were filtered based on keywords, such as misinformation and misinformation intervention strategies. These selection criteria were applied uniformly across all platforms. Each retrieved post was then manually reviewed by two researchers in our research team to confirm its relevance to misinformation interventions.

Given the potential immediate and salient impact of these features, it is important to understand their role in fostering a trustworthy digital ecosystem. Characterizing the changes in social media platforms also reveals incremental adjustments that may have important implications in the fight against misinformation.

Then, we conducted a rapid qualitative analysis [23] of these posts. Rapid qualitative analysis is a method to obtain targeted qualitative data and comparative results when data collection targets and processes are highly structured [23]. Research has demonstrated the effectiveness and rigor of rapid qualitative analysis to be comparable to traditional qualitative analysis, despite the streamlined process present in the former [45]. Two researchers first examined the data in its entirety to establish a general understanding of the problem space. Then, given the major overlap in changes made across platforms, the two researchers independently categorized the major features and changes that social media platforms implemented to combat misinformation into higher-level themes.

Overall, several major themes were identified across the existing misinformation interventions using our rapid qualitative analysis: 1) labeling/tagging, 2) credible information curation, and 3) actionable external source verification, as briefly illustrated in Figure 2. Specifically, **(A)** Labeling/Tagging Features include the labeling or tagging of potentially misleading or false information shared on social media platforms. Labels or tags can provide additional context and warnings to help users discern the credibility of the content. **(B)** Curation Features (e.g., Credible Information Centers) are dedicated spaces or sections in social media platforms that curate and showcase credible information from authoritative sources. **(C)** Verification Features (e.g., Clickable External Links) enable access to additional information associated with a particular post or content. These links can lead to additional external resources, fact-checking websites, or trusted sources of information, allowing users to verify the accuracy and credibility of the shared content. These themes also align with misinformation intervention approaches seen in existing literature [2].

3.2 Survey Study

3.2.1 Overall Study Design. Following our analysis of changes across social media platforms, we incorporated key findings into our survey design. Our survey aimed to investigate people's trust and distrust in social media in the context of misinformation countermeasures.

To ensure the validity of our survey, we conducted multiple rounds of pilot studies. First, we conducted three informal pilot studies with our research team and colleagues. During the pilot studies, we identified several areas for refinement. For example, we simplified the language in our questions for clarity to avoid academic jargon, ensuring they were understandable to a general audience. Additionally, we included clarifying examples (e.g., the screenshots shown in Figure 2) next to complex questions to reduce potential misinterpretation.

Subsequently, we conducted a formal pilot test with 100 respondents using the Qualtrics online research panel [2]. The feedback and responses received during the pilot tests were invaluable in refining and developing the final survey instrument. Respondents were compensated based on the estimated time required to complete the survey [3].

We also included one attention check question in the survey to improve the scale validity further. This attention-check question

[2]https://www.qualtrics.com/research-services
[3]The incentive structures were determined by Qualtrics and specific compensation details were not disclosed to the research team

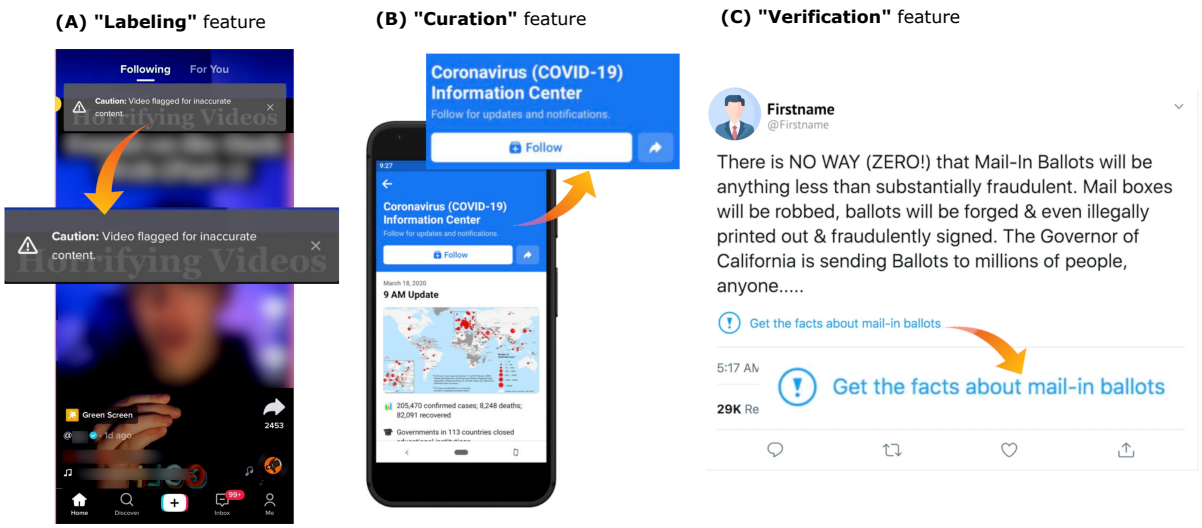

Figure 2: Examples of social media features that seek to combat misinformation: (A) Labeling/Tagging Features, (B) Curation Features, and (C) Verification Features.

was placed early in the survey and had an obviously correct response to identify inattentive respondents. This process allowed us to screen out these respondents prior to conducting any analyses.

3.2.2 Survey Study Recruitment & Overview of Participants. We recruited survey respondents from a third-party service, Qualtrics online research panel, in March of 2023. Qualtrics recruited participants from a nationally representative pool, based on the following **inclusion criteria**: participants were adults (aged 18+) who had used at least one of the four social media platforms of interest (Facebook, Twitter, YouTube, or TikTok) within the last three months. These four platforms were chosen due to their significant impact on global information dissemination, their distinct modes of user interaction, and their important influence on the spread of misinformation at the time of our study [43]. In our study, participants were only asked about platforms they had engaged with in the past three months. Specifically, the reason is that we were interested in answers from people who recently experienced the platforms and the features, as this should yield more accurate and reliable perceptions of misinformation intervention features and trust/distrust.

Before participants were asked to assess their trust concerning misinformation interventions, we presented them with vignettes (e.g., as shown in Figure 2). These vignettes were designed to provide a concrete example of how the misinformation intervention functions, which can help minimize interpretation variability and enhance the accuracy of participants' responses.

To ensure data quality, we excluded participants who met one or more of the following **exclusion criteria**: 1) those who completed the survey in under 10 minutes, which was considered "speeding" based on our pilot test results; 2) those who provided gibberish or unrelated responses to the open-ended question on the definition of misinformation, including nonsensical words like *"gllllsscc"*; and 3) those who took part in any of the previous pilot tests.

Participants Overview. In total, 1,769 participants from the United States were included in our final dataset for analysis. Detailed demographic characteristics of the participants are shown in Table 1. This study contained a nationally representative sample, meaning the demographic distribution closely aligned with the demographic distribution of the United States population.

The majority of respondents identified as women (59%), followed by men (40%) and non-binary or undisclosed (1%). The average age of respondents was 48 years (SD=17). Likewise, the largest proportion of respondents identified as Caucasian/White (50%), followed by African American or Black (26%), Asian (12%), American Indian or Alaskan Native (9%), and Native Hawaiian or Pacific Islander (1%). 13% of the participants indicated that they were Hispanic or Latino. Furthermore, 36% of respondents had a high school diploma or less, 19% had an associate degree, 28% had a bachelor's degree, and 17% had a postgraduate degree. Moreover, 44% of respondents identified themselves as Democrats or Lean Democrats, 28% claimed to be Independent, and 26% saw themselves as Republicans or Lean Republicans. Our respondents also varied across all annual household income levels, which we categorized into low-income (27%) and moderate-to-high-income (73%) based on the 2022 U.S. Federal Poverty Level (185%) Guidelines [1].

3.2.3 Measures. Overall, our survey questions focused on participants' trust and distrust of social media, their experiences with misinformation intervention features, and their demographic background. Below, we provide detailed measurements and scales used in this study.

Trust. Trust in social media was measured using four survey items with a five-point Likert scale, with responses ranging from

Table 1: Demographic characteristics of our participants.

Demographic	Response Options	Number of Participants (Total N = 1,769)	Percentage (%)
Gender	Female	1037	59%
	Male	717	40%
	Prefer not to answer or Non-binary	15	1%
Age*	18-24	168	9%
	25-34	321	18%
	35-44	379	22%
	45-54	214	12%
	55-64	296	17%
	65+	391	22%
	* Mean = 48, SD = 17, Age range = [18, 90]		
Ethnicity & Race	Hispanic or Latino	227	13%
	Not Hispanic or Latino	1541	87%
	Prefer not to answer	1	.05%
	African American or Black	466	26%
	American Indian or Alaskan Native	154	9%
	Asian	213	12%
	Native Hawaiian or Pacific Islander	26	1%
	White	890	50%
	Other	11	1%
	Prefer not to answer	9	1%
Education	Less than high school	39	2%
	High school graduate	597	34%
	Associate degree	331	19%
	Bachelor's degree	503	28%
	Postgraduate degree	299	17%
Political Ideology	Democrat/ Lean Democrat	779	44%
	Independent	494	28%
	Republican/ Lean Republican	461	26%
	Other, please describe	35	2%
Household Income	Less than $25,143	282	16%
	$25,143 - $33,874	141	8%
	$33,875 - $42,606	113	6%
	$42,607 - $51,338	95	5%
	$51,339 - $60,070	130	7%
	$60,071 - $68,802	73	4%
	$68,803 - $77,534	83	5%
	$77,535 - $86,266	66	4%
	$86,267 - $94,998	51	3%
	$94,999 - $103,730	117	7%
	$103,731 - $112,462	77	4%
	$112,463 - $121,194	66	4%
	$121,195 - $129,926	67	4%
	$129,927 - $138,658	55	3%
	$138,659 and above	350	20%
	Prefer not to answer	3	.2%
	Low-income**	481	27%
	Moderate-to-high income**	1281	73%
	Can not be defined**	7	.4%
	** Income levels were determined based on the 2022 U.S. Federal Poverty Level (185%) Guidelines [1].		

strongly disagree (1) to strongly agree (5). This four-item measurement was adapted from prior work [18, 77], which emphasized the importance of specifying the trustee (i.e., the object of trust) and the context of the study when measuring trust. In our research, we operationalize the context as the situations in which users encounter misinformation on social media. The items of trust measurement correspond to four trust dimensions detailed in a systematic review, including benevolence, integrity, competence, and reliance [77]. These four items were utilized to create the following four statements that comprised the measurement of trust in social media in our study:

(1) *I believe that <social media platform>* [4] *cares about helping me avoid misinformation.*
Rationale: This measure reflects the dimension of benevolence, which relates to people's perceptions about the intentions of social media platforms, and their perceptions about platforms' levels of concern for users' well-being in the misinformation age.
(2) *<Social media platform> is reliable because it attempts to combat the spread of misinformation.*

[4]The placeholder term *<social media platform>* was replaced with specific platform names (i.e., Facebook, Twitter, TikTok, and YouTube) in the actual survey.

Rationale: This measure corresponds to the dimension of reliability. In this context, reliability refers to the perceived effectiveness and commitment of a social media platform in combating misinformation.

(3) *I feel very confident about <social media platform>'s ability to address misinformation.*
Rationale: This measure aligns with the dimension of competence. Competence refers to the level of confidence that people place in <social media>'s ability to tackle misinformation effectively.

(4) *I am willing to act upon the information I get on <social media platform>.*
Rationale: This measure relates to the dimension of reliance. Reliance signifies the level of trust users place in the information received from <social media>, and whether or not they trust that information enough to take action based on it.

Note that in the actual survey study, the term "*<social media platform>*" was replaced with specific platform names, including Facebook, Twitter, TikTok, and YouTube. This customization allowed for a more targeted assessment of participants' trust perceptions towards different social media platforms, which aligned with findings from the aforementioned systematic review [77] in that trust in social media may differ depending on the platform.

