Who is benefitting from fact-checking on social media – user or platform? Examining the impact of different fact-checking approaches on social media platforms on user's perception of trust

Who is benefitting from fact-checking on social media – user or platform?

Examining the impact of different fact-checking approaches on social media platforms on user's perception of trust

Alexander Acht

A Thesis in the Field of Psychology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2024

Abstract

This thesis explores the effectiveness of fact-checking in increasing trust in social networking websites (SNS) and its effect on overall media trust. It seeks to study how fact-checking can alleviate the issue of misinformation, particularly on social media platforms. The research is motivated by the spread of fake news, which has contributed to polarization and mistrust in institutions. The study begins by discussing the problem of misinformation and how it can shape public opinion and trust. It highlights the significant role of trust in the functioning of social media sites and notes how controversies and misinformation have led to a decline in trust in these platforms.

The hypothesis is that fact-checking, particularly by external, credible sources, can positively influence users' trust in the platform where it occurs, but it may also decrease trust in media at large by highlighting misinformation. An experimental design is used to test the hypothesis with simulated social media environments where participants are exposed to various fact-checking scenarios, including internal and external corrections of misinformation.

The results reveal that credible external fact-checking organizations can significantly increase trust in social media platforms. However, the impact on overall media trust is complex, suggesting that fact-checking can increase awareness of misinformation, but it does not uniformly enhance trust across all media. The findings emphasize the importance of source credibility and the presentation of fact-checked information in shaping public trust.

The thesis contributes to the understanding of fact-checking's potential as a tool for combating misinformation and enhancing platform trust. It calls for a multi-faceted approach to addressing misinformation, combining credible fact-checking with efforts to promote media literacy and critical information evaluation skills among the public. By highlighting the importance of source credibility and the method of information correction, the study offers insights for policymakers, social media companies, and the broader public on strategies to enhance trust and counter misinformation in the digital age.

Table of Contents

List of Tables

List of Figures

Chapter I.

Introduction

In his recent publication entitled "21 Lessons for the 21st Century," Yuval Harari (2018) underscores the pressing issue of fake news as a major challenge of our time. The author delves into the pervasive problem of misinformation and propaganda in contemporary society, with a particular focus on the impact of information overload in the context of social media. Harari (2018) argues that our brains, which tend to be more inclined towards absorbing captivating narratives, struggle to process and analyze factual information, numbers, and statistics. While not explicitly suggesting fact-checking as a solution, the author emphasizes the significance of Artificial Intelligence and the scientific community's increased involvement in sharing their knowledge through easily comprehensible narratives with the public.

The Significance of Fact-Checking

Disseminating false information poses a significant obstacle for contemporary societies, particularly on social media platforms. Fake news refers to the deliberate dissemination of false and biased information disguised as legitimate news with the intention of manipulating public opinion or damaging the reputation of individuals, companies, organizations, or subjects (Lazer et al., 2018; Pennycook et al., 2018). One example of a false claim is the alleged plan by Microsoft co-founder Bill Gates to implant microchips through vaccines. Despite the lack of evidence, the portrayal of this claim as

true news on social media has had a detrimental impact on the reputation of Bill Gates and his organizations. Based on a YouGov poll, it was found that 28% of Americans support this claim. However, when focusing solely on Republicans, the percentage rises to 44% (Goodman & Carmichael, 2020). Considering the widespread reliance on social media for news consumption, the dissemination of false information can result in an overwhelming number of misunderstandings and perplexities among individuals, ultimately causing societal divisions and eroding trust in both institutions and the media.

The influence of misinformation on Social Networking Sites [SNS] in shaping public debate is significant, given that social media platforms are the primary sources of daily news for a majority of individuals. According to Bridgman et al. (2020), 61% of 18-29-year-olds and 54% of 30-44-year-olds rely on social media for news, with only individuals over 55 mentioning network news before turning to social media. This highlights the significance of news and misinformation on social media platforms, as a considerable 74% of Americans already view fake news on social media as a significant issue (Gallup and Knight Foundation, 2020). Throughout the COVID-19 pandemic, the spread of false information about vaccines on social networking sites had a direct and detrimental effect on people's attitudes toward vaccination. This, in turn, hindered the efforts to combat COVID-19 and tragically led to more deaths, as noted by Zhang et al. (2021).

As stated, inaccurate information can have serious consequences, particularly when it leads individuals to make poor decisions that put their health, safety, or overall welfare at risk. For example, spreading inaccurate information about vaccinations could lead to individuals avoiding essential immunizations. Misinformation can also erode trust

in crucial institutions like the government and the media, ultimately undermining social cohesion and impeding the smooth functioning of a community. Correcting misinformation is crucial as it ensures that individuals have accurate information and can make informed judgments. Thus, it becomes crucial to investigate the efficacy of fact-checking labels and their influence on trust in media or social networking sites.

The Significance of Trust for Social Networking Sites

In recent years, the reputation of social media has experienced a decline due to a range of challenges, including misinformation, lack of transparency, data protection issues, societal polarization, cyberbullying, and concerns regarding the safety of minors. The industry's reputation has been dramatically affected by this array of issues. Notably, in the wake of trust violations like the Facebook–Cambridge Analytica data scandal or the intense public scrutiny that social networking sites face during televised congressional hearings, the significance of establishing trust becomes evident. In the wake of the Cambridge Analytica Scandal, where 87 million Facebook user data was misused for political purposes, resulting in a substantial breach of trust and a decline in Facebook's reputation, researchers turned to the Fuoli and Paradis' trust-repair discourse model. This model focuses on neutralizing the negative aspects and highlighting the positive aspects to rebuild trust for Facebook (Amran, 2016).

According to the research conducted by Ipsos' Global Trustworthiness Monitor, the social media sector consistently ranks poorly in terms of trustworthiness, sharing the lowest position with the government. According to the most recent Ipsos Global Trustworthiness Monitor, the perception of social media companies as trustworthy among

the global public stands at a mere 22%. This finding highlights a persistent decline in public confidence in social media platforms (Ipsos, 2023).

Several crucial factors underscore the significance of social media companies in building and maintaining public trust. Trust is crucial in fostering user engagement and retention on online platforms, enabling increased user participation and interaction (Bright et al., 2021). After a privacy violation, the repair of trust in social media platforms and their associated services or products is closely linked to behavioral integrity. The perception of an entity's integrity greatly impacts trust repair, while the subsequent actions taken after an apology play a crucial role in rebuilding credibility. Simultaneously, an organization must ensure that its words are in harmony with its actions in order to effectively showcase integrity, rebuild trust, and uphold relationships with users (Ayaburi & Treku, 2020; Warner-Søderholm et al., 2018). This highlights the rationale behind SNS openly sharing and displaying their fact-checking efforts on their platforms, ensuring that their actions align with their communication.

In addition, the reliability of information shared on social media, which is a primary source for many users, plays a crucial role in curbing the spread of misinformation (Pennycook et al., 2018). The viability of these platforms' businesses is closely tied to trust, which significantly impacts their attractiveness to advertisers, investors, and users. Establishing trust is of utmost importance when adhering to privacy and data protection regulations, especially for social media companies operating in ever more rigorous regulatory landscapes (Mantelero, 2018). In addition, academic discourse highlights the significance of trust in cultivating favorable online communities, promoting the use of new features and services, and addressing concerns like

misinformation and online harassment. Trust shapes how people perceive a company's social responsibility and ethical practices. In addition, the trust that social media platforms build with users can significantly impact their competitive advantage, as highlighted by McKnight et al. (2011). Ultimately, establishing a solid public trust influences user behavior and has extensive consequences for regulatory compliance, business sustainability, and the ethical reputation of social media companies within the broader societal framework.

Contrary to popular belief, the connection between trust and usage of SNS is more complex than it may seem. While trust can have a positive influence on user behavior on these platforms, such as engagement, interaction, or usage time (Wang et al., 2016), being a frequent user does not necessarily imply a high level of trust with other users (Lan & Tung, 2024). In order to comprehensively analyze the usage patterns of participants, the present study has gathered data and aims to integrate these findings.

The significance of trust as a measurable and influential element in the business objectives of social media companies is further highlighted when examining the external communication of SNS. TikTok's website emphasizes their commitment to creating the most trusted entertainment platform (TikTok, 2023). Meanwhile, Meta, the parent company of Facebook, Instagram, and WhatsApp, has taken the initiative to establish Trust, Transparency, and Control Labs (TTC Labs) as a cross-industry project aimed at enhancing trust, transparency, and control in digital products (TTC Labs, 2023).

Extensive research has been conducted on the efficacy of fact-checking in correcting individuals' misconceptions. However, the impact of fact-checking on trust in the SNS and media remains relatively unexplored. I argue that social media platforms use

fact-checking to increase users' trust in the platform. This is because the bare encounter with fake news can already negatively impact user's perception of trust towards a platform while the perception of an attempted correction or moderation of content can be beneficial (Ayaburi & Treku, 2020; Lan and Tung, 2024; Warner-Søderholm et al., 2018). On the contrary, in some cases, users perceive information correction and content moderation, such as decreasing its visibility, as censorship, ultimately decreasing trust. (Stewart 2021). The positive effect of fact-checking on the level of trust towards a social media platform is influenced by the source of the correction (Kim & Dennis, 2019; Margolin et al., 2018; Schwarz & Newman, 2017; Young et al., 2018; Zhang et al., 2021) and extend of fact-checking, i.e., labelling or flagging misleading information as false versus presenting corrected information (Berinsky, 2015; Byrne & Hart, 2009; Lazer et al., 2018; Nisbet et al., 2015; Porter & Wood, 2021; Schwarz et al., 2016). While I expect that fact-checking will increase trust in the specific site where it occurs, I think it will decrease trust in media in general because it alerts people to the fact that there is misinformation in the media (Brenan, 2020; Lan & Tung, 2024; Melki et al. 2021; Primig (2022).