Respondents also rated their level of trust in social media when it contains a particular type of misinformation intervention feature, from strongly disagree (1) to strongly agree (5). Specifically, we provided four-item measurements including *"In your opinion, when a social media platform has this <misinformation intervention feature>"... 1) it shows that the platform cares about helping me avoid misinformation."*, 2) *"it shows that the platform is reliable because it attempts to combat the spread of misinformation"*, 3) *"it makes me feel more confident in the platform's ability to address misinformation"*, and 4) *"I am more willing to act upon the information I get on this platform."*. In this set of questions, the placeholder term "*<misinformation intervention feature>*" was replaced with specific feature names (i.e., labeling, curation, verification features as shown in Figure 2).

Distrust. We measured distrust in social media using four survey items with a five-point Likert scale, ranging from strongly disagree (1) to strongly agree (5). This measurement of distrust was inspired by the systematic review in social media [77]. Four items comprised the measurement of distrust in social media in our study, including skepticism, dishonesty, malevolence, and fear, which were utilized to create the following four statements:

(1) *I am skeptical about whether <social media platform> keeps my interests in mind when it makes decisions on addressing misinformation.*
Rationale: This measurement aligns with the dimension of skepticism. It reflects distrust in <social media platform>'s intentions among users, raising doubts about whether the platform prioritizes the user's interests when making decisions related to addressing misinformation.

(2) *<Social media platform> intentionally allows misinformation to stay on its platform.*
Rationale: This measurement indicates a dimension of dishonesty. If a user believes that <social media> knowingly permits misinformation to persist on its platform, it may lead to distrusting social media.

(3) *<Social media platform> transmits misinformation for its own interests.*
Rationale: This measurement also relates to the dimension of malevolence. It implies the perception that <social media platform> propagates misinformation to serve its own interests, shedding light on the belief that it prioritizes its own agenda over providing accurate information.

(4) *The prevalence of misinformation on <social media platform> makes me fear using this platform.*
Rationale: This measurement maps to the dimension of fear. It reflects a failure to address misinformation may lead to fear or apprehension towards using social media due to the widespread of misinformation.

Additionally, respondents rated their level of distrust in social media when it contains a particular type of misinformation intervention feature, from strongly disagree (1) to strongly agree (5). The four-item measurements of distrust include *"In your opinion, when a social media platform has this <labeling/tagging feature>"... 1) I am skeptical about whether the platform keeps my interest in mind when it makes decisions about addressing misinformation"*, 2) *"it shows that the platform intentionally allows misinformation to stay on it."*, 3) *"it shows that the platform spreads misinformation for its own interests.'*, and 4) *"it shows that misinformation is widespread on the platform, making me fear using it"*. Similarly, in this set of questions, the placeholder term "*<misinformation intervention feature>*" was also replaced with specific feature names (i.e., labeling, curation, verification features as shown in Figure 2).

Frequency of exposure to misinformation intervention features. To understand how often participants were exposed to misinformation intervention features (i.e., Labeling/Tagging Features, Curation Features, and Verification Features), they were asked the question, *"How often do you see the above kind of feature on social media?"*. Note that participants were shown example images of the individual features (see Figure 2) alongside this question for reference. Ratings were provided ranging from (1) never to (5) always.

Relationship between experiences with misinformation intervention features and people's attitudes and behaviors. We also aimed to explore the potential influence that participants' prior experiences with the misinformation intervention features had on their awareness of misinformation, information-sharing intentions, and desire to receive information from that social media platform. For example, with respect to labeling/tagging features, we used the prompt *"Thinking about your experiences with this feature, please indicate how much you agree or disagree with the following statements about this labeling/tagging feature."*. Participants were asked to rate their level of agreement, ranging from strongly disagree (1) to strongly agree (5), with three separate statements, including:

(1) Overall, these labeling/tagging features make me more aware of misinformation.

(2) *I am more likely to share posts from social media platforms that have these labels/tags.*
(3) *When a social media platform has these labeling/tagging features, it makes me want to receive more information from the platform.*

We replicated these questions to ask participants to reflect on the rest of the misinformation feature categories (i.e., "labeling/tagging features" in the above prompt was replaced with "curation features" and "verification features"). This approach enabled us to examine and compare the extent to which participants' experiences with each type of misinformation feature impacted their misinformation awareness, information-sharing intentions, and desire to receive information on platforms.

Demographic Background. Participants were asked a few questions about their demographic background, including age, sex, race and ethnicity, education, political ideology, and income. Participant demographic data is summarized in Table 1.

Age. Participants were asked to provide their age in the survey (as a numeric value). Table 1 shows grouped categories of age distribution.

Sex. Participants were provided options of "Female", "Male", and "Prefer to self-describe", and "Prefer not to answer".

Race & Ethnicity. Participants were first asked "*Do you consider yourself Hispanic or Latino?*". Then, they were asked to choose one or more races with which they most closely identify. Response options included "African American or Black", "American Indian or Alaskan Native", "Asian", "Native Hawaiian or Pacific Islander", and "White", with the additional options of "Prefer not to answer" or "Self-describe".

Education. Participants were asked about their education level using the question, "*What is the highest degree or level of school you have completed?*" Response options included "Less than high school", "High school graduate", "Associate degree", "Bachelor's degree", and "Postgraduate degree".

Political Ideology. Participants were asked to describe the political viewpoint with which they most closely aligned. Response options included "Democrat/Lean Democrat", "Independent", "Republican/Lean Republican", and "Other, please describe."

Income. Participants were asked about their income level using the statement, "*Please indicate the answer that most closely matches your entire household income in 2022 before taxes.*". Response options included "$25,143 - $33,874", "$33,875 - $42,606", "$42,607 - $51,338", "$51,339 - $60,070", "$60,071 - $68,802", "$60,071 - $68,802", "$68,803 - $77,534", "$77,535 - $86,266", "$86,267 - $94,998", "$94,999 - $103,730", "$103,731 - $112,462", "$112,463 - $121,194", "$121,195 - $129,926", "$129,927 - $138,658", and "$138,659 and above." These brackets of income levels were based on the 2022 U.S. Federal Poverty Level Guidelines [1], which allowed us to group participants' income into categories (e.g., low- vs. moderate-to-high income) for our subsequent analysis. These categorizations account for both income and household size. Specifically, we categorized participants as having low income if their household income adjusted for inflation was equal to or below 185% of the federal poverty level, following the criteria used to determine eligibility for government assistance programs.

3.2.4 Data Analysis. We used a variety of statistical analyses to investigate the relationships between demographics, information practices, and trust and distrust. First, we calculated descriptive statistics for all variables, including means and standard deviations for continuous variables and frequency with percentages for categorical variables. This allowed us to gain a general understanding of the distribution of our variables of interest.

To explore the relationships between variables, we used correlation analyses. We calculated Spearman's correlation matrices for continuous variables and chi-squared tests for categorical variables. To explore the relationship between trust and distrust, we employed factor analysis (using R package nFactors [53]) to group correlated factors together into a few factors. For clustering analysis, we used the Gaussian Mixture Model (GMM) [55] (using R package mclust [61] and Python package sklearn.mixture [49]). The GMM adeptly captures intricate distributions by accommodating multiple Gaussian distributions, since some data points sit ambiguously between distinct patterns. GMM's soft clustering assigns probabilities to these observations, effectively addressing their inherent ambiguity. We then delved into a detailed analysis of the specific group, employing descriptive statistics and a General Linear Model (GLM) to thoroughly examine key demographic variables. Additionally, we aimed to uncover insights into the user's behavioral patterns and habits. We also conducted multiple regression analyses to examine the effects of demographic variables, such as age, gender, and education, on trust and distrust behaviors across different platforms as well as with different misinformation interventions.

To examine differences between groups, we performed independent-sample t-tests and Kruskal-Wallis Rank Sum Test (using R package stats [52]). In cases where significant differences were found, we performed post-hoc tests using Dunnett's test (using the R package FSA [47]) to determine groups that differed significantly. We considered statistical significance at a significance level of $p < 0.05$ and reported effect sizes where applicable to indicate the strength of the relationship between variables.

4 RESULTS

We first investigate the complex dynamics of trust and distrust (subsection 4.1). We then show results regarding the platform differences and demographic differences in trust and distrust in social media (subsection 4.2). After that, we present results related to people's use of misinformation intervention features on social media and how their use of these features is related to their attitudes and trust in social media (subsection 4.3).

4.1 Dynamics of Trust and Distrust in Social Media

4.1.1 Validity and Reliability of Social Media Trust and Distrust Scale (SMTDS). One of the core components of this study was to evaluate trust and distrust in social media. Since the measurement used in our study was a combination of previous work, we assessed the validity and reliability of our measurement with Factor Analysis and Cronbach's α, tests commonly used for multiple Likert questions in a survey that form a scale [67].

The correlation graph (see Figure 3) presents four distinct pairwise correlation matrices, each for Facebook, TikTok, Twitter, and

YouTube. These matrices reveal that within each social media platform, the different aspects of trust are positively correlated with each other. These same patterns are observed for the aspects of distrust, where different aspects of distrust are correlated among each other. However, despite these correlations within trust and distrust dimensions, there is a noticeable separation between the trust and distrust items in the matrices. Specifically, the reason is that we were interested in answers from people who recently experienced the platforms and the features, as this should yield more accurate and reliable perceptions of misinformation intervention features and trust/distrust. This separation implies that trust and distrust are related but distinct constructs: they are interconnected yet different, particularly in how they manifest across various social media platforms.

We used Factor Analysis [58], a technique to group correlated factors into a few factors, to evaluate our survey construct validity. If trust and distrust, based on how we measure them, are two separate constructs, then we would expect the four trust questions to load into one factor and the four distrust questions to load into another factor. The value of Kaiser–Meyer–Olkin (KMO) test is 0.85. The result of Bartlett's test of Sphericity was significant (χ^2 = 15521.21, $df = 28$, $p < .01$), suggesting that there was a substantial correlation in the data that we could summarize using factor analysis. Our scree plot (see Figure 11 in Appendix A) suggests using two factors, with eigenvalues greater than one. Parallel analysis also indicates two factors.

The type of Factor Analysis that we used was Maximum Likelihood Factor Analysis, specifically `factanal` function in R. For rotation, we used `promax` as we did not expect the factors to be totally independent but slightly correlated, as observed in the correlation results. Table 2 shows the results for two factors. We can see that the four trust measurements load into one factor, with factor loading ranging from 0.81 to 0.89, and the other four distrust measurements load into another factor, with factor loading ranging from 0.60 to 0.86. The total variation explained (TVE) is 0.67, which is acceptable based on Hair [21].

Table 2: Factor Analysis results on trust and distrust questions.

Variable	Factor 1 (Trust)	Factor 2 (Distrust)
Trust: Competence	0.89	
Trust: Benevolance	0.86	
Trust: Reliability	0.91	
Trust: Reliance	0.81	
Distrust: Dishonesty		0.86
Distrust: Skepticism		0.60
Distrust: Malevolence		0.87
Distrust: Fear		0.69

For the four-item survey instrument measuring trust in social media, we achieved an excellent Cronbach's alpha of 0.92 on all platforms tested in our study, indicating strong internal consistency. Specifically, Cronbach's alpha for the trust in Facebook was $\alpha = 0.92$, trust in TikTok was $\alpha = 0.91$, trust in Twitter was $\alpha = 0.93$, and trust in YouTube was $\alpha = 0.90$. Similarly, we achieved a good Cronbach's alpha of 0.84 in our four-item survey instrument measuring distrust in social media. Specifically, Cronbach's alpha for the distrust in Facebook was $\alpha = 0.82$, distrust in TikTok was $\alpha = 0.84$, distrust in Twitter was $\alpha = 0.85$, and distrust in YouTube was $\alpha = 0.84$.

Collectively, these results suggested that our constructed survey is valid and reliable in measuring trust and distrust in social media.

4.1.2 Relationship of Trust and Distrust in Social Media. To examine the relationship between trust and distrust in social media, we first tested the correlation between these two concepts, to answer RQ1. If trust and distrust are opposites of the same construct, they would be perfectly or near-perfectly negatively correlated. In other words, we would expect the correlation between trust and distrust to be very close to -1 if they were truly opposites of the same concept. Meanwhile, a strong negative correlation might suggest they are opposites on a continuum; a weak or no correlation may imply they are separate constructs.