Effectiveness of Fact-checking and Its Challenges

While numerous studies have investigated the effectiveness of fact-checking in rectifying people's misunderstandings, the influence of fact-checking on trust in SNS and media has yet to be fully explored. Fact-checking can be prone to pitfalls such as mistrust in the organization that corrected false information, overall mistrust in media, or selective sharing but in addition, psychological factors such as confirmation bias or reactance can influence the effectiveness of fact-checking.

Nevertheless, fact-checking in the form of labeling or removing content has become an industry standard for most platforms (Garrett & Poulsen, 2019), it is unclear whether the implemented tools are effective in correcting false beliefs or whether they are merely a PR and marketing claim, i.e., bluewashing, to appease public policymakers. Nevertheless, research indicates that the current measures are insufficient (Chou et al., 2021). Extraneous factors such as the perceived credibility of the correcting source (Kim & Dennis, 2019; Zhang et al., 2021), the presentation of the corrected information (Nassetta & Gross, 2020), or social connections and political affiliations (Margolin et al, 2018) can negatively impact the effectiveness of fact-checking tools. Furthermore, psychological factors such as confirmation bias (Kim & Dennis, 2019) or reactance (Garrett & Poulsen, 2019) can have negative consequences by, for example, reinforcing false information, implying that the act of correcting false information may already have a negative impact on the effectiveness of fact-checking (Nisbet et al., 2015; Schwarz et al., 2016).

The increasing relevance and usage of social media networks, which causes fake news to spread faster than ever, emphasizes the need to understand the consequences of fact checking, including its effects on trust perception. Ineffective measures against this spread could impact general education and trust in media sources which influences the democratic discourse once there is no everyday basis for mutual exchange. Especially if fact-checking measures would only help SNS building trust with their users but not effectively correct false believes. In the past years, it has become more difficult to identify and correct fake news (Dimock, 2019) and most news on social media is already

perceived as false by young adults (Gallup and Knight Foundation, 2020) highlighting wider relevance and possible implications for society.

In a recent study, Gallup looked deeper into the division of society, fake news, and trust in media. The issue becomes very clear when looking at the trust in mass media concerning political party affiliation. In 1998, 59% of Democrats and 52% of Republicans stated that they trust the media. In the recent survey from Gallup, trust in media from Democrats increased to 73% but decreased to only 10% for Republicans (Brenan, 2020). This is also considered a "political divide" and points out the fundamental problem of media consumption today: A significant portion of society does not believe the information they receive from media channels or avoids them entirely, instead depending mainly on information in their social media feeds, emphasizing the need for effective fact-checking measures. This "information gerrymandering" can distort social media opinions because people establish opinions based on the media they consume or the people with whom they engage, while social networking sites or their algorithms, as well as confirmation bias, create an "information bubble."

When people are within their information bubble, they cannot pay attention to all news because of an information overload. Whenever our brain is overwhelmed by the influx of stimuli, it uses cognitive heuristics to ease our thinking. People prefer easily consumable stories and information from people they trust (from in-group sources), which they are more inclined to share with others, as noted by Harari (2018). Also, due to confirmation bias, people search for information that supports their beliefs and are more likely to remember them (Menczer & Hills, 2020). Consequently, when people distrust the media and are in their information bubble, they are more likely to believe and share

fake news (Ognyanova et al., 2020). This exemplifies the fundamental problem with false information on social networking sites and the significance of understanding the efficacy of fact-checking. Fake news spreads faster than the truth, producing a downward spiral, when users are overloaded with too much information, unable to verify its authenticity, and more willing to share the information from their relatives and friends (Vosoughi et al., 2018). The use of fact-checking tools could be a viable method for reversing this downward spiral and assisting users in the validation process, without necessarily preventing the rapid spread of information, but rather reducing the distribution of incorrect information.

Fact-Checking as a Tool to Increase Trust

This paper contends that social media platforms employ fact-checking mechanisms as a strategic means to bolster user trust. The rationale behind this assertion is grounded in the observation that mere exposure to misinformation can detrimentally affect users' trust perceptions towards a platform. Conversely, the perception that a platform is making efforts to correct or moderate content is seen to have a positive impact. This argument is supported by the findings of Ayaburi & Treku (2020), Lan and Tung (2024), and Warner-Søderholm et al. (2018), indicating that proactive content moderation strategies, including fact-checking, are instrumental in enhancing trust among social media users.

Ayaburi & Treku (2020) delve into the realm of crisis management following privacy violations on social media, aiming to shed light on how organizations navigate the complex process of addressing privacy concerns, restoring trust, and retaining users in the aftermath. Central to the investigation is the concept of behavioral integrity, which

emerges as a pivotal factor in the repair of trust within the context of social media platforms and their related services or products. The research underscores that the perception of an entity's integrity by users plays a crucial role in the trust repair process, with actions taken subsequent to an apology significantly contributing to the restoration of credibility. The study illuminates the integral role of apologies in the trust repair mechanism, the critical importance of behavioral integrity, and how these elements may influence users' perceptions of other services offered by the same entity. It posits that while privacy concerns themselves may not have a direct impact on the restoration of trust, they are modulated by users' perceptions of the entity's behavioral integrity. This highlights the necessity for a congruence between what an organization says and what it actually does in the aftermath of a crisis to successfully rebuild trust and sustain user engagement. In the realm of fact-checking, platforms can exhibit behavioral integrity by articulating the significance of countering disinformation and concurrently implementing prominent methods to regulate its dissemination, so fostering trust. Concluding, the study emphasizes the indispensable role of behavioral integrity as a mediator between the effectiveness of an apology and the repair of trust, suggesting avenues for future research such as exploring various crisis response strategies, the optimal timing for apologies, and the consequences of fact-checking on trust.

Accusations of Bias and Censorship

Maintaining behavioral integrity while combating misinformation can have a positive impact on trust. However, if corrections are perceived as biased or if content moderation seems like censorship, the trust can decline. Therefore, it is crucial to ensure unbiased fact-checking mechanisms and transparent content moderation policies to foster

trust among users. Stewart (2021) highlights the critical tension between the need for content moderation and the accusations of bias and censorship such efforts may invite. Central to the discussion are two main issues: defining problematic content warranting moderation and the feasibility of categorizing content impartially. The paper outlines the intrinsic value judgments involved in identifying what constitutes "fake news," suggesting that any attempt at content moderation inherently involves bias, potentially eroding user trust in the platform and its fact-checking processes. The researcher delves into the policy and labeling challenges in content moderation, emphasizing the difficulties in achieving consensus on what content should be targeted and accurately identifying instances of fake news without imposing subjective biases. Stewart (2021) proposes that resolving these issues necessitates biased decisions that prioritize certain values over others, which, while not discrediting the need for content moderation, calls for cautious implementation to maintain user trust. The paper suggests employing diverse moderators, limiting the scope of moderation to specific problems, and enhancing transparency about moderation processes as strategies to mitigate bias and foster trust.

Pitfalls of Fact-Checking

Fact-checking, while important in countering false information, can have negative impacts on trust when accusations of bias and censorship arise. Additionally, there are potential pitfalls such as mistrust in the organization conducting the fact-checking, overall mistrust in media, and selective sharing. Psychological factors, such as confirmation bias and reactance, can impact the effectiveness of fact-checking and its influence on trust.

Although fact-checking measures such as labeling or removing content have become a common practice for many online platforms (Garrett & Poulsen, 2019), it remains uncertain whether these tools are truly effective in rectifying false beliefs or whether they are merely employed as a public relations and marketing strategy, also known as "bluewashing," to satisfy public policymakers.

According to research, the measures currently in place to fact-check information may not be enough (Chou et al., 2021). Other factors such as the perceived credibility of the source correcting the information (Kim & Dennis, 2019; Zhang et al., 2021), the way in which the corrected information is presented (Nassetta & Gross, 2020), or social connections and political affiliations (Margolin et al, 2018) can all negatively impact the effectiveness of fact-checking tools. Additionally, psychological factors such as confirmation bias (Kim & Dennis, 2019) or reactance (Garrett & Poulsen, 2019) can also have negative consequences, potentially reinforcing false information. This implies that simply correcting false information may actually have a negative impact on the effectiveness of fact-checking (Nisbet et al., 2015; Schwarz et al., 2016).

The increasing relevance and usage of social media networks, which causes fake news to spread faster than ever, emphasizes the need to understand the effectiveness of fact checking. Ineffective measures against this spread could impact general education and trust in media sources which influences the democratic discourse once there is no everyday basis for mutual exchange. In the past years, it has become more difficult to identify fake news (Dimock, 2019) and most news on social media is already perceived as false by young adults (Gallup and Knight Foundation, 2020) highlighting wider relevance and possible implications for society.