Correlation Analysis. Our results show there is a weak but statistically significant negative relationship between trust and distrust with a Pearson correlation coefficient of -0.27 and a Spearman correlation coefficient of -0.23 across all platforms. In detail, we observe correlation coefficients as follows: in TikTok (r = -0.12, $p < 0.05$), in Twitter (r = -0.24, $p < 0.05$), in YouTube (r = -0.25, $p < 0.05$), and in Facebook (r = -0.39, $p < 0.05$). In other words, these results suggest that those who reported trusting TikTok, Twitter, YouTube, and Facebook distrust them less, and vice versa. Yet, while significant, these relationships are far from -1.

The scatterplot in Figure 4 (A) shows a negative correlation, which suggests that for many users, high levels of trust correspond with low levels of distrust, and vice versa, in the context of social media. This supports the conventional wisdom that trust and distrust may be inversely related. However, the presence of data points in the upper-right corner, where both trust and distrust levels are high, indicates that there is a subset of the population for which trust and distrust coexist.

Furthermore, our results also show that data points across the four platforms exhibit similar patterns, as shown in Figure 4 B-E. However, when looking at each platform individually, we observed that a greater number of people exhibit high distrust and low trust in Facebook (see the upper-left-hand corner in each graph), as shown in Figure 4 (B). On the other hand, people display a moderate level of both trust and distrust in YouTube, as shown in Figure 4 (E), evident from the concentrated points in the center of the graph. **Collectively, these results suggest that while trust and distrust can be viewed as related constructs, they can also be distinct**, with additional evidence provided shortly with clustering analysis.

Clustering Analysis. As previously mentioned, our correlation results and factor analysis suggest that trust and distrust could be related but distinct concepts. To further elaborate on and validate this finding, we performed a clustering analysis—a process of grouping data based on the information describing them and their relationships within the data [12]. Through clustering, we can

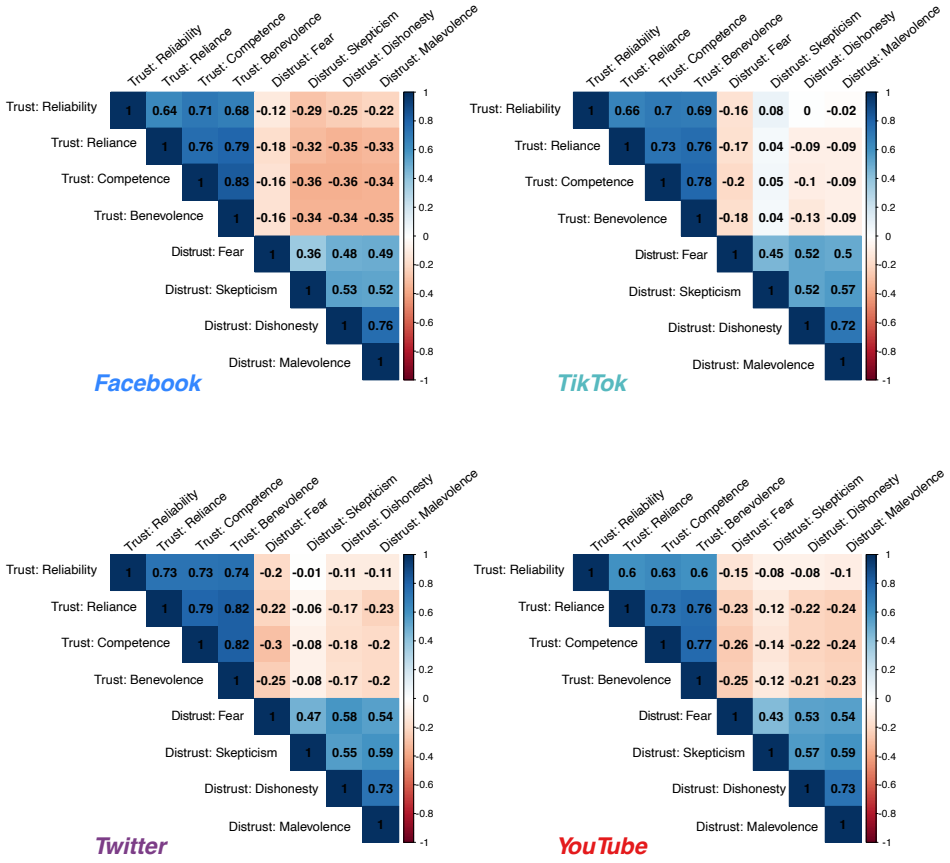

Figure 3: Correlation matrix between all individual items for the trust and distrust scales across four social media platforms (i.e., Facebook, TikTok, Twitter, YouTube). Each matrix shows the correlation coefficients between dimensions of trust (i.e., reliability, reliance, competence, and benevolence) and aspects of distrust (i.e., fear, skepticism, dishonesty, and malevolence).

identify distinct groups of users who may share similar levels of trust and distrust.

To guide us in estimating the appropriate number of clusters, we used Bayesian Information Criterion (BIC) [17]. Our analysis showed that the BIC curve remains relatively flat beyond four clusters, suggesting that using four clusters for the model fit is appropriate (with more details available in Figure 12 in Appendix A Appendices). Therefore, we chose to run our clustering model with four clusters.

The clustering analysis with four clusters (Figure 5) resulted in groups consisting of users with low trust but high distrust and users with high trust and low distrust, indicating that trust and distrust exhibit diametrically opposing behaviors among participants in our study, further suggesting the two extremes continuum.

However, two distinct clusters also emerged in our results, consisting of individuals with both high trust and high distrust (see upper-right corner) and individuals that spanned around the center of the cluster. The Spearman correlation between trust and distrust within the high-trust-high-distrust group is 0.82 , suggesting a strong positive correlation between these variables, indicating a high degree of association between them. When high distrust is present, we find that high trust can also be present, which differs from the green cluster. Excluding the group with high trust and high distrust, the Spearman correlation between trust and distrust stands at -0.52. While this correlation is slightly more negative compared to that observed within the full respondent pool, it remains moderately inverse, suggesting that trust and distrust, although related, may not constitute diametrically opposed constructs on a single continuum. In other words, trust and distrust not only exist at the two extremes of a continuum; instead, they can also be ambivalent, exhibiting a more complex relationship.

The demographic breakdown (see Table 3) of the high-trust-high-distrust group shows that males constituted the majority with 57%, followed by females at 42%, and a small segment (1%) preferring not to answer or identifying as non-binary. The average participant was 37 years old, with the largest age groups being 35-44 (37%) and 25-34 (35%). Ethnicity was predominantly non-Hispanic or Latino (87%), with 40% identifying as African American or Black and 36% as White. Educational attainment varied, with 35% holding a bachelor's degree and 30% having completed high school. In terms of political affiliation, 63% leaned towards the Democratic side. Household income levels were predominantly moderate-to-high (67%) as per the U.S. Federal Poverty Level guidelines.

Table 3: Demographic characteristics of the high-trust-high-distrust group

Demographic	Response Options	Number of Participants (Total N = 246)	Percentage (%)
Gender	Female	104	42%
	Male	141	57%
	Prefer not to answer or Non-binary	1	1%
Age*	18-24	30	12%
	25-34	86	35%
	35-44	90	37%
	45-54	22	9%
	55-64	11	4%
	65+	7	3%
	* Mean = 37, SD = 11, Age range = [18, 79]		
Ethnicity & Race	Hispanic or Latino	31	13%
	Not Hispanic or Latino	215	87%
	African American or Black	99	40%
	Asian	28	11%
	White	88	36%
	Other	31	13%
Education	Less than high school	9	4%
	High school graduate	74	30%
	Associate degree	38	15%
	Bachelor's degree	85	35%
	Postgraduate degree	40	16%
Political Ideology	Democrat/ Lean Democrat	155	63%
	Independent	44	18%
	Republican/ Lean Republican	44	18%
	Other, please describe	3	1%
Household Income	Low-income**	81	33%
	Moderate-to-high income**	164	67%
	Can not be defined**	1	.4%
	** Income levels were determined based on the 2022 U.S. Federal Poverty Level (185%) Guidelines [1].		

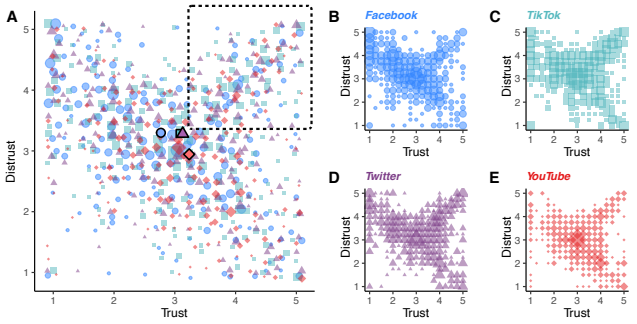

Figure 4: Distribution of trust and distrust among our participants across social media platforms: (A) Four Platforms and their centroids. Data points inside the dotted black square ☐ at the upper-right corner deviate from the linear pattern, indicating elevated levels of both trust and distrust, signaling a particular group of users who simultaneously hold trust and distrust in complex ways. (B) Facebook ●, (C) TikTok ■, (D) Twitter ▲, and (E) YouTube ◆. The larger the data point, the greater the number of people it represents.

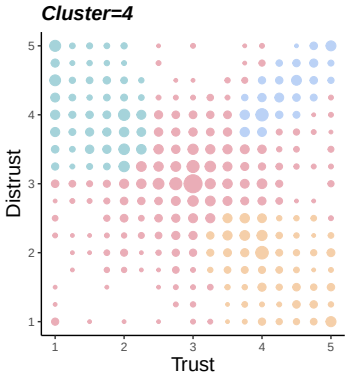

Figure 5: Clustering using the Gaussian Mixture Model with distinct color-coded clusters ●●● (i.e., each color represents a cluster).

To further discern the differences between the high-trust-high-distrust group ● and other groups, we first focus on the demographic patterns observed within this cluster of users ●. To achieve this, we used a generalized linear model, specifically logistic regression, to estimate the probability of membership in the high-trust-high-distrust group ● based on a range of demographic variables. After adjusting for multiple comparisons using the Bonferroni method (see Table 6 in Appendices), we found that age significantly influenced group classification, with a coefficient (log-odds) of -0.06 ($p < 0.001$). This result suggests that the likelihood of being classified as non-high-trust-high-distrust groups increases with age.

Educational attainment was also a significant factor: high school graduates were less likely to be in the high-trust-high-distrust group than those with less than a high school education, indicated by a coefficient of -1.46 ($p < 0.05$). Gender differences also emerged, with females showing lower odds of the high-trust-high-distrust group membership than males, as reflected by a coefficient of -0.88 ($p < 0.05$). Politically, individuals identifying as Independent or Republican were less likely to be in the high-trust-high-distrust group compared to Democrats, with coefficients of -0.93 ($p < 0.05$) and -0.64 ($p < 0.05$), respectively. Our analysis suggests that the remaining variables examined did not significantly affect the probability of being in the high-trust-high-distrust group.

Our analysis also revealed a tendency for younger individuals to be part of the high-trust-high-distrust group. To empirically test this observation, we conducted a non-parametric comparison using Dunn's test, which is appropriate for assessing differences between groups' medians without assuming a normal data distribution. The results confirmed (see Figure 13 in Appendix A), with high statistical significance ($p < 0.05$), that the median age of the high-trust-high-distrust group is substantially lower at 35.5 years, compared to the other groups' median age of 50 years. This finding suggests that high trust and high distrust in social media are characteristics particularly prevalent among younger individuals, notably within the 25-44-year age bracket. Such a demographic pattern emphasizes the need for more research into the social and psychological factors that foster these attitudes toward social media among younger populations.

In our analysis (see Figure 6), we examined the relationship between intervention observation frequency and social media usage within the high-trust-high-distrust group compared to other groups. Dunn's test revealed a significant difference at the 0.001 level between the high-trust-high-distrust group and others in terms of intervention observation. Focusing first on the frequency of intervention observation, within the high-trust-high-distrust group, the median frequency was found to be 3.33, suggesting that individuals in this group sometimes observe the interventions. In contrast, the median for the other group was 2.33, indicating that they rarely notice these interventions. Turning to social media usage, we observed that the median frequency within the high-trust-high-distrust group stood at 4.5, indicating that members of this group typically use social media multiple times a day. On the other hand, the median for the other groups was 4. Overall, we found that the high-trust-high-distrust group was more likely to engage with social media and, relatedly, with the misinformation interventions presented on social media, potentially affecting their perceptions and behaviors in this online environment.

Collectively, our results show that trust and distrust in social media are not simply opposites. Instead, trust and distrust may co-exist, manifesting in complex ways among users, who notably largely engage with social media. These findings point toward a unique demographic and behavioral pattern among the **high-trust-high-distrust group**, particularly among younger individuals, suggesting the need for further research into the factors driving these attitudes toward social media.

4.2 Platform & Demographic Differences in Trust and Distrust in Social Media

4.2.1 Platform Differences in Trust and Distrust in Social Media. To answer RQ2a, we present the results of multiple comparisons between trust and distrust levels across the four social media platforms studied. Specifically, our results show that respondents' trust (χ^2=92.34, $p < 0.05$) and distrust (χ^2=95.82, $p < 0.05$) in social media significantly differ across platforms. Our post-hoc tests (see Figure 7) further reveal that respondents significantly trust TikTok, Twitter, and YouTube more than Facebook ($p < 0.001$). The median trust level between TikTok and Twitter, as well as between Twitter and YouTube were not statistically different ($p = 0.34$, $p = 0.24$). Moreover, we found that respondents trusted YouTube significantly more than TikTok, with the median difference between the two platforms at 0.25 ($p < 0.01$).

As for distrust, our results showed significant differences between Facebook, Twitter, and TikTok with YouTube, respectively. The results showed that participants exhibit lower distrust towards YouTube than other platforms tested in our study. More precisely, the distrust difference is 0.25 between Facebook and YouTube ($p < 0.001$), Twitter and YouTube ($p < 0.001$), as well as TikTok and YouTube ($p < 0.001$). Comparisons between other platforms (Facebook-Twitter, Facebook-TikTok, Twitter-TikTok) did not present significant results in our study. More details are available in Table 7 in Appendix A Appendices.

In short, when comparing social media platforms on trust and distrust levels, we found that Facebook was significantly less trusted and YouTube was significantly less distrusted.

4.2.2 Demographic Differences in Trust and Distrust in Social Media. To answer RQ2b, we presented results of multiple regression models, examining how respondents' trust and distrust in Facebook, TikTok, Twitter, and YouTube is associated with their age, education, gender, income, race, and ethnicity, as well as political ideology.

Demographic Differences in Trust. Table 4 reveals a consistent pattern where age, coded as a continuous numerical variable, is inversely related to trust across all social media platforms at a high significance level ($p < 0.001$). For instance, with each incremental year in age, trust in Facebook diminishes by $\beta = -0.016 \pm 0.002$, in TikTok by $\beta = -0.020 \pm 0.003$, in Twitter by $\beta = -0.019 \pm 0.003$, and in YouTube by $\beta = -0.008 \pm 0.002$.

Regarding education, those with higher educational attainment reported lower trust in Facebook compared to those with less than a high school degree, with a bachelor's degree, in particular, presenting a significant decrease in trust ($\beta = -0.662$, $p < 0.05$). The association between education and trust did not reach significance for other platforms, indicating a nuanced impact of education on trust in social media.

In terms of race and ethnicity, the findings are mixed. Black participants showed significantly higher trust in Facebook ($\beta = 0.354$, $p < 0.01$) and YouTube ($\beta = 0.413$, $p < 0.001$) compared to White participants. Other racial comparisons did not yield significant results. In terms of political ideology, Independents showed significantly lower trust in YouTube ($\beta = -0.318$, $p < 0.001$), and those identifying with political affiliations other than Democrat or Republican expressed markedly lower trust in YouTube ($\beta = -1.061$, $p < 0.001$). Gender differences were not statistically significant in

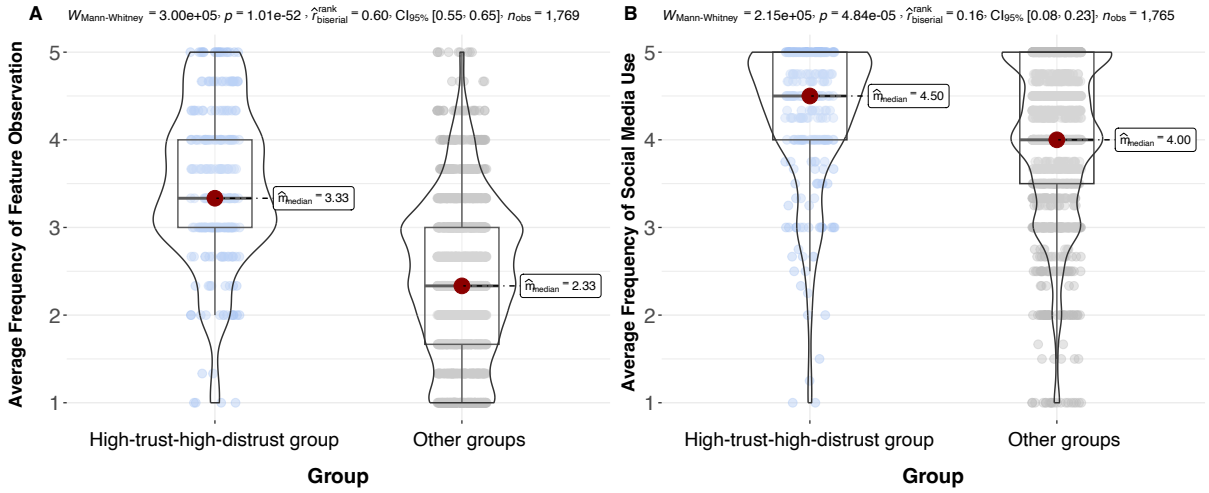

Figure 6: Non-parametric pairwise comparison test (Dunn's test) that shows the differences in an average frequency of (A) feature observation and (B) social media use across different groups.

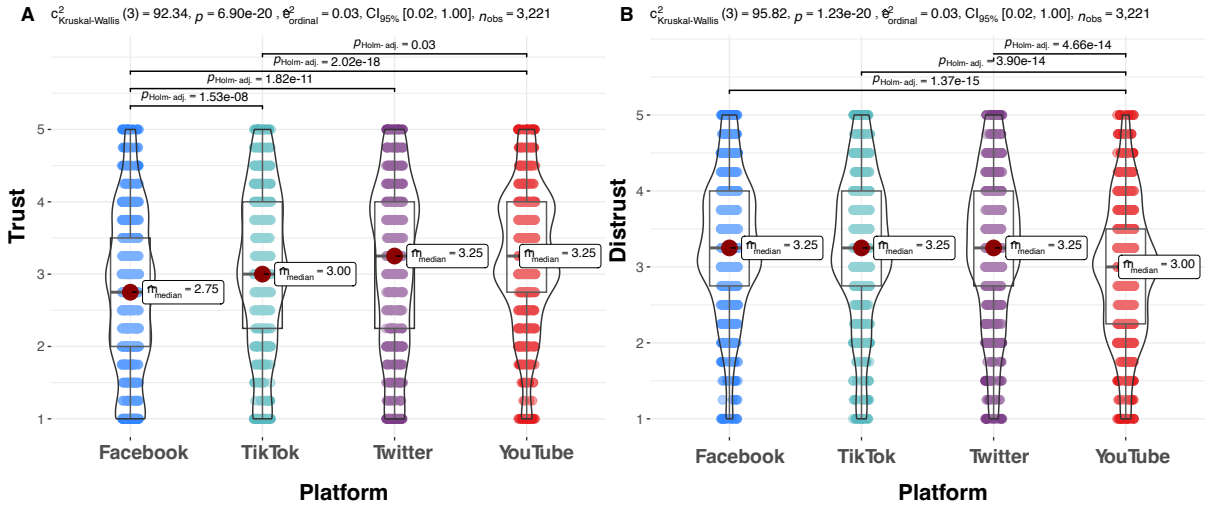

Figure 7: Non-parametric pairwise comparison test (Dunn's test): Differences in (A) trust and (B) distrust levels across different social media platforms.

respondents' trust in social media. Likewise, no significant relationship was observed between income levels and trust in social media.

Demographic Differences in Distrust.

Table 5 presents the results of multiple regression models, examining how respondents' distrust in Facebook, TikTok, Twitter, and YouTube is associated with their age, education, gender, income, race and ethnicity, and political ideology.

Age appears to have a significant negative correlation with distrust in YouTube ($\beta = -0.006$, $p < 0.01$), indicating that older individuals tend to distrust YouTube less. No other significant age-related effects are observed for distrust across Facebook, TikTok, or Twitter. Education does not appear to be significantly associated with distrust in any social media platforms at the $p < 0.05$ level. Gender differences in distrust toward social media were not statistically significant, suggesting that distrust does not vary markedly between genders. In terms of income, no significant effects were observed, indicating that income levels do not play a substantial role in the level of distrust in social media. When examining race and ethnicity, significant findings include that Hispanic respondents exhibit less distrust in Facebook ($p < 0.05$), and Black respondents show less distrust in TikTok and Twitter ($p < 0.05$). Additionally, respondents from 'Other' racial groups indicate significantly higher levels of distrust in Twitter ($p < 0.001$). We also found that Republican or Lean Republican respondents have significantly greater distrust in Twitter than Democrat or Lean Democrat respondents ($p < 0.001$). No other significant differences were noted across the political ideologies for the remaining social media platforms.

Table 4: Multiple regression models explaining respondents' trust in social media (Significance level: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Overall, our results show that older individuals generally trust social media less, and those with higher education also exhibit lower trust, especially for Facebook. Black respondents tend to have higher trust than White respondents, and political affiliation significantly influences trust levels in social media.

	Dependent Variable			
	Trust in Facebook β (Std. Error)	Trust in TikTok β (Std. Error)	Trust in Twitter β (Std. Error)	Trust in YouTube β (Std. Error)
const	4.380 (0.256)***	4.512 (0.318)***	3.619 (0.402)***	4.280 (0.236)***
Age	-0.016 (0.002)***	-0.020 (0.003)***	-0.019 (0.003)***	-0.008 (0.002)***
Education				
(*Reference: Less than high school*)				
High school	-0.543 (0.215)	-0.682 (0.278)	0.508 (0.376)	-0.501 (0.199)
Associate	-0.532 (0.224)	-0.579 (0.284)	0.372 (0.383)	-0.354 (0.205)
Bachelor	-0.662 (0.223)*	-0.511 (0.285)	0.520 (0.382)	-0.419 (0.205)
Postgraduate	-0.644 (0.229)	-0.428 (0.295)	0.596 (0.388)	-0.567 (0.211)
Gender				
(*Reference: Male*)				
Female	-0.142 (0.068)	-0.181 (0.083)	-0.169 (0.084)	-0.166 (0.063)
Prefer not to answer or Non-binary	-0.098 (0.463)	-0.399 (0.405)	-0.904 (0.456)	-1.065 (0.363)
Income				
(*Reference: Low income*)				
Moderate-to-high income	-0.126 (0.086)	-0.094 (0.100)	-0.149 (0.109)	-0.080 (0.079)
Race & Ethinicity				
(*Reference: Non Hispanic*)				
Hispanic	0.112 (0.110)	-0.136 (0.110)	-0.181 (0.126)	0.081 (0.098)
(*Reference: White*)				
Asian	0.269 (0.108)	-0.328 (0.144)	-0.093 (0.145)	0.068 (0.100)
Black	0.354 (0.099)**	0.272 (0.104)	0.275 (0.109)	0.413 (0.089)***
Other	0.032 (0.100)	0.143 (0.118)	0.089 (0.129)	0.079 (0.089)
Political Ideology				
(*Reference: Democrat / Lean Democrat*)				
Republican / Lean Republican	-0.177 (0.084)	-0.103 (0.103)	-0.168 (0.109)	-0.181 (0.081)
Independent	-0.199 (0.082)	-0.256 (0.098)	-0.650 (0.103)	-0.318 (0.075)***
Other	-0.460 (0.233)	-0.356 (0.342)	-0.113 (0.341)	-1.061 (0.217)***

Collectively, our results show associations between demographic factors—such as age, education, gender, race and ethnicity, and political ideology—and levels of trust and distrust in social media. The strength and nature of these associations can vary across platforms.