In a recent study, Gallup looked deeper into the division of society, fake news, and trust in media. The issue becomes very clear when looking at the trust in mass media concerning political party affiliation. In 1998, 59% of Democrats and 52% of Republicans stated that they trust the media. In the recent survey from Gallup, trust in media from Democrats increased to 73% but decreased to only 10% for Republicans (Brenan, 2020). This is also considered a "political divide" and points out the fundamental problem of media consumption today: A significant portion of society does not believe the information they receive from media channels or avoids them entirely, instead depending mainly on information in their social media feeds, emphasizing the need for effective fact-checking measures. This "information gerrymandering" can distort social media opinions because people establish opinions based on the media they consume or the people with whom they engage, while social networking sites or their algorithms, as well as confirmation bias, create an "information bubble."

When people are within their information bubble, they cannot pay attention to all news because of an information overload. Whenever our brain is overwhelmed by the influx of stimuli, it uses cognitive heuristics to ease our thinking. People prefer easily consumable stories and information from people they trust (from in-group sources), which they are more inclined to share with others, as noted by Harari (2018). Also, due to confirmation bias, people search for information that supports their beliefs and are more likely to remember them (Menczer & Hills, 2020). Consequently, when people distrust the media and are in their information bubble, they are more likely to believe and share fake news (Ognyanova et al., 2020). This exemplifies the fundamental problem with false information on social networking sites and the significance of understanding the efficacy

of fact-checking. Fake news spreads faster than the truth, producing a downward spiral, when users are overloaded with too much information, unable to verify its authenticity, and more willing to share the information from their relatives and friends (Vosoughi et al., 2018). The use of fact-checking tools could be a viable method for reversing this downward spiral and assisting users in the validation process, without necessarily preventing the rapid spread of information, but rather reducing the distribution of incorrect information.

This study explores the effectiveness of fact-checking tools on social media depending on their type and source by assessing their effect on users' perceived trust in a social media platform and media in general. Based on research by Nassetta and Gross (2020) which indicate a relation between effective fact-checking and the presentation of the corrected information, the type of fact-checking was manipulated in this study by using presentation conditions. Additionally, research showed that the effectiveness of fact-checking can also depend on the source that correct the information (Kim & Dennis, 2019; Zhang et al., 2021). As a result, this study included two different sources as the sender of the fact-check. The source of the correction will be labeled as internal correction by the social media platform or external correction by, for example, a bipartisan NGO.

Chan et al.'s (2017) meta-analysis of the psychological efficacy of messages that counter belief in misinformation questions Facebook's fact-checking procedure and indicates that fact-checking labels that only label fake news as incorrect have a comparably weaker effect than sharing broader context or empirical evidence to debunk the false information. Verifications from reputable health institutes or colleges were

especially useful in debunking vaccination misconceptions (Chan et al., 2017). According to Bode and Vraga (2017), effectiveness of fact-checking labels in correcting misinformation on health issues is dependent on the sender of the correction, while distribution of this correction via comments under a posting or an algorithm was equally effective. The findings show that, at least for health issues, comments under the posting are as effective as algorithmic fact-checks at correcting health misinformation. The recommendation suggests that public health officials should use a communication strategy to combat misinformation by providing straightforward, factually supported rebuttals with credible references. To reach a wider audience, they should expand the range of social media platforms used to spread accurate information, inviting user-generated corrections. However, relying on proprietary algorithms and algorithmic interventions could undermine the effectiveness of this approach due to their opaque nature and the potential for distrust. Future studies should examine the comparative dynamics of corrections made via social interactions versus those made by algorithms.

A recently published study by Zhang et al. (2021) examined the effect of fact-checking vaccine disinformation supports the claim that the efficacy of fact-checking is dependent on the source of the correction. They discovered that subjects were more likely to show positive attitudes toward vaccinations when asked to rank the benefit or usefulness of vaccines on a questionnaire, and that labels from health institutions and research universities were more effective than fact-checking organizations, news media or algorithms. This example demonstrates the significance of the subject matter, as WHO, HSPH, and the CDC have identified viral misinformation as one of the most severe threats associated with the COVID-19 pandemic.

Even while fact-checking labels are influential in certain instances, the correction of misinformation has a minimal effect on a person's real belief (Garrett, 2011) and is dependent on the relationship between the person being corrected and the provider of the correction (Margolin, 2018). Fact-checking had a negligible effect on false information congruent with a person's worldviews and beliefs (Einwiller & Kamins, 2008). The confirmation bias amplifies this tendency, as individuals tend to consume media that confirms their perspective and avoid sources that might challenge it (Bessi et al., 2015). For example, a person who is vulnerable to conspiracy theories is also inclined to distrust the government and established institutions, making them ineffective as a source to correct misinformation and increasing their belief in the misinformation when fact-checked by a distrusted institution (Larson et al., 2016). The interplay of various psychological factors and personal predispositions exemplifies the complicated nature of misinformation, media mistrust, and fact-checking approaches on social media. Due to the fact that 62% of people acquire their news from social media, it is even more important to comprehend the effects of fact-checking and the implications of disinformation on SNS (Shearer & Gottfried, 2017). Concerningly, inaccurate or misleading content that evades fact-checking and is not tagged can produce an implied truth effect, in which users perceive the untagged information as accurate since it is not labeled as fake news (Pennycook & Rand, 2017).

According to research, individuals selectively share fact-checked information if it supports their political beliefs, reinforcing the bias of friends and followers and creating information bubbles. Shin and Thorson (2017) studied how Twitter users shared and commented on three large fact-checking Twitter accounts in October 2012. Data was

gathered from a vast collection of political tweets, and each fact-check was coded for a political party that acquired a relative advantage, as well as the fact-valence checks toward Obama and Romney. A measure incorporating three computational approaches was devised to determine the partisanship of each Twitter user who commented on or retweeted a fact-checking post. The study looked at how social media users reacted to political fact-checking statements. Drawing on Social Identity Theory, the study examined how partisan bias influences the selective dissemination of information, finding that individuals were inclined to share fact-checked content that aligned with their political stance while disregarding content that favored the opposing viewpoint. Democrats were more likely to share fact-checked information, while Republicans were more hostile to fact-checks. According to the study, selective sharing may further polarize audiences and decrease faith in fact-checking (Shin & Thorson, 2017). This selective sharing of content amplifies the bias of friends and followers, resulting in an information bubble. Shin and Thorson (2017) also demonstrated that political identification can affect the likelihood that content will be spread. Negative fact checks on President Obama were more likely to be shared by Republicans, and negative fact checks about Senator Romney were more likely to be shared by Democrats. Based on their findings, the researchers hypothesize a spiraling relationship between conservative attitudes, media bias, affective polarization, and selective sharing (Shin & Thorson, 2017). It can be difficult to correct someone's false beliefs because beliefs are deeply personal and can be influenced by a range of variables such as upbringing, cultural or religious beliefs, and personal experiences. People may be hard to changing their views if they have held them for a long time, and they may be even more resistant if their beliefs

appear to be under attack. Both Lewandowsky et al. (2012) and Schwarz et al. (2007) acknowledge that it is exceedingly challenging to rectify someone's incorrect information. Debiasing strategies, such as fact-checking labels, must be easily available to the brain in order to be useful from a cognitive perspective (Schwarz et al. 2007). Individuals have difficulty detecting inaccurate information and are more likely to accept than reject new information. Correcting inaccurate information involves personal relevance, high degrees of attention, and an understanding of the subject at hand. Correcting misinformation is difficult if the issue is unimportant or if the needed level of attention to comprehend the topic is not available (Lewandowsky et al. 2012). It is possible to minimize the acceptance of fake news by warnings, but it is difficult to alter the information after it has been presented. The most common and effective strategy is to counteract false information with the truth, but this must occur quickly after its disclosure (Schwarz et al., 2016) indicating that fact-checking tools could be more effective if they intervene right after the exposure to false information.

Impact of Misinformation and Fact-checking on Media Trust

It is widely acknowledged that fact-checking can be an effective tool for correcting misconceptions and thus influencing trust perception, provided that negative influencing factors such as the source of information, its visual representation, and the perception of fact-checking as censorship or partiality are considered. Additionally, the mere perception of misinformation can create a general mistrust in media as a whole, as it alerts people to the fact that there is misinformation present in the media, which can lead to the propagation and acceptance of fake news. These concerns have been highlighted by

various studies, including Brenan (2020), Lan and Tung (2024), Melki et al. (2021), and Primig (2022).

According to a recent Brenan (2020) research, 70% of Americans are concerned that media owners influence coverage, and 83% blame the media for political polarization in the United States. Due to the widespread skepticism of the media, especially the mainstream media, there is an increase in the adoption of fake news from alternative, nonmainstream media sources. A functioning democracy is dependent on media credibility because voters want accurate information from reputable sources in order to make informed decisions (Jones, 2004). Prior to the introduction of social media, media mistrust was already prevalent. According to Jones (2004), Americans disliked the media because many journalists focus on why a politician selects or proposes a policy as opposed to the politician's actual behavior, which Jones directly links to mistrust in the media. According to the results of his study, mistrust of politicians is somewhat connected with suspicion of the media as a whole. In this perspective, it is vital to recognize that government credibility has declined in recent years, directly affecting media credibility (Jones, 2004). Another element of distrust in media is media bias, which results from the selection of articles, subjects, and events covered by news outlets. It may also be influenced by the tastes of the intended audience, the ownership of a media firm, or governmental factors (Jones, 2004; Shin & Thorson, 2017) In addition, Jenkins (2018) identified information overload and political partisanship as possible causes of media distrust.