4.3 Perceptions of and Experiences with Misinformation Interventions

4.3.1 Relationship Between Experiences with Misinformation Intervention Features and People's Attitudes. 91% of our respondents reported seeing misinformation intervention features on social media. Of these, 54% frequently saw the labeling feature, 56% saw the curation feature, and 56% noticed the verification feature.

Among those who had seen the misinformation interventions, we examined the relationship between participants' experiences with these interventions and their awareness of misinformation (Figure 8A), likelihood to share posts from social media (Figure 8B), and likelihood to receive more information from the social media platform (Figure 8C).

Of those who had seen the misinformation interventions, 71% agreed or strongly agreed that labeling increased their awareness of misinformation on social media, while 8% disagreed or strongly disagreed. In addition, 61% and 55% agreed or strongly agreed that the curation feature and the verification feature increased their awareness of misinformation, respectively. A small percentage (10% on curation and 14% on verification) disagreed that these features increased their awareness of the matter.

Furthermore, 31% of participants agreed or strongly agreed that they were more likely to share information from social media with the labeling feature, with a larger proportion of individuals disagreeing (22%) or strongly disagreeing (15%). In comparison, curation and verification features were more likely to make participants want to share information from social media, with 45% and 41% agreeing or strongly agreeing, respectively.

Finally, regarding the likelihood of receiving information from social media, 50% of respondents agreed or strongly agreed that curation features influenced them, while only 40% and 42% of respondents felt the same regarding labeling and verification features, respectively. Concurrently, 27% of respondents reported that they disagreed or strongly disagreed that labeling made them want to

Table 5: Multiple regression models explaining respondents' distrust in social media. (Significance level: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$) The analysis suggests that distrust in social media varies less with age but is significantly influenced by race (with Hispanic respondents showing lower distrust in Facebook, and Black respondents displaying lower distrust on TikTok and Twitter) and political ideology (where Republicans exhibit less distrust in Twitter).

	Dependent Variable			
	Distrust in Facebook β (Std. Error)	Distrust in TikTok β (Std. Error)	Distrust in Twitter β (Std. Error)	Distrust in YouTube β (Std. Error)
const	3.285 (0.223)***	3.851 (0.288)***	4.112 (0.352)***	3.028 (0.236)***
Age	-0.001 (0.002)	-0.002 (0.003)	0.001 (0.002)	-0.006 (0.002)**
Education				
(*Reference: Less than high school*)				
High school	0.096 (0.187)	-0.367 (0.251)	-0.495 (0.329)	0.403 (0.197)
Associate	0.226 (0.195)	-0.303 (0.257)	-0.264 (0.336)	0.333 (0.204)
Bachelor	0.330 (0.199)	-0.278 (0.258)	-0.320 (0.335)	0.399 (0.204)
Postgraduate	0.294 (0.059)	-0.226 (0.267)	-0.476 (0.340)	0.461 (0.210)
Gender				
(*Reference: Male*)				
Female	-0.070 (0.367)	-0.075 (0.075)	-0.162 (0.074)	-0.053 (0.063)
Prefer not to answer or Non-binary	0.301 (0.075)	0.131 (0.367)	0.342 (0.400)	0.468 (0.367)
Income				
(*Reference: Low income*)				
Moderate-to-high income	0.039 (0.075)	0.038 (0.091)	-0.118 (0.095)	-0.085 (0.079)
Race & Ethinicity				
(*Reference: Non Hispanic*)				
Hispanic	-0.292 (0.096)*	0.034 (0.100)	-0.011 (0.110)	0.186 (0.099)
(*Reference: White*)				
Asian	-0.107 (0.094)	0.129 (0.131)	0.044 (0.127)	0.056 (0.101)
Black	-0.110 (0.086)	-0.288 (0.095)*	-0.298 (0.095)*	-0.189 (0.090)
Other	-0.065 (0.087)	-0.227 (0.107)	-0.450 (0.113)***	-0.159 (0.090)
Political Ideology				
(*Reference: Democrat / Lean Democrat*)				
Republican / Lean Republican	-0.026 (0.073)	-0.094 (0.094)	-0.430 (0.095)***	-0.030 (0.082)
Independent	-0.160 (0.072)	-0.051 (0.088)	-0.158 (0.090)	-0.061 (0.076)
Other	0.147 (0.202)	0.203 (0.310)	0.153 (0.299)	0.012 (0.214)

receive more information, 19% felt similarly on curation, and 22% on verification.

In summary, while most of our participants agreed that misinformation interventions heightened their awareness of misinformation, many participants were neutral or disagreed that these interventions enhanced their likelihood to share and receive information. Nonetheless, individuals in our study were more inclined to share and receive information on platforms that employ curation features than any other misinformation interventions, whereas the labeling feature made our participants become more aware of misinformation than other features.

4.3.2 Relationship between Misinformation Intervention Features and Trust and Distrust in Social Media. The correlation matrices show the interplay between trust and distrust among social media users under various intervention strategies: Labeling, Curation, and Verification (see Figure 9). Trust dimensions (Reliance, Benevolence, Competence, Reliability) consistently demonstrate moderate to strong positive correlations with one another, underscoring a cohesive construct of trust. For example, users who perceive a platform as benevolent also tend to regard it as competent. In terms of distrust (Skepticism, Malevolence, Dishonesty, Fear), there are also positive correlations, particularly strong between Skepticism and Malevolence, suggesting that users skeptical of the platform's intent may also view it as malevolent. Notably, the correlations between trust and distrust dimensions are generally weak, indicating that these constructs may operate independently; an increase in trust does not necessarily equate to a decrease in distrust.

As shown in Figure 10, our respondents indicated levels of trust in the labeling, curation, and verification features used to address misinformation on social media platforms. Participants' trust in social media platforms was assessed by their agreement with the platforms' anti-misinformation efforts. Labeling features had an average trust score of 3.50 ($SD = 0.93$), curation features scored slightly higher at 3.51 ($SD = 0.92$), and verification features came in at 3.43 (SD = 0.92). This indicates a general trust in these features, with Curation slightly more trusted than Labeling and Verification. The small standard deviations suggest that most respondents consistently agree that these features show the platforms' commitment to reducing misinformation and enhancing user engagement.

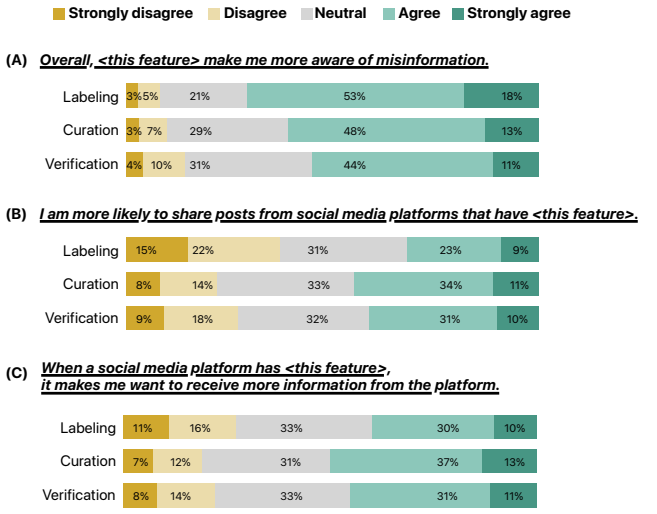

Figure 8: Participants' prior experiences with misinformation intervention features and their attitudes, including awareness of misinformation, likelihood of sharing information from social media, and intention to receive information from social media.

Conversely, distrust across all intervention features is moderate, with average scores indicating a neutral attitude—neither strong agreement nor disagreement. Mean distrust scores are 3.08 (SD = 0.89) for labeling, 3.04 (SD = 0.93) for curation, and 3.01 (SD = 0.94) for verification, showing consistent skepticism across interventions. This uniformity suggests that users are cautiously skeptical about the platforms' anti-misinformation efforts, questioning whether platforms are effectively combating misinformation or acting in their own interests.

4.3.3 Demographic Differences. Table 8 in Appendix A reports the factors related to respondents' trust in social media interventions, with a focus on the Labeling, Curation, and Verification features. Age is inversely related to trust across the Labeling, Curation, and Verification features; specifically, for each incremental year in age, trust decreases by $\beta = -0.0102 \pm 0.0013$ for Labeling, $\beta = -0.0099 \pm 0.0013$ for Curation, and $\beta = -0.0088 \pm 0.0013$ for Verification, all with a significance level of $p < 0.001$. Gender differences are also pronounced; non-binary respondents exhibit significantly less trust in the Curation feature with coefficient value β = -0.7003 ($p < 0.05$). Likewise, racial and ethnic disparities are evident; Black respondents show a higher level of trust in Labeling ($\beta = 0.1770, p < 0.05$) and Curation features ($\beta = 0.1868, p < 0.05$). Political affiliations also reveal a strong correlation, where, for instance, Republicans demonstrate significantly less trust in the Verification features with coefficient value β = -0.3669 ($p < 0.001$). Education and income show less pronounced effects on trust levels, lacking significant correlations with trust, suggesting that education levels and economic status do not play major roles in trust towards social media interventions.

Similar to trust, from Table 9 (in Appendices), we can see that age is a significant factor in distrust but with a smaller effect. For example, for Curation, each additional year in age is associated with a β = -0.0069 ($p < 0.001$), and for Verification, the corresponding β = -0.0055 ($p < 0.001$). Gender differences are also seen, with females exhibiting a greater tendency towards distrust in social media misinformation interventions. In the Curation task, the β = -0.1673 ($p < 0.001$); for Verification, the β = -0.1600 ($p < 0.01$); and for Labeling, the β = -0.1322 ($p < 0.1$). This suggests that female respondents in our study generally have less distrust of these misinformation interventions than males. Meanwhile, other demographic variables such as education, income, and race do not demonstrate consistent patterns of significance in influencing distrust.

5 DISCUSSION

Our analysis attempted to tease out the complex relationship between trust and distrust and how this relationship differs demographically and across social media platforms. Meanwhile, we contributed a set of validated survey scales to measure trust and distrust that future research can use. Collectively, we contribute to further theorizing the dynamics of trust and distrust. Furthermore, we examined people's perceptions of and experiences with misinformation intervention and how their prior experiences with misinformation intervention may influence their trust and distrust in social media. Building upon our results, we discuss the significance of theorizing the relationship between trust and distrust, as well as the practical and design implications arising from our work.

5.1 The Significance of Theorizing the Relationship between Trust & Distrust in Social Media

Through an empirical study, we investigated the nuanced relationship between trust and distrust in the realm of social media. Our results showed weak yet statistically significant negative correlations between trust and distrust in social media (e.g., Facebook, Twitter, TikTok, and YouTube). These results suggest that the dynamics of trust and distrust might coexist within our participants. This multifaceted interplay offers a fresh perspective in understanding these constructs.