Melki et al. (2021) and Primig (2022) linked fake news to a lack of media trust and demonstrated that participants were more inclined to believe misleading information

if they lacked media trust. Therefore, people with higher trust in social media platforms were more prone to believe false news. Carson et al. (2022) identified a negative relationship between fact-checking measures and media trust. When information was verified by a third party, reader confidence in the news declined. The impression of fact checkers as a power elite evaluating the facts on behalf of government propaganda may be a contributing factor (Hanitzsch and Vos, 2018; Fawzi, 2020). In addition to information overload and the novelty or surprise component of fake news, mistrust in the media is a key reason in its propagation and why users believe misleading information. When social media users lack trust in the traditional media landscape and view their news as government-controlled and potentially restricted, they are more likely to believe fake news. Although fact-checking is frequently adopted to regulate the spread of disinformation, third-party fact checkers might be regarded as a government-controlled elitist organization tasked with promoting propaganda. Primig (2022) showed that media trust is relevant not only in the evaluation and identification of false information, since people are more likely to identify fake news if they trust traditional media channels, but also in the strengthening of trust in fact-checking measures. This suggests that fact-checking procedures are more effective when users have greater trust in the media.

Study Aims and Hypotheses

The aim of this study is to assess the impact of fact-checking labels and corrections on users' trust in SNS and the general media. In recent years, social media has become an essential source of information for individuals worldwide. However, the spread of misinformation on these platforms has become a significant concern, leading to a loss of trust in both social media platforms and the broader media landscape. This study

aims to understand how different fact-checking mechanisms employed by social media platforms influence the perceived trustworthiness of the platforms themselves as well as the broader media landscape. Most studies suggest that fact-checking can improve or maintain belief in correct information. However, there is a debate on the effectiveness of fact-checking mechanisms, which raises the question of why social media companies continue to invest in them. One possible reason is that building and maintaining trust has become a key goal for these companies, and fact-checking will increase trust in the platform (Hypothesis 1). Fact-checking initiatives can bolster trust in social media companies by demonstrating their commitment to combating misinformation and promoting accuracy on their platforms. When social media companies implement robust fact-checking processes and partnerships with credible organizations, users are more likely to perceive them as responsible stewards of information. By flagging or removing false content, these companies signal their dedication to providing users with reliable and trustworthy information, thus enhancing their credibility and integrity. Moreover, transparent communication about fact-checking efforts and their impact can further reassure users and foster a sense of accountability (Ayaburi & Treku, 2020; Lan and Tung, 2024; Warner-Søderholm et al., 2018). As users increasingly prioritize accuracy and reliability in their online interactions, social media companies that prioritize fact-checking can distinguish themselves as trustworthy platforms, ultimately strengthening user trust and loyalty.

One of the study's objectives is to explore the role of the source of fact-checking (platform vs. NGO) in shaping users' perceptions of trust in the information provided on social media platforms. The effectiveness of fact-checking mechanisms employed by

social media platforms is contingent on several factors, including the source of the fact-checking. This study aims to investigate whether fact-checking by the platform itself (internal) versus an external, possibly bipartisan, organization (NGO) affects trust differently. The study hypothesizes that external sources of correction may be perceived as more objective or trustworthy than internal sources (Hypothesis 2).

Another objective of this study is to investigate the effect of the presentation of fact-checked information (labeling vs. providing corrected information) on the effectiveness of fact-checking in establishing trust. This aim examines how different methods of presenting fact-checks influence users' acceptance of the corrections and whether these methods impact trust in the platform and media. The study hypothesizes that providing context and corrected information will be more persuasive and restore trust more effectively than simple labels (Hypothesis 3).

Exposure to false or misleading information can significantly damage the trust people have in traditional news sources (Bachmann & Valenzuela, 2023; Majerczak & Strzelecki, 2022; Tandoc et al., 2021). It can cast doubt on their accuracy and reliability, leading people to question their motives and perceive bias or incompetence. As misinformation spreads unchecked across various platforms, it fuels skepticism and uncertainty about where to find trustworthy information (Flintham et al., 2018; Tanzer et al., 2021). This can prompt people to seek alternative sources for news and information, further diminishing the influence and credibility of traditional media outlets (Karduni et al., 2018). The sharing of misinformation on social media exacerbates these effects, contributing to a broader atmosphere of distrust towards media sources. Exposure to misinformation can harm people's trust in the media, which is a crucial foundation of our

information ecosystem. To restore confidence in the media, it's important to promote media literacy and transparency. Therefore, this study hypothesizes that exposure to false information has a negative impact on overall trust in media (Hypothesis 4a).

Exposure to misinformation can lead to a decrease in trust in traditional media. Therefore, fact-checking initiatives are essential to counteract this trend (Amazeen et al., 2018; Pingree et al., 2014). By implementing a thorough fact-checking process, traditional media outlets can show their commitment to accuracy and integrity in reporting. Prompt identification and correction of misinformation not only prevent the spread of false information but also demonstrate the media organization's dedication to upholding journalistic standards. Communicating the results of fact-checking transparently can further enhance trust by providing insight into the verification process and reinforcing the credibility of the information presented (Nieminen & Rapeli, 2019). By consistently prioritizing factual accuracy and accountability, traditional media outlets can mitigate the negative impact of misinformation on trust, positioning themselves as reliable sources of information in an era plagued by falsehoods and disinformation (Pingree et al., 2018). Therefore, I hypothesize that even though exposure to false information has a negative effect on trust in media, fact-checking is an effective tool to control this effect (Hypothesis 4b).

Chapter II.

Method

The study was conducted in an online setting using a Qualtrics survey. The survey consisted of several questions that were answered by the participants before they were directed towards a simulated social media feed within the Qualtrics platform. Once in the simulated social media environment, subjects were instructed to carefully read fictional social media postings on the platform. Following this interaction, participants were asked to assess their experience within this experimental social media environment on two different scales, the Platform Trust Scale (PTS) and the Media Trust Scale (MTS - post). The targeted sample for this study was a minimum of 153 adult male and female social media users between the ages of 18 to 55 years old.

Participants

A total of 285 participants took part in the study. However, 64 of them were excluded because they met at least one of the following criteria: 49 participants did not answer questions on the Platform Trust Scale and Media Trust Scale (post), meaning they didn't enter any experimental conditions. Twelve subjects spent less than 20 seconds in the study, which made it impossible for them to fully engage with the experimental condition. Additionally, 36 participants failed the attention check. It is worth noting that some of the excluded participants might have met more than one exclusion criterion, such as spending less than 20 seconds and not entering the experimental condition. The final sample consisted of 221 participants. Out of the total participants, the majority of 132 individuals identified as female, accounting for 59.7% of the sample. There were 88

participants who identified as male, making up 39.8% of the sample. Only one person chose not to disclose their gender. All participants were above 18 years old and of legal adult age at the time of their involvement in the study. Out of the total participants, 5 individuals (2.3%) were aged between 18-20, 92 individuals (41.6%) were aged between 21-29, 74 individuals (33.5%) were aged between 30-39, 36 individuals (11.8%) were aged between 40-49, 11 individuals (5.0%) were aged between 50-59, and 13 participants (5.9%) were 60 years old or older. When examining the educational level, 70.1% of the participants in this study had a bachelor's degree or a higher level of education.

Procedure

Participants will access the study online using Qualtrics, where they will be greeted with a brief introduction to the experiment and subsequent procedure. In order to prevent response bias, ethically permissible deception techniques, such as incomplete disclosure, will be employed in the introduction to disguise the research's objective and so decrease the possible adverse effect of demand characteristics. The participants will then be asked to complete a demographic questionnaire, followed by an evaluation of their social media usage (see Appendix A). Before subjects are randomly assigned to one of six experimental conditions or the control group, they will complete the General Trust Scale (see Appendix B) and Media Trust Scale (see Appendix C).

After answering both questions, participants are randomly assigned to one of the experimental conditions or the control group and will be shown a fictitious social networking website called “βetaSocial” within Qualtrics. Depending on the experimental or control group, participants will receive fictitious social media posts about climate change and vaccinations. Participants are required to thoroughly explore their fictitious

social media feed and read all postings. Within the experimental groups, false information will be presented as corrected or checked by either the platform (internal) or a non-governmental organization (external). This results in a 2x2 factorial experimental design with the independent variables "type of fact-check" and "source of correction." The control group will be exposed to the identical social media postings including correct and false information, but incorrect postings will not be labelled or corrected.

After completing the experimental conditions or the control group, all participants will be asked to evaluate the fictitious social media platform on the Platform Trust Scale (see Appendix D). The participants will then be asked to complete the Media Trust Scale once more, serving as both a test-retest and a test of the manipulation.

The deception will be revealed at the end of the study based on the APA standards for the use of deception by debriefing participants on the purpose of the research, any potential hazards of the deception methods used. Participants will have the option of contacting the researcher with further questions and opting in for a sharing of results. All responses and data will be stored in Qualtrics's online database.

Fictive Social Media Environment

The major social networking sites don't allow external access to their application programming interface (API). This means that it's not possible to manipulate or influence the displayed content. To demonstrate the experimental conditions, a fictitious social media environment called "βetaSocial" was be used. It was integrated into the web-based survey tool "Qualtrics," similar to research from Zhang et al. (2021). This method has a positive impact on the internal validity of the study by allowing better control of experimental manipulations or potential confounding variables. Additionally, using a

fictitious social media environment offers greater flexibility in designing and implementing experimental conditions, and can lead to more accurate and reliable results. This method also reduces the likelihood of bias or interference from external factors that could affect the study's outcomes.

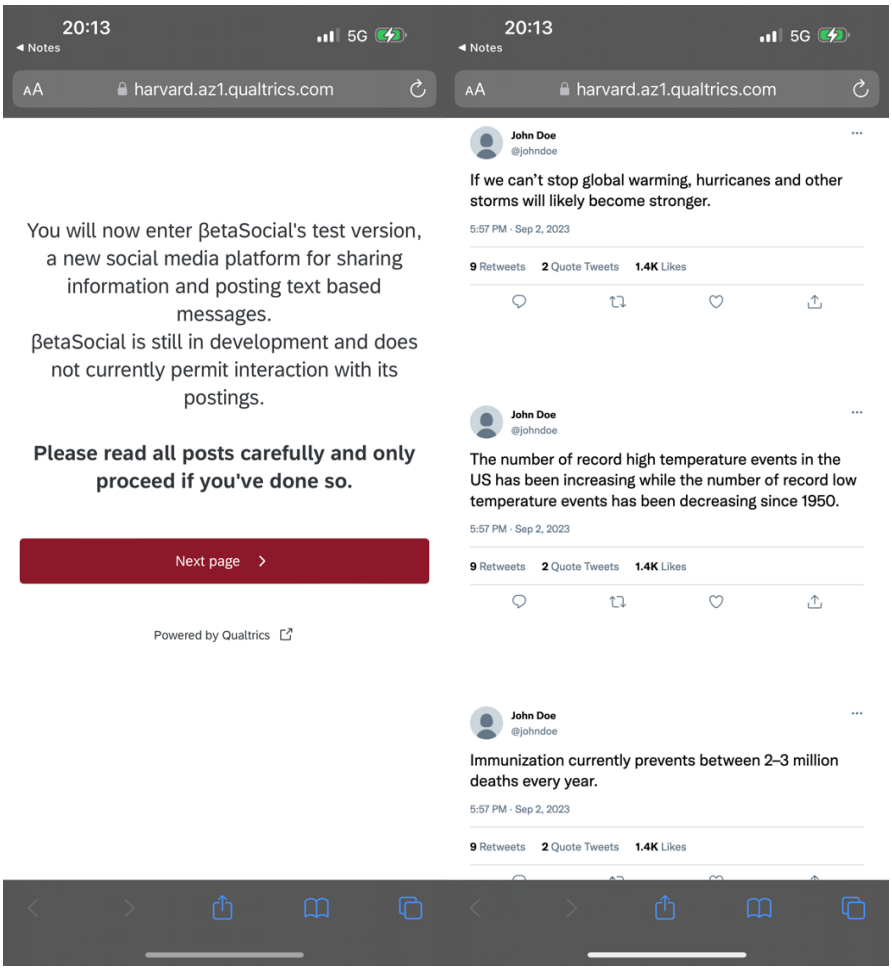

Figure 1. Fictive Social Media Environment of βetaSocial

The left side shows a screenshot of the page participants saw after answering the demographic questions, Social Media Usage Scale, General Trust Scale and Media Trust Scale (pre). Once participants clicked on "next page", they were randomly assigned to one of the experimental conditions and entered βetaSocial with the fictional postings.

Social Media Content

The content displayed in the fictitious social media ecosystem focuses on vaccinations and climate change. Both topics were selected due to their timeliness in current society and because studies have classified them as either prone to misinformation or highly debated in society. According to a study by Nisbet et al. (2015), media content can have a significant impact on climate change understanding, and misleading information on climate change is prevalent in particular media outlets, resulting in a communication gap. Additionally, van der Linden et al. (2017) demonstrated that providing respondents with knowledge regarding the scientific consensus on climate change was helpful in combatting misinformation, which may be an indication that fact-checking may also be an effective method for combating misleading information on SNS. Larson et al. (2016) consider vaccination hesitancy as a thread to global public health and link vaccination reluctance with false information, whereas Bridgman et al. (2020) indicate that relying on information from social media, which is more likely to be false than news from traditional media outlets, directly correlates with misconceptions about COVID-19. Zhang et al. (2021) found that fact-checking can positively affect vaccination willingness, which further supports the inclusion of the topic in this study.

Individual postings were created by the researcher and are based on information from the Intergovernmental Panel on Climate Change (IPCC) for climate change information and World Health Organization (WHO) for vaccination information. (see Appendix D)

Figure 2. Example of False Information Labelled by IPCC

Figure 3: Example of False Information Labelled by WHO

Figure 4. Example of False Information Labelled by βetaSocial

Figure 5. Example of False Information Labelled and Corrected by IPCC

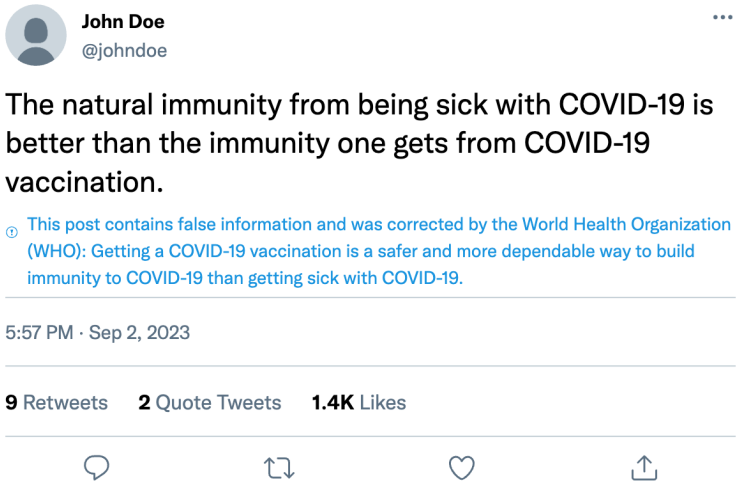

Figure 6. Example of False Information Labelled and Corrected by WHO

Figure 7. Example of False Information Labelled and Corrected by βetaSocial

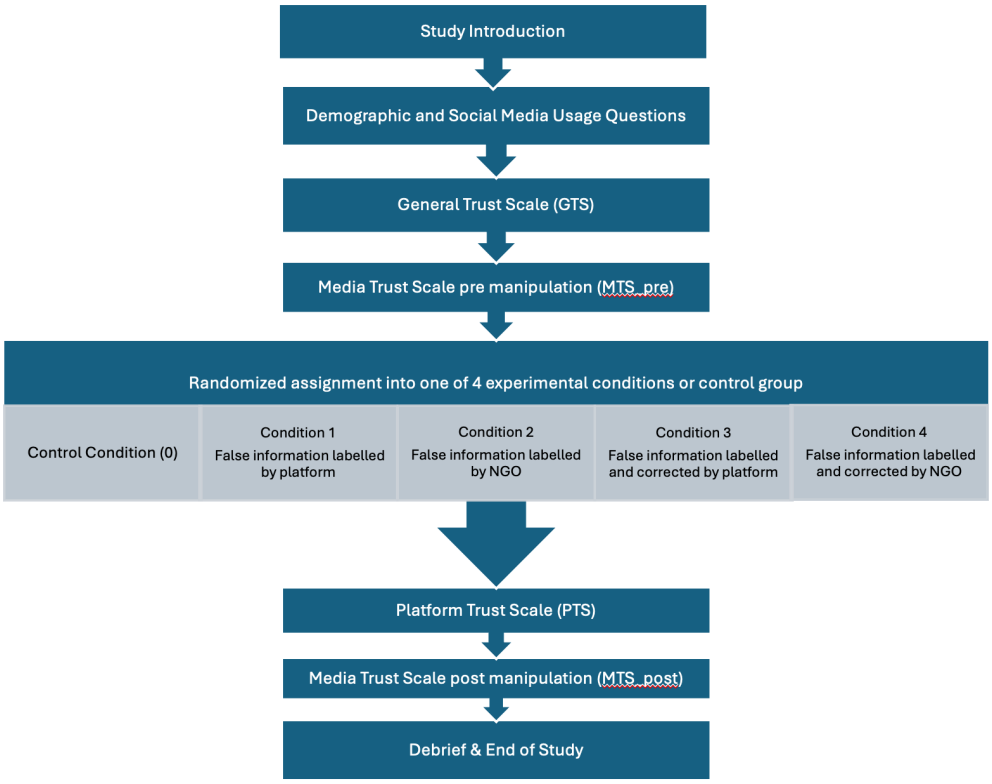

Figure 8. Study Procedure

Measures

This study employs a multi-dimensional approach to measure social media usage and trust. Firstly, I used a dual-question survey, inspired by the method of Clayton et al. (2020), to determine the frequency and types of engagement that users have with social media networks. Secondly, I incorporated the General Trust Scale (GTS) to evaluate the baseline levels of trustworthiness of individuals, which could affect their interactions and perceptions within the social media domain. Additionally, the Media Trust Scale, adapted from the Edelman Trust Barometer, aims to quantify trust in media sources before and after the experimental manipulation. Lastly, the Platform Trust Scale, a novel instrument

developed for this study, seeks to assess trust in a fictitious social media platform used during the experiment. Together, these scales form a comprehensive framework for assessing the interplay between social media usage and trust, providing insights into the potential implications for both individuals and organizations.