5.1.1 Dynamics of Trust & Distrust in Social Media: In this paper, we presented analyses to showcase the complex relationship that exists between trust and distrust, and further shed light on new research opportunities to explore the heterogeneity of "trust and distrust profiles" among users. For example, our clustering analysis shows that users have diverse trust dynamics, and this variation is crucial to understanding their attitudes and behavior towards social media usage. Specifically, our results show an obvious cluster of people with high trust and high distrust. This finding may indicate that while users might trust specific aspects or functionalities of a platform, they simultaneously remain wary or distrustful of other elements of the platform. Such coexistence of high trust and distrust could arise from users discerning between the credibility of information sources (i.e., other users on the platform) and the reliability of the platform to continue providing trustworthy information (i.e., misinformation interventions). This pattern may also indicate that many users, rather than being universally skeptical or trusting,

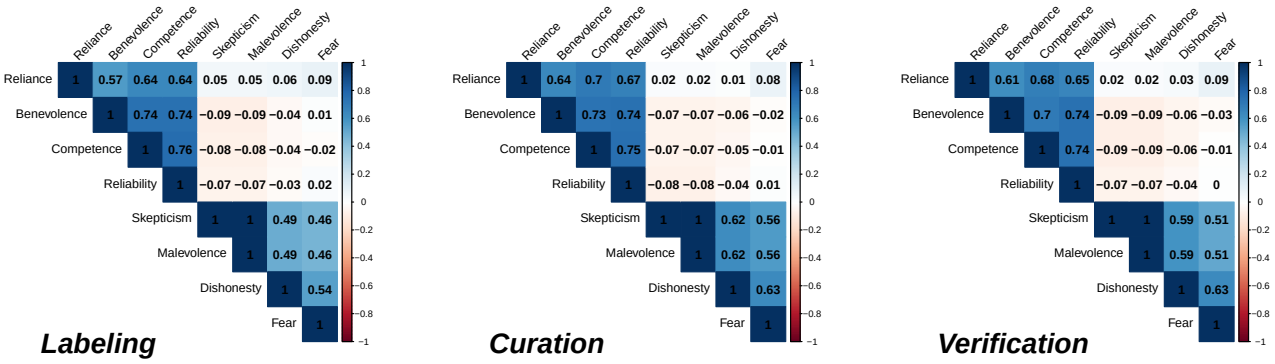

Figure 9: Correlation matrix for the trust and distrust with different misinformation interventions.

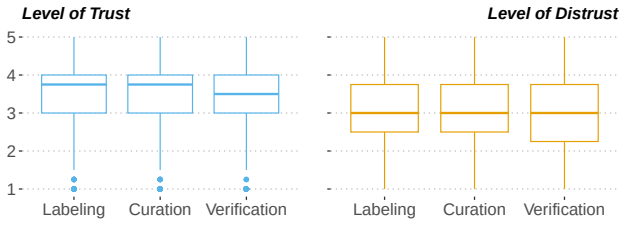

Figure 10: Respondents' average trust and distrust in social media with different misinformation intervention features.

have a much more nuanced view of their online experience. Results from our analyses also help future researchers reconsider the use of a single trust-distrust Likert scale, instead utilizing a set of valid and reliable questions that represent trust and distrust separately, as presented in our SMTDS (see subsubsection 4.1.1).

Nonetheless, our clustering analysis did not reveal a low trust, low distrust group. This may be attributed to our recruitment strategy, where we focus on participants who are currently using social media. We assume that those who use social media, possess some trust in the platform and thus, continue using the platform. If we were to recruit participants who did not currently use social media, we might be able to identify the low trust and low distrust cluster in our results. Nevertheless, we suggest future work continue to dig into the nuanced characteristics of social media users and define their placement within specific trust and distrust clusters. Some additional user characteristics worth examining may include content consumption preferences, external information sources [27], and psychological traits [6] (e.g., risk aversion or propensity for skepticism can also influence trust dynamics). By dissecting user characteristics, we can better understand the diverse trust and distrust landscapes on social media platforms.

Additionally, our findings provide several trajectories for future theoretical work. First, the lack of a universally accepted definition of distrust in the academic community presents challenges in operationalizing this concept. Our study examines this understudied topic of distrust. While we aimed to explore the boundaries of distrust in the social media context, we recognize that this approach may limit the scope of our findings. Moving forward, we suggest future research to dive deeper into the nuances of distrust, exploring its various facets in different contexts. By expanding the definition and measurement of distrust, subsequent studies can offer a more holistic understanding of how trust and distrust coexist and interact in different environments. Such research will complement our findings and contribute to the discourse on trust dynamics.

Second, future work should work to develop theoretical models that encapsulate the various trust-distrust profiles. These models would serve as a foundational framework, elucidating how and why certain profiles form, their stability over time, and their responsiveness to external stimuli or platform changes. Another avenue for future research is to examine the temporal evolution of these profiles. Key questions to explore may include: Do these profiles remain consistent over time? Or do they shift due to broader societal shifts, individual experiences, or changes made to the platforms? Examining the temporal dynamics of trust profiles provides academic insights and has practical implications in shaping users' experiences, informing platform design, and understanding societal shifts.

5.1.2 Situated Perspective in Understanding Trust & Distrust in the Social Media Misinformation Age: What further adds a layer of complexity is the backdrop against which these relationships exist: the context of misinformation. Our findings underscore the growing relevance and importance of misinformation intervention features. A striking 91% of respondents reported encountering such features on their chosen social media platforms. This finding suggests the pervasive nature of misinformation on social media and the subsequent efforts made by these companies to combat it. While these features are being encountered, the frequency for which they are encountered is sparse, as our results showed that many users still either rarely or never see these features. This might indicate that these features are not uniformly distributed or that they only activate under certain conditions or algorithms, suggesting areas for potential improvement in deployment.

Our results also empirically showed that these misinformation intervention features raised the awareness of misinformation on the platform. While, in theory, this could potentially decrease trust, our results showed that the features actually increased trust on average instead. A user who inherently distrusts a platform might view its fact-checking interventions skeptically, whereas one who trusts the platform may perceive it as a seal of authenticity. Simultaneously, those who inhabit the gray zone of both trust and distrust may weigh these interventions differently, oscillating between acceptance and doubt. Our results showed that the labeling features offer users a direct way of discerning potentially harmful information and that a significant 71% felt it increased their awareness. This suggests that immediate and visually recognizable cues are crucial for aiding user judgment in information consumption. However, the fact that curation—a proactive and possibly more nuanced approach to misinformation—is associated with a higher likelihood of sharing and receiving information is noteworthy. This might indicate that users appreciate and trust pre-vetted content aggregated for accuracy and relevance.

Collectively, by situating our understanding of trust and distrust within the context of misinformation and its countermeasures, we deepen our theoretical understanding and provide insights that could shape the future of digital information design and dissemination. Our findings highlight the importance of continuous efforts and innovations to preserve the integrity of digital information spaces to maintain user trust and address issues of distrust. Looking ahead, specific avenues of investigation warrant attention in future research. For example, as social media platforms evolve, they will undoubtedly introduce a new generation of misinformation countermeasures. As such, a longitudinal study that tracks the efficacy of these interventions over time, alongside shifts in user trust and distrust, could be crucial.

5.2 Practical & Design Implications

We provide practical implications in real-world contexts in multiple ways, ranging from designing tools that enable people to navigate areas of skepticism and distrust to implications regarding policy and regulations.

Our results provide initial evidence regarding different types of dual trust-distrust profiles, suggesting that understanding the feedback loop between user-generated content and trust and distrust dynamics could be interesting. How do certain types of content being either flagged or endorsed shape users' trust, as well as their distrust in the platform and their subsequent content creation or sharing behaviors? Answering these questions will shed light on the complex interplay between content consumption, content creation, and the evolving perceptions of users. Therefore, to dig into the underlying reasons for forming the high-trust-high-distrust group, qualitative follow-up studies such as interviews can help unpack the nuances behind their simultaneous trust and distrust in social media, providing rich, contextual insights that quantitative data alone cannot reveal. This knowledge can guide platforms in designing more effective interventions to combat misinformation and foster a more trusting and informed user community. Moreover, it could influence content moderation strategies, ensuring a healthier digital ecosystem and more responsible user engagement.

Additionally, given our findings concerning the nuances across demographics, design indicators may resonate with specific cultural or contextual sentiments, such as local endorsements, regional checks, or community-driven verifications. In our study, we found significant differences in trust levels in social media platforms across political ideologies. Specifically, in our study, Independents exhibited significantly lower trust in YouTube, while those with political affiliations other than Democrat or Republican showed even lower trust in this platform. Regarding Twitter, Republican or Lean Republican respondents showed significantly greater distrust than their Democrat or Lean Democrat counterparts. Additionally, Republicans demonstrated significantly less trust in the Verification features of social media platforms. Our findings contribute to the existing body of research on how political affiliations shape interactions with online content, particularly in the context of misinformation, moderation, and trust. For example, Sharevski et al.'s work [76] suggested that Republicans and Independents were more likely to perceive misleading tweets as "somewhat accurate" compared to Democrats, who view them as "not very accurate," which aligns with our observation of varying trust levels across political ideologies. In another work, Zannettou [62] has found that most tweets with warning labels are shared by Republicans, while Democrats are more engaged in commenting on these tweets. While this work examined the relationship between user engagement (e.g., sharing and commenting) and political ideology, our work specifically focused on trust and distrust in social media. Collectively, these findings highlight the importance of acknowledging and engaging with the nuanced perceptions that characterize different subpopulations. These insights also suggest a tailored approach to designing and implementing platform moderation strategies that are informed by an understanding of the diverse and complex landscape of user trust.

Our findings also echo a recent review paper focused on misinformation interventions [2]. The authors argue that existing misinformation interventions have primarily focused on individualistic approaches, ignoring community factors, such as the role of social norms [37, 57]. Therefore, to ensure the efficacy of future intervention designs in the realm of misinformation, it is important to integrate both individual and community-based perspectives, anchoring them in the diverse sociocultural contexts of the user base.

Last but not least, policymakers and regulators might also benefit from our work. Instead of drafting policies that singularly focus on enhancing trust, it might be equally crucial to devise strategies that address sources of distrust. For example, a more comprehensive regulatory framework, which promotes trustworthy practices while curbing elements that seed distrust, is essential for fostering a robust online information ecosystem. Another direction to move forward is collaborative policy drafting [63]. For example, policymakers could collaborate with social media platforms, content creators, and users to draft regulations. Such a collaborative approach ensures that policies resonate with real-world challenges and user sentiments. Additionally, we suggest that future research could pioneer the concept of "distrust audits." Similar to how platforms undergo privacy or security evaluations, these audits would systematically assess features or areas within a platform that might

induce user skepticism. By identifying and addressing these potential pitfalls, platforms may proactively cultivate a trustworthy digital environment.

6 LIMITATIONS

While our empirical study provides valuable insights into the effects of misinformation interventions on people's trust in social media, some limitations should be acknowledged. First, our study was conducted in the United States, limiting our findings' generalizability to other countries or cultural contexts. Future work should expand the scope of studied populations to include participants from other countries to better understand how misinformation interventions on social media influence people's trust and distrust across different cultures and societies. In addition, we acknowledge that the trust and distrust scales used in our study require ecological validation to ensure their reliability and effectiveness in other real-world settings (e.g., diverse social contexts and different populations). Moreover, our study focused primarily on a subset of visible misinformation features and several main social media platforms. Our work did not explicitly examine other types of interventions that social media platforms use. We also noted a limitation in the study design regarding the evaluation of trust and distrust. Our survey captured respondents' perceptions of their experience with platforms' already-deployed misinformation intervention features, but did not contrast these with a baseline from platforms without such measures. To further determine the effects of misinformation interventions, future work may consider conducting experimental design studies for comparative analysis. Therefore, future work can further study the effects of these other misinformation interventions on people's trust and distrust in social media and expand the scope of the social media platforms examined.

7 CONCLUSION

Our extensive research, conducted through a large-scale survey involving 1,769 participants in the U.S., has revealed several crucial insights into the dynamics of trust and distrust in social media. Our results show that trust and distrust can be two concepts rather than the two sides of a singular spectrum. This dual lens enriches our theoretical understanding of online trust dynamics. Our findings further classify users based on varied trust and distrust intensities. Moreover, we highlight that both trust and distrust perceptions can shift depending on the platform and are influenced by demographic factors. Additionally, while misinformation interventions can elevate users' misinformation awareness and bolster trust in platforms, they don't necessarily reduce distrust. Our research suggests that focusing solely on trust is insufficient; rather, distrust can be regarded as a distinct concept that requires dedicated attention in the future.

ACKNOWLEDGMENTS

This material is based on work that is funded by an unrestricted gift from Google. We thank our anonymous reviewers for their reviews.