Social Media Usage

Similar to Clayton et al. (2020), participants respond to two questions concerning their use of various social media platforms (see Appendix B). The purpose of the first question is to collect information on which SNS the participants use and how frequently they use it. The second question asks particularly about proactive contributions, such as sharing or posting on each of the platforms specified in the first question. The self-report assessment employs a seven-point Likert scale, with values ranging from 1 = "daily" to 7 = "never." Utilization of social media is a possible confounding variable that may influence the dependent and independent variables. It is essential to control external variables to ensure internal validity.

General Trust Scale

The General Trust Scale (GTS) was developed by Yamagishi and Yamagishi (1994) and evaluates an individual's general level of trustworthiness towards others (α =.83). The Likert scale consists of six items, ranging from 1 = "completely disagree" to 5 = "strongly agree" (see Appendix B). The GTS was validated in cross-cultural settings by Jasielska et al. (2021) and will be used as a pre-treatment measure to account for potential differences in participants' general trust levels, which could influence both independent and dependent variables.

Media Trust Scale

A question from the Edelman Trust Barometer, an annual global poll of over 36,000 respondents from 28 countries released by the US public relations and marketing company Edelman (2023), seeks to assess participants' levels of trust in general media (see Appendix C). On a 9-point Likert scale ranging from 1 = "low trust" to 9 = "high trust," participants are asked to rank their trust in various media sources for news and information, with 1 = "low trust" and 9 = "high trust." According to Edelman's (2023) instructions, the top four box scores are summarized as "trust." The assessment will be administered before the subjects enter the experimental condition and again at the end of the study. This test-retest will serve as a manipulation test. Differences between pre- and post-experimental condition scores suggest that experimental manipulation influences the dependent variable "trust."

Platform Trust Scale

This scale will assess participants' trust in the fictitious social media environment employed in the experimental condition (see Appendix D). It was developed by the researcher and is based on the GTS (Yamagishi and Yamagishi, 1994) with a 5-point Likers scale ranging from 1 = "completely disagree" to 5 = "strongly agree." The 12 items of the Platform Trust Scale are based on GTS items and items from Cummings and Bromiley's (1996) Organizational Trust Inventory (OTI), which evaluates trust between individuals and organizations ($\alpha = .76$). Trust is characterized by the expectation that an individual or entity will act with integrity in honoring both stated and understood promises, will maintain transparency during the formulation of these agreements, and will refrain from exploiting others even when circumstances may allow for it (Cummings

& Bromiley, 1996). Items from both scales were adapted to the context of platform trust, i.e., they ask about general trust in the given platform or organizational trust in a specific organization, which in the experimental conditions is the SNS.

Chapter III.

Results

The following analyses were conducted using IBM SPSS Statistics 29. The analysis of the data reveals several key findings regarding the relationship between social media usage and trust in media platforms. Firstly, a positive correlation was identified between social media usage and posting behaviors, indicating that individuals who frequently use social media are also more likely to engage in posting activities. However, the relationship between these behaviors and trust in media and platforms showed varying degrees of correlation, suggesting a complex interaction that warrants further exploration. Significantly, changes in media trust following exposure to fact-checking interventions were found to be closely linked to both initial levels of media trust and trust in the social media platforms themselves. This relationship was characterized by negative correlations between trust scores and social media usage/posting scores, highlighting areas for potential in-depth study.

Table 1. Descriptive Statistics

Variables	Mean	*SD*	*N*
Social Media Usage Score	3.3564	1.22271	221
Social Media Post Score	5.7830	1.15894	221
General Trust Score	3.3250	.65022	221
Media Trust Score (pre)	4.3139	1.25128	221
Platform Trust Score	2.9084	.57699	221
Media Trust Score (post)	4.2424	1.28673	213

Table 2. Descriptive Statistics by Condition

Variables	Condition	Mean	*SD*
Social Media Usage Score	Control	3.3595	1.35705
	False Information (platform)	3.2435	1.19399
	False Information (NGO)	3.5354	1.27349
	False Information - corrected (platform)	3.4000	1.19406
	False Information - corrected (NGO)	3.2591	1.12194
	Total	3.3564	1.22271
Social Media Post Score	Control	5.8636	1.12789
	False Information (platform)	5.6978	1.21883
	False Information (NGO)	5.9012	1.09410
	False Information - corrected (platform)	5.7652	1.15724
	False Information - corrected (NGO)	5.7000	1.22265
	Total	5.7830	1.15894
General Trust Score	Control	3.3864	.65963
	False Information (platform)	3.4493	.54758

	False Information (NGO)	3.2114	.68315
	False Information - corrected (platform)	3.3261	.62554
	False Information - corrected (NGO)	3.2386	.72817
	Total	3.3250	.65022
Media Trust Score (pre)	Control	4.6648	1.17964
	False Information (platform)	4.1223	1.28222
	False Information (NGO)	3.9756	1.17434
	False Information - corrected (platform)	4.4511	1.35131
	False Information - corrected (NGO)	4.3352	1.18732
	Total	4.3139	1.25128
Platform Trust Score	Control	2.6534	.51242
	False Information (platform)	2.9330	.52104
	False Information (NGO)	2.9045	.51625
	False Information - corrected (platform)	2.9551	.63805
	False Information - corrected (NGO)	3.0928	.61506
	Total	2.9084	.57699
Media Trust Score (post)	Control	4.6192	1.13289
	False Information (platform)	4.0994	1.41744
	False Information (NGO)	3.8656	1.17887
	False Information - corrected (platform)	4.3214	1.35029
	False Information - corrected (NGO)	4.2841	1.26427
	Total	4.2424	1.28673

Table 3. Correlations

Variables		SMU	SMP	GTS	MTS (pre)	PTS	MTS (post)
SMU	Pearson Correlation	1	.413**	.081	-.342**	-.051	-.314**
	Sig. (2-tailed)		<.001	.232	<.001	.452	<.001
	N	221	221	221	221	221	213
SMP	Pearson Correlation	.413**	1	.032	-.235**	-.070	-.242**
	Sig. (2-tailed)	<.001		.635	<.001	.302	<.001
	N	221	221	221	221	221	213
GTS	Pearson Correlation	.081	.032	1	.085	.003	.086
	Sig. (2-tailed)	.232	.635		.209	.964	.211
	N	221	221	221	221	221	213
MTS (pre)	Pearson Correlation	-.342**	-.235**	.085	1	.138*	.896**
	Sig. (2-tailed)	<.001	<.001	.209		.040	<.001
	N	221	221	221	221	221	213
PTS	Pearson Correlation	-.051	-.070	.003	.138*	1	.145*
	Sig. (2-tailed)	.452	.302	.964	.040		.034
	N	221	221	221	221	221	213
MTS (post)	Pearson Correlation	-.314**	-.242**	.086	.896**	.145*	1
	Sig. (2-tailed)	<.001	<.001	.211	<.001	.034	
	N	213	213	213	213	213	213

**. *Correlation is significant at the 0.01 level (2-tailed).*
*. *Correlation is significant at the 0.05 level (2-tailed).*

Testing Hypothesis 1: Fact-Checking Will Increase Trust in the Platform.

I set out to test the hypothesis that fact-checking would increase trust in social media platforms. To do this, I conducted a one-way ANOVA which revealed significant effects of condition on participants' scores on the Platform Trust Score ($F = 3.52$, $p = .008$). To understand which specific fact-checking conditions affected the trust scores, I delved into the multiple comparisons section. Here, using the Tukey HSD post-hoc test, I compared the control group against each of the fact-checking conditions. It was clear that the 'False Information - corrected (NGO)' condition significantly increased the Platform Trust Score when compared to the control group, with a mean difference of -0.43 and a *p*-value of .003. However, the other fact-checking conditions, such as "False Information (platform)," "False Information (NGO)," and "False Information - corrected (platform)," did not show a significant difference from the control group in terms of improving trust scores.

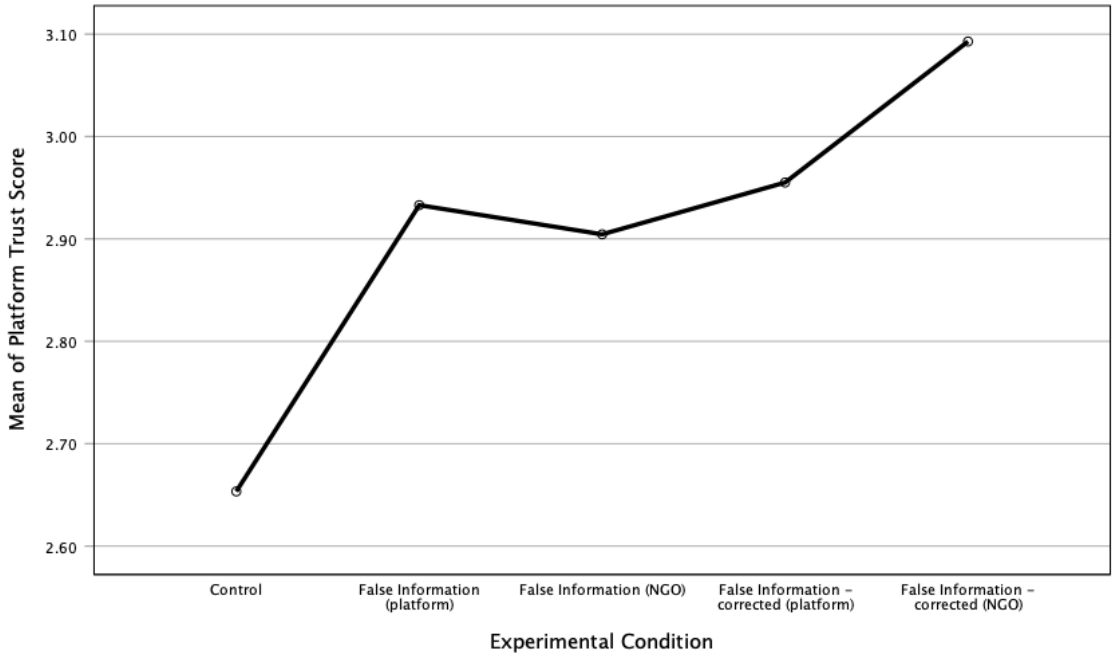

Figure 9. Mean of Platform Trust Score per Experimental Condition

Shows the Mean Scores on the PTS for each experimental condition. A higher score on the PTS indicates higher trust levels.

Testing Hypothesis 2 and 3

I hypothesized that external sources of correction through NGOs will be perceived as more trustworthy than internal sources, i.e. correcting from the platform, and that providing context and corrected information will be more persuasive and restore trust more effectively than simply labeling false information. To test these hypotheses, I conducted a 2 (Source: External vs. Internal) X 2 (Method: Flagging vs. Correction) ANOVA, excluding the control condition.

There was not significant main effect of source ($F = .396$, $p = .530$). This non-significant result indicates that the data does not provide sufficient evidence to support Hypothesis 2. The mean trust scores for the platform and NGO sources were *2.944* and

2.999 respectively. Therefore, based on this dataset, I cannot conclude that external sources of correction through NGOs are perceived as more trustworthy than internal platform sources.

Similarly, there was no main effect of method of flagging misinformation, F value of 1.472 with a significance level of .227. The mean trust scores for flagging alone and flagging with correction were 2.919 and 3.024 respectively. Consequently, the data does not provide sufficient evidence to confirm Hypothesis 3, as the improvement in trust from flagging alone to flagging with correction does not reach statistical significance. Besides that, there was no significant interaction between the different flagging conditions and source conditions ($F(1, 176) = .918$, $p = .339$).

Testing Hypothesis 4a

I hypothesize that being exposed to false information has a negative impact on the overall trust people have in the media. However, even though false information can negatively affect trust in media, fact-checking can be an effective tool to control this effect. Based on the within-subjects ANOVA (table 5) of the effects of exposure to false information on media trust, I have found a significant main effect of time ($F(1, 208) = 4.58$, $p = .034$), which supports my first hypothesis (Hypothesis 4a) that false information can have a negative impact on trust in media. This finding shows that false information has the potential to erode the foundation of trust that individuals have in media outlets. Moving on to my second hypothesis (Hypothesis 4b), which suggested that fact-checking could be an effective countermeasure to the negative effects of false information, the statistical evidence was not as clear. The interaction between time and condition was not statistically significant ($F(4, 208) = .37$, $p = .828$), which led me to conclude that fact-

checking did not significantly reduce the negative impact of false information on trust. These results suggest that while trust in media can be influenced by misinformation, the effectiveness of fact-checking as a tool to remedy the negative effects of false information is not conclusive.

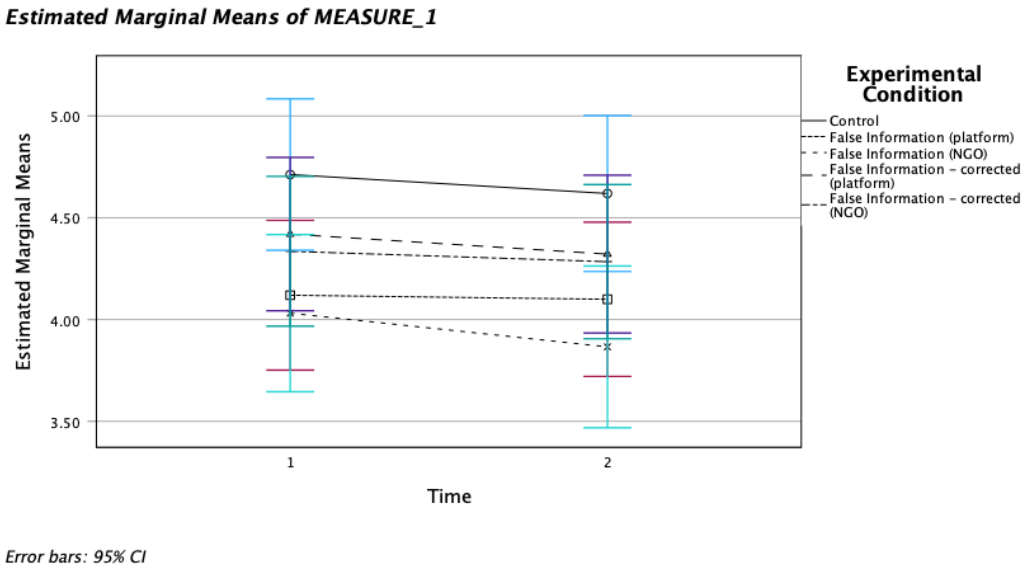

Figure 10. Estimated Marginal Means of Media Trust Score (Pre & Post)

Represents the differences of means for each experimental condition from before entering the experimental condition (1 - MTS pre) and after (2 – MTS post).

Table 4. One-Way ANOVA Pairwise Comparison Results

(I) Experimental Condition	(J) Experimental Condition	Mean Difference (I-J)	Std. Error	Sig.
Control	False Information (platform)	-.27956*	.11897	.020
	False Information (NGO)	-.25106*	.12247	.042
	False Information - corrected (platform)	-.30166*	.11897	.012
	False Information - corrected (NGO)	-.43939*	.12029	<.001
False Information (platform)	Control	.27956*	.11897	.020
	False Information (NGO)	.02850	.12118	.814
	False Information - corrected (platform)	-.02210	.11764	.851
	False Information - corrected (NGO)	-.15983	.11897	.181
False Information (NGO)	Control	.25106*	.12247	.042
	False Information (platform)	-.02850	.12118	.814
	False Information - corrected (platform)	-.05060	.12118	.677
	False Information - corrected (NGO)	-.18833	.12247	.126
False Information - corrected (platform)	Control	.30166*	.11897	.012
	False Information (platform)	.02210	.11764	.851
	False Information (NGO)	.05060	.12118	.677
	False Information - corrected (NGO)	-.13773	.11897	.248
False Information - corrected (NGO)	Control	.43939*	.12029	<.001
	False Information (platform)	.15983	.11897	.181
	False Information (NGO)	.18833	.12247	.126
	False Information - corrected (platform)	.13773	.11897	.248

Table 5. Within-Subjects ANOVA

Source		*df*	Mean Square	*F*	Sig.
Time	Sphericity Assumed	1	.779	4.576	.034
	Greenhouse-Geisser	1.000	.779	4.576	.034
	Huynh-Feldt	1.000	.779	4.576	.034
	Lower-bound	1.000	.779	4.576	.034
Time * Condition	Sphericity Assumed	4	.063	.372	.828
	Greenhouse-Geisser	4.000	.063	.372	.828
	Huynh-Feldt	4.000	.063	.372	.828
	Lower-bound	4.000	.063	.372	.828
Error(Time)	Sphericity Assumed	208	.170		
	Greenhouse-Geisser	208.000	.170		
	Huynh-Feldt	208.000	.170		
	Lower-bound	208.000	.170		

Chapter IV.

Discussion

This study investigated how fact-checking interventions affect users' trust in social media. The research showed that the type and source of fact-checking play a significant role in determining trust outcomes. Although the findings suggest that fact-checking can influence trust in media, the effects differ depending on the situation and are not consistent. The study aimed to understand the complex dynamics of social media use, posting behavior, and levels of trust in media platforms, focusing particularly on the impact of fact-checking interventions. The results revealed a nuanced view of these relationships, emphasizing the multifaceted nature of trust in the digital era.

According to Hypothesis 1, fact-checking can increase trust in media platforms. The data partially supports this hypothesis, as the fact-checking condition “False Information - corrected (NGO)” was found to significantly improve trust in platforms as compared to the control group. This suggests that the credibility of the fact-checking source is crucial, as corrections from an NGO were more effective than those from the platform itself. This finding has significant implications for social media platforms, indicating that partnering with external, reputable organizations may be an effective strategy to combat misinformation.

Contrary to Hypothesis 2, there was no significant difference found between the trustworthiness of external sources (NGOs) and internal platform sources for fact-checking. This could imply that the credibility of fact-checking may not solely depend on the source but may also be influenced by other factors, such as the presentation of the facts, the perceived impartiality of the source, or the pre-existing biases of users.