REFERENCES

[1] Assistant Secretary for Planning and Evaluation (ASPE). 2022. 2022 Poverty Guidelines: 48 Contiguous States. https://aspe.hhs.gov/sites/default/files/documents/4b515876c4674466423975826ac57583/Guidelines-2022.pdf

[2] Zhila Aghajari, Eric PS Baumer, and Dominic DiFranzo. 2023. Reviewing Interventions to Address Misinformation: The Need to Expand Our Vision Beyond an Individualistic Focus. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–34. https://doi.org/10.1145/3579520

[3] Mabrook S Al-Rakhami and Atif M Al-Amri. 2020. Lies kill, facts save: detecting COVID-19 misinformation in twitter. *Ieee Access* 8 (2020), 155961–155970. https://doi.org/10.1109/ACCESS.2020.3019600

[4] Joseph B Bak-Coleman, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. 2022. Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour* 6, 10 (2022), 1372–1380. https://doi.org/10.1038/s41562-022-01388-6

[5] Megan Boler. 2008. *Digital media and democracy: Tactics in hard times.* Mit Press.

[6] Tom Buchanan and Vladlena Benson. 2019. Spreading disinformation on facebook: Do trust in message source, risk propensity, or personality affect the organic reach of "fake news"? *Social media+ society* 5, 4 (2019), 2056305119888654. https://doi.org/10.1177/2056305119888654

[7] Celeste Campos-Castillo and Denise Anthony. 2019. Situated trust in a physician: Patient health characteristics and trust in physician confidentiality. *The Sociological Quarterly* 60, 4 (2019), 559–582. https://doi.org/10.1080/00380253.2018.1547174

[8] Yang Cheng and Zifei Fay Chen. 2020. Encountering misinformation online: antecedents of trust and distrust and their impact on the intensity of Facebook use. *Online Inf. Rev.* 45 (2020), 372–388. https://api.semanticscholar.org/CorpusID:229422065

[9] Jinsook Cho. 2006. The mechanism of trust and distrust formation and their relational outcomes. *Journal of retailing* 82, 1 (2006), 25–35.

[10] Delonia Cooley and Rochelle Parks-Yancy. 2019. The effect of social media on perceived information credibility and decision making. *Journal of Internet Commerce* 18, 3 (2019), 249–269. https://doi.org/10.1080/15332861.2019.1595362

[11] Rachel Croson and Nancy Buchan. 1999. Gender and culture: International experimental evidence from trust games. *American Economic Review* 89, 2 (1999), 386–391.

[12] D. Divjak and N. R. J. Fieller. 2014. Cluster analysis: Finding structure in linguistic data. https://api.semanticscholar.org/CorpusID:57469984

[13] Greg Elmer. 2017. Precorporation: or what financialisation can tell us about the histories of the Internet. *Internet Histories* 1, 1-2 (2017), 90–96. https://doi.org/10.1080/24701475.2017.1308197

[14] Gunther Eysenbach and Christian Köhler. 2002. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *Bmj* 324, 7337 (2002), 573–577.

[15] Andrew J Flanagin and Miriam J Metzger. 2000. Perceptions of Internet information credibility. *Journalism & mass communication quarterly* 77, 3 (2000), 515–540.

[16] Joshua Fogel and Elham Nehmad. 2009. Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in human behavior* 25, 1 (2009), 153–160. https://doi.org/10.1016/j.chb.2008.08.006

[17] Chris Fraley and Adrian Raftery. 2007. Model-based methods of classification: using the mclust software in chemometrics. *Journal of Statistical Software* 18 (2007), 1–13.

[18] David Gefen, Elena Karahanna, and Detmar W Straub. 2003. Trust and TAM in online shopping: An integrated model. *MIS quarterly* 27, 1 (2003), 51–90. https://doi.org/10.2307/30036519

[19] Carolin Gerlitz and Anne Helmond. 2013. The like economy: Social buttons and the data-intensive web. *New media & society* 15, 8 (2013), 1348–1365. https://doi.org/10.1177/1461444812472322

[20] Ashish Goel and Latika Gupta. 2020. Social media in the times of COVID-19. *Journal of clinical rheumatology* (2020). https://doi.org/10.1097/RHU.0000000000001508

[21] Joseph F Hair. 2009. Multivariate data analysis. (2009).

[22] Mark A Hall, Elizabeth Dugan, Beiyao Zheng, and Aneil K Mishra. 2001. Trust in physicians and medical institutions: what is it, can it be measured, and does it matter? *The milbank quarterly* 79, 4 (2001), 613–639.

[23] Alison Hamilton. 2013. Qualitative methods in rapid turn-around health services research.

[24] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.

[25] Russell Hardin. 2002. *Trust and trustworthiness.* Russell Sage Foundation.

[26] Marc J Hetherington. 2005. *Why trust matters: Declining political trust and the demise of American liberalism.* Princeton University Press.

[27] Itai Himelboim, Ruthann Weaver Lariscy, Spencer F Tinkham, and Kaye D Sweetser. 2012. Social media and online political communication: The role of interpersonal informational trust and openness. *Journal of Broadcasting & Electronic Media* 56, 1 (2012), 92–115. https://doi.org/10.1080/08838151.2011.648682

[28] Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media.

[29] Petros Iosifidis and Nicholas Nicoli. 2020. The battle to end fake news: A qualitative content analysis of Facebook announcements on how it combats disinformation. *International Communication Gazette* 82, 1 (2020), 60–81. https://doi.org/10.1177/1748048519880729

[30] Will Jennings, Gerry Stoker, Hannah Bunting, Viktor Orri Valgarðsson, Jennifer Gaskell, Daniel Devine, Lawrence McKay, and Melinda C Mills. 2021. Lack of trust, conspiracy beliefs, and social media use predict COVID-19 vaccine hesitancy. *Vaccines* 9, 6 (2021), 593.

[31] Alexandra D Kaplan, Theresa T Kessler, J Christopher Brill, and PA Hancock. 2021. Trust in artificial intelligence: Meta-analytic findings. , 00187208211013988 pages. https://doi.org/10.1177/00187208211013988

[32] Dan J Kim, Donald L Ferrin, and H Raghav Rao. 2009. Trust and satisfaction, two stepping stones for successful e-commerce relationships: A longitudinal exploration. *Information systems research* 20, 2 (2009), 237–257.

[33] Kyung-Sun Kim, Sei-Ching Joanna Sin, and Tien-I Tsai. 2014. Individual differences in social media use for information seeking. *The journal of academic librarianship* 40, 2 (2014), 171–178. https://doi.org/10.1016/j.acalib.2014.03.001

[34] John Koetsier. 2020. Reddit, Facebook, Twitter Worst For Mental Health Post-Coronavirus; YouTube Best. Retrieved 2021-11-19 from https://www.forbes.com/sites/johnkoetsier/2020/04/26/reddit-worst-for-mental-health-for-covid-19-news-consumption-survey-says

[35] Roderick M Kramer and Tom R Tyler. 1996. *Trust in organizations: Frontiers of theory and research*. Sage.

[36] Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems* 16, 10 (2015), 1.

[37] Maria Knight Lapinski and Rajiv N Rimal. 2005. An explication of social norms. *Communication theory* 15, 2 (2005), 127–147. https://doi.org/10.1111/j.1468-2885.2005.tb00329.x

[38] Roy J. Lewicki, Daniel J. McAllister, and Robert J. Bies. 1998. Trust and Distrust: New Relationships and Realities. *The Academy of Management Review* 23, 3 (1998), 438–458. http://www.jstor.org/stable/259288

[39] D. Mcknight and Norman Chervany. 2001. *Trust and distrust definitions: One bite at a time*. 27–54.

[40] Paul Mena. 2020. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & internet* 12, 2 (2020), 165–183. https://doi.org/10.1002/poi3.214

[41] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.

[42] Sadiq Muhammed T and Saji K Mathew. 2022. The disaster of misinformation: a review of research in social media. *International journal of data science and analytics* 13, 4 (2022), 271–285. https://doi.org/10.1007/s41060-022-00311-6

[43] Steven Lee Myers. 2022. How Social Media Amplifies Misinformation More Than Information. Retrieved 2023-09-01 from https://www.nytimes.com/2022/10/13/technology/misinformation-integrity-institute-report.html

[44] Peter Nannestad. 2008. What have we learned about generalized trust, if anything? *Annu. Rev. Polit. Sci.* 11 (2008), 413–436.

[45] Andrea L. Nevedal, Caitlin M. Reardon, Marilla A. Opra Widerquist, George L. Jackson, Sarah L. Cutrona, Brandolyn S. White, and Laura J. Damschroder. 2021. Rapid versus traditional qualitative analysis using the Consolidated Framework for Implementation Research (CFIR). *Implementation Science* 16, 1 (12 2021), 1–12. https://doi.org/10.1186/S13012-021-01111-5/TABLES/5

[46] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[47] Derek H. Ogle, Jason C. Doll, A. Powell Wheeler, and Alexis Dinno. 2023. *FSA: Simple Fisheries Stock Assessment Methods*. https://CRAN.R-project.org/package=FSA R package version 0.9.4.

[48] Samantha R Paige, Janice L Krieger, and Michael L Stellefson. 2017. The influence of eHealth literacy on perceived trust in online health communication channels and sources. *Journal of health communication* 22, 1 (2017), 53–65. https://doi.org/10.1080/10810730.2016.1250846

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[50] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[51] Robert D Putnam. 2000. *Bowling alone: The collapse and revival of American community*. Simon and schuster.

[52] R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[53] Gilles Raiche, David Magis, and Maintainer Gilles Raiche. 2020. Package 'nFactors'. *Repository CRAN* (2020), 1–58.

[54] Peter Railton. 2014. Reliance, trust, and belief. *Inquiry* 57, 1 (2014), 122–150.

[55] Douglas A Reynolds et al. 2009. Gaussian mixture models. *Encyclopedia of biometrics* 741, 659-663 (2009).

[56] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[57] Jon Roozenbeek and Sander Van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5, 1 (2019), 1–10. https://doi.org/10.1057/s41599-019-0279-9

[58] Rudolf J Rummel. 1988. *Applied factor analysis*. Northwestern University Press.

[59] Paola Sapienza, Anna Toldra-Simats, and Luigi Zingales. 2013. Understanding trust. *The Economic Journal* 123, 573 (2013), 1313–1332.

[60] Subhro Sarkar, Sumedha Chauhan, and Arpita Khare. 2020. A meta-analysis of antecedents and consequences of trust in mobile commerce. *International Journal of Information Management* 50 (2020), 286–301. https://doi.org/10.1016/j.ijinfomgt.2019.08.008

[61] Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal* 8, 1 (2016), 289–317. https://doi.org/10.32614/RJ-2016-021

[62] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. 2022. Misinformation warnings: Twitter's soft moderation effects on covid-19 vaccine belief echoes. *Computers & security* 114 (2022), 102577. https://doi.org/10.1016/j.cose.2021.102577

[63] Margaret S Sherraden, Betsy Slosar, and Michael Sherraden. 2002. Innovation in social policy: Collaborative policy advocacy. *Social Work* 47, 3 (2002), 209–221.

[64] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36. https://doi.org/10.1145/3137597.3137600

[65] Edson C Tandoc Jr, Darren Lim, and Rich Ling. 2020. Diffusion of disinformation: How social media users respond to fake news and why. *Journalism* 21, 3 (2020), 381–398.

[66] Jiliang Tang, Xia Hu, and Huan Liu. 2014. Is Distrust the Negation of Trust? The Value of Distrust in Social Media. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (Santiago, Chile) *(HT '14)*. Association for Computing Machinery, New York, NY, USA, 148–157. https://doi.org/10.1145/2631775.2631793

[67] Mohsen Tavakol and Reg Dennick. 2011. Making sense of Cronbach's alpha. , 53 pages.

[68] United Nations. 2023. Secretary-General Urges Broad Engagement from All Stakeholders towards United Nations Code of Conduct for Information Integrity on Digital Platforms. Retrieved 2023-09-03 from https://press.un.org/en/2023/sgsm21832.doc.htm

[69] Eric M Uslaner. 2002. The moral foundations of trust. *Available at SSRN 824504* (2002).

[70] José Van Dijck. 2013. *The culture of connectivity: A critical history of social media*. Oxford University Press.

[71] Valentina Vellani, Sarah Zheng, Dilay Ercelik, and Tali Sharot. 2023. The illusory truth effect leads to the spread of misinformation. *Cognition* 236 (2023), 105421. https://doi.org/10.1016/j.cognition.2023.105421

[72] Hilde AM Voorveld, Guda Van Noort, Daniël G Muntinga, and Fred Bronner. 2018. Engagement with social media and social media advertising: The differentiating role of platform type. *Journal of advertising* 47, 1 (2018), 38–54. https://doi.org/10.1080/00913367.2017.1405754

[73] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[74] Yiran Wang and Gloria Mark. 2013. Trust in Online News: Comparing Social Media and Official Media Use by Chinese Citizens. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 599–610. https://doi.org/10.1145/2441776.2441843

[75] Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter* 21, 2 (2019), 80–90. https://doi.org/10.1145/3373464.3373475

[76] Savvas Zannettou. 2021. " I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 865–876. https://doi.org/10.1609/icwsm.v15i1.18110

[77] Yixuan Zhang, Joseph D Gaggiano, Nutchanon Yongsatianchot, Nurul M Suhaimi, Miso Kim, Yifan Sun, Jacqueline Griffin, and Andrea G Parker. 2023. What Do We Mean When We Talk about Trust in Social Media? A Systematic Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany). ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3544548.3581019

[78] Yixuan Zhang, Nurul M Suhaimi, Nutchanon Yongsatianchot, Joseph D Gaggiano, Miso Kim, Shivani A Patel, Yifan Sun, Stacy Marsella, Jacqueline Griffin, and Andrea G Parker. 2022. Shifting Trust: Examining How Trust and Distrust Emerge, Transform, and Collapse in COVID-19 Information Seeking. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New Orleans,

LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, 21 pages. https://doi.org/10.1145/3491102.3501889

A APPENDICES

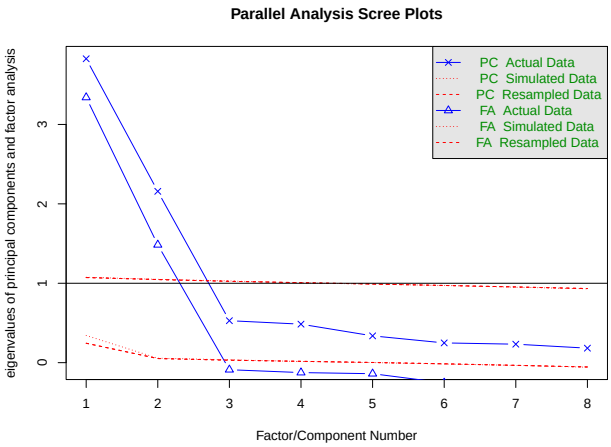

Figure 11: Parallel analysis to determine the number of components to keep in the factor analysis

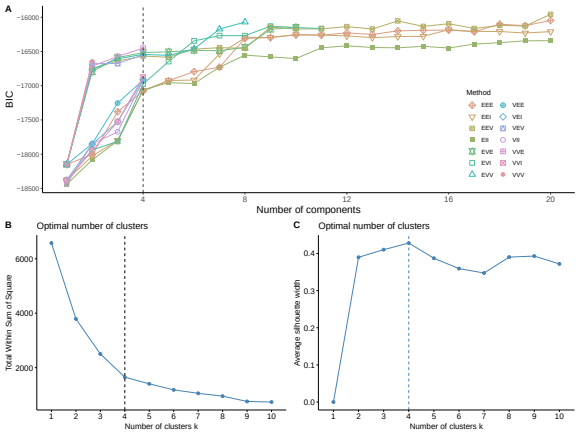

Figure 12: Cluster number definition criteria: (A) BIC value, (B) With-in-Sum-of-Squares (WSS) and (C) Average Silhouette Method.

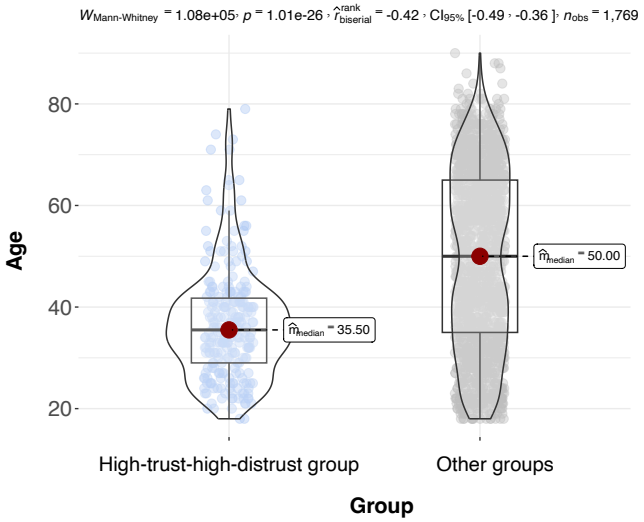

Figure 13: Non-parametric comparison test (Dunn's test) that shows the differences in age across different groups.

Table 6: Logistic regression model coefficients with demographic predictors for being in the high-trust-high-distrust group

Predictor	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.717	0.538	5.050	0.000***
Age	-0.059	0.006	-10.040	0.000***
Education				
(*Reference: Less than high school*)				
High school graduate	-1.459	0.442	-3.300	0.015*
Associate degree	-1.089	0.459	-2.370	0.281
Bachelor's degree	-0.637	0.453	-1.410	1.000
Postgraduate degree	-0.658	0.472	-1.390	1.000
Gender				
(*Reference: Male*)				
Female	-0.883	0.154	-5.720	0.000***
Prefer not to answer or Non-binary	-1.353	1.077	-1.260	1.000
Income				
(*Reference: Low income*)				
Moderate-to-high income	-0.268	0.189	-1.420	1.000
Ethnicity				
(*Reference: Non-Hispanic*)				
Hispanic	-0.242	0.235	-1.030	1.000
Race				
(*Reference: White*)				
Asian	-0.181	0.256	-0.710	1.000
Black	0.230	0.192	1.200	1.000
Other	-0.542	0.247	-2.190	0.454
Political Ideology				
(*Reference: Democrat / Lean Democrat*)				
Independent	-0.935	0.195	-4.800	0.000***
Other, please describe	-0.767	0.665	-1.150	1.000
Republican / Lean Republican	-0.642	0.204	-3.150	0.026*

Table 7: Median difference and Dunn's Test result of Trust & Distrust Levels in Between-group multiple comparisons (Significance level: $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$)

	Platform	Median diff.	Z-score (P-value)
Trust	Facebook - TikTok	-0.25	-5.89 (< 0.001 ***)
	Facebook - Twitter	-0.5	-6.95 (< 0.001 ***)
	Facebook - YouTube	-0.5	-8.96 (< 0.001 ***)
	TikTok - Twitter	-0.25	-0.95 (0.34)
	TikTok - YouTube	-0.25	-2.54 (< 0.01 **)
	Twitter - YouTube	0.00	-1.56 (0.24)
Distrust	Facebook - TikTok	0.00	-0.17 (1.00)
	Facebook - Twitter	0.00	-0.08 (0.94)
	Facebook - YouTube	0.25	8.21 (< 0.001 ***)
	TikTok - Twitter	0.00	0.09 (1.00)
	TikTok - YouTube	0.25	7.77 (< 0.001 ***)
	Twitter - YouTube	0.25	7.72 (< 0.001 ***)

Table 8: Multiple regression models explaining respondents' trust in social media interventions, represented by Labeling, Curation, and Verification tasks. (Significance level: $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$) The analysis suggests that trust decreases with age and varies significantly with political affiliation, and significant gender differences as non-binary users exhibit markedly lower trust, and Black respondents show more trust in the Labeling and Curation feature.

	Dependent Variable		
	Labeling β (Std. Error)	Curation β (Std. Error)	Verification β (Std. Error)
const	4.1823 (0.1734)***	4.2507 (0.1671)***	4.1207 (0.1707)***
Age	-0.0102 (0.0013)***	-0.0099 (0.0013)***	-0.0088 (0.0013)***
Education			
(*Reference: Less than high school*)			
High school	-0.0333 (0.1518)	-0.1782 (0.1462)	-0.0417 (0.1494)
Associate	-0.0587 (0.1564)	-0.1250 (0.1507)	-0.0450 (0.1539)
Bachelor	0.0063 (0.1564)	-0.0934 (0.1506)	-0.0003 (0.1539)
Postgraduate	0.0240 (0.1604)	-0.0484 (0.1545)	0.0175 (0.1578)
Gender			
(*Reference: Male*)			
Female	-0.0480 (0.0441)	-0.0993 (0.0430)	-0.0424 (0.0434)
Non-binary	-0.7043 (0.2422)	-0.7003 (0.2362)*	-0.4980 (0.2383)
Income			
(*Reference: Low income*)			
Moderate-to-high income	-0.0326 (0.0556)	-0.0213 (0.0542)	0.0118 (0.0548)
Race & Ethnicity			
(*Reference: Non Hispanic*)			
Hispanic	0.0821 (0.0669)	0.0302 (0.0653)	0.0638 (0.0659)
(*Reference: White*)			
Asian	0.0533 (0.0719)	0.0366 (0.0702)	-0.0204 (0.0707)
Black	0.1770 (0.0604)*	0.1868 (0.0589)*	0.1601 (0.0594)
Other	0.0171 (0.0648)	0.0436 (0.0632)	0.0188 (0.0638)
Political Ideology			
(*Reference: Democrat*)			
Republican	-0.3301 (0.0555)***	-0.3709 (0.0542)***	-0.3669 (0.0547)***
Independent	-0.3048 (0.0525)***	-0.3524 (0.0513)***	-0.3085 (0.0517)***
Other	-0.5776 (0.1636)**	-0.6575 (0.1573)***	-0.6704 (0.1610)***

Table 9: Multiple regression models explaining respondents' distrust in social media interventions, represented by Labeling, Curation, and Verification tasks. (Significance level: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$) The analysis indicates that older age groups show decreased distrust; females generally exhibit less distrust than males, especially in the Labeling and Verification feature.

	Dependent Variable		
	Labeling β (Std. Error)	Curation β (Std. Error)	Verification β (Std. Error)
const	3.4840 (0.1705)***	3.5693 (0.1791)***	3.6325 (0.1770)***
Age	-0.0033 (0.0013)	-0.0069 (0.0014)***	-0.0055 (0.0014)***
Education			
(Reference: Less than high school)			
High school	-0.2248 (0.1492)	-0.2058 (0.1567)	-0.3129 (0.1549)
Associate	-0.1571 (0.1538)	-0.1063 (0.1616)	-0.2184 (0.1597)
Bachelor	-0.1297 (0.1537)	-0.1360 (0.1615)	-0.2149 (0.1596)
Postgraduate	-0.2244 (0.1577)	-0.2936 (0.1657)	-0.3445 (0.1637)
Gender			
(Reference: Male)			
Female	-0.1322 (0.0439)*	-0.1673 (0.0461)***	-0.1600 (0.0456)**
Non-binary	0.2789 (0.2410)	0.2090 (0.2623)	0.3474 (0.2503)
Income			
(Reference: Low income)			
Moderate-to-high income	-0.0172 (0.0554)	0.0070 (0.0582)	-0.0024 (0.0575)
Race & Ethnicity			
(Reference: Non Hispanic)			
Hispanic	0.1439 (0.0668)	0.1305 (0.0700)	0.0769 (0.0693)
(Reference: White)			
Asian	0.1119 (0.0718)	0.1352 (0.0753)	0.0836 (0.0743)
Black	0.0466 (0.0602)	0.1002 (0.0631)	0.1295 (0.0625)
Other	-0.0795 (0.0645)	-0.0436 (0.0677)	-0.0285 (0.0669)
Political Ideology			
(Reference: Democrat)			
Independent	-0.1020 (0.0524)	-0.1143 (0.0550)	-0.0988 (0.0544)
Republican	0.1076 (0.0553)	0.1085 (0.0581)	0.0982 (0.0574)
Other	0.0859 (0.1605)	0.0332 (0.1687)	-0.0709 (0.1666)