The data did not support Hypothesis 3, which proposed that providing context and corrections would be more effective than simply labeling information as false. The absence of a significant difference could be attributed to various factors, such as the likelihood that users are already skeptical or have formed their opinions on certain subjects, which may reduce the impact of additional context. This emphasizes the challenge that platforms face in not only identifying false information but also in persuading users to reconsider their perspectives.

The study conducted an examination of the impact of false information on media trust (Hypothesis 4a), which revealed a significant negative effect over time. This highlights the detrimental influence of misinformation on public trust. However, the data did not provide robust support for the efficacy of fact-checking as a tool to mitigate this effect (Hypothesis 4b). The lack of a significant interaction suggests that while fact-checking may not reverse the damage caused by exposure to misinformation, it does not necessarily exacerbate distrust either. It's possible that fact-checking serves more as a preventative measure rather than a restorative one.

The study emphasizes the need for ongoing research on how trust is affected in the digital information ecosystem. While fact-checking interventions are promising, it is essential to consider their implementation and the context in which users receive them. The study shows a correlation between social media usage and posting behavior, as well as varying degrees of trust in media, indicating that individual differences and user engagement levels are crucial factors to consider in future research.

Furthermore, the study uncovered negative associations between trust levels and social media usage or posting frequency. These findings suggest that excessive usage of

social media could lead to skepticism or fatigue, which could explain why some types of fact-checking did not significantly improve trust levels after exposure to false information. To sum up, this study provides a fundamental understanding of the complex and context-dependent relationship between social media behavior and trust in the media. It also highlights the need for further research into the effectiveness of different fact-checking methods and sources, as well as the psychological factors that underpin trust in the digital era. Ultimately, the goal is to cultivate an informed and critical-thinking society in an age where information, both accurate and inaccurate, is readily available.

Theoretical Implications

The present study investigates the impact of fact-checking on social media users' trust and sheds light on several theoretical implications that align with and extend the current literature on misinformation, trust, and media integrity. The research findings suggest that fact-checking, especially when conducted by external organizations such as NGOs, can significantly enhance trust in social media platforms. This supports the hypothesis that fact-checking is an effective tool to boost user trust, which is consistent with the works of Ayaburi & Treku (2020), Lan and Tung (2024), and Warner-Søderholm et al. (2018), who argue for the positive impact of content moderation strategies on trust. The study also contributes to the discourse on the importance of source credibility in the fact-checking process, which resonates with the works of Chan et al. (2017) and Zhang et al. (2021), highlighting the differential impact of fact-checking based on the source's perceived authority and neutrality. The research underscores the superiority of external over internal corrections, which advocates for the strategic

engagement of reputed third-party organizations in fact-checking initiatives to leverage their credibility for trust enhancement.

However, the present findings challenge the notion of a straightforward positive relationship between fact-checking and media trust, as suggested by some previous research (e.g., Brenan, 2020; Lan & Tung, 2024). The consistent change in media trust scores across different experimental conditions, regardless of fact-checking intervention, suggests a more complex and nuanced effect of fact-checking on media trust. This observation invites a reevaluation of the direct impact of fact-checking on media trust perceptions, hinting at the multifaceted nature of trust dynamics and the potential for other intervening factors not captured in this study. The study's insights into the limited role of social media usage and general trust levels in moderating the impact of fact-checking interventions align with recent discussions on the complex interplay between user engagement, trust predispositions, and information credibility (Pennycook & Rand, 2017; Shearer & Gottfried, 2017). These findings contribute to a broader understanding of trust as a multidimensional construct influenced by a variety of factors beyond mere exposure to or correction of misinformation.

In conclusion, this research enriches the existing body of literature by providing a nuanced understanding of the impact of fact-checking on trust in social media platforms and the media at large. The study offers valuable insights for scholars, policymakers, and practitioners aiming to navigate the challenges of misinformation in the digital age by elucidating the significance of source credibility and the context-dependent nature of fact-checking's effectiveness. The findings advocate for a comprehensive approach to trust restoration that goes beyond simple fact-checking to encompass the broader socio-

technical and cognitive dimensions influencing public trust in media and information sources. As we move forward, it becomes imperative to continue exploring the intricate dynamics of misinformation, fact-checking, and trust within the ever-evolving landscape of social media and digital communication.

Practical Implications

The findings of my study have practical implications in several areas, such as policymaking, governance of social media platforms, and business practices. These findings provide valuable insights to enhance trust, not only in social media platforms but also in the wider media landscape. My research highlights the efficacy of fact-checking interventions, especially those performed by external organizations. It suggests that social media companies can benefit from forming partnerships with reputable third-party fact-checking organizations. This collaboration will not only enhance the credibility and neutrality of the fact-checking process but also boost user trust in the platforms.

To ensure transparency and accountability, social media platforms are encouraged to implement clear fact-checking policies and openly share their criteria for identifying misinformation. Such openness about their efforts to combat misinformation and the involvement of external partners in these efforts could significantly increase user confidence in the content moderation process. Additionally, the study emphasizes the importance of user education initiatives to inform users about fact-checking and guide them in critically evaluating information. This will foster a more informed user community. Policymakers and regulators can develop regulatory frameworks that support fact-checking practices on social media. This support will promote the participation of external, credible organizations and ensure transparency, impartiality, and consistency in

fact-checking processes. Supporting independent fact-checking organizations through funding and resources is another crucial area. Investing in media literacy programs to educate the public on assessing information sources critically is also essential.

Businesses and advertisers should consider the implications of their advertising placements on social media platforms. They should focus on platforms committed to fighting misinformation through credible fact-checking to enhance brand safety and align with ethical advertising principles. Integrating support for accurate information and fact-checking into corporate social responsibility initiatives could also play a vital role in promoting accurate information on topics relevant to their industry.

In conclusion, my study highlights the critical role of credible, transparent, and effective fact-checking in building trust in social media platforms and the media at large. Adopting a strategic approach that involves multiple stakeholders in fact-checking and misinformation management can help create a more informed, trustworthy, and resilient digital information ecosystem.

Limitations

This study, while contributing valuable insights into the dynamics of fact-checking interventions on social media platforms, encounters several limitations that necessitate careful consideration. Firstly, the absence of a direct measure of the interventions' effectiveness in altering beliefs represents a significant limitation. The research's focus was primarily on the impact of fact-checking on trust towards the platform, without an explicit evaluation of how these interventions influenced users' beliefs regarding the accuracy of information. Consequently, it remains unclear whether

the observed increase in platform trust correlates with an enhanced ability to discern false from accurate information.

Secondly, the study's reliance on specific external sources for fact-checking interventions raises questions about the generalizability of the findings. The effectiveness of these interventions might not extend across diverse contexts or platforms, where other sources of fact-checking could be employed. This specificity limits the applicability of the study's conclusions beyond the examined scenarios.

Furthermore, the exploration of internal fact-checking's effectiveness was conducted within a hypothetical platform context, potentially not reflecting the dynamics present in platforms where users have established levels of trust. The assumption suggests that internal fact-checking might exhibit higher efficacy in environments where the platform already enjoys a degree of user trust. This could also apply to the relevance of the fact-checking source. The effectiveness of a fact-check may depend on the source from which it comes. This is particularly relevant in real world scenarios involving social media platforms that may be either untrusted or highly trusted. Corrections from a highly trusted platform may be more effective than those from an untrusted source.

Notably, the simulated setting was solely text-oriented, resembling the platform X (formerly Twitter), and lacked any photo- or videographic elements commonly found on sites like YouTube, Instagram, or TikTok. This methodological choice restricts the findings' applicability to real-world scenarios, as the controlled experimental setting may not accurately reflect the complex dynamics of social media use. Additionally, the investigation's focus on immediate responses to fact-checking does not account for the potential long-term effects on users' trust and behaviors, omitting an exploration of how

perceptions of trustworthiness may evolve or persist over time. The study's narrow focus on internal versus external sources of fact-checking, along with its limitation to specific topics like climate change and vaccinations, may not encompass the wide variety of sources and misinformation subjects present on social media. Furthermore, the homogeneity of the participant sample in terms of demographics and social media habits could limit the study's insights into diverse user responses to fact-checking efforts. Psychological and cultural factors, which could significantly influence individuals' openness to fact-checking, were not extensively examined, leaving out potential mediators of fact-checking's effectiveness.

Lastly, the approach to measuring general media trust in this study, which aggregates trust across various media types, may obscure the specific effects of fact-checking on particular media forms. This broad measurement approach could conceal nuanced effects, potentially misleading the interpretation of fact-checking interventions' impact on trust in media.

Future Directions

Considering the limitations identified in this study, there are several avenues for future research that promise to enrich our understanding of fact-checking's efficacy on social media platforms. These directions not only aim to address the gaps in the current literature but also to explore new territories that could provide deeper insights into the mechanisms and outcomes of fact-checking interventions. Looking forward, future research could benefit from longitudinal studies to observe the long-term impacts of fact-checking on trust and misinformation beliefs, providing a deeper understanding of the persistence of these interventions. Expanding the scope to include a broader array of fact-

checking sources and a wider variety of misinformation topics could offer a more comprehensive view of the effectiveness of fact-checking across different contexts. Cross-cultural studies could illuminate the influence of cultural differences on the perception and success of fact-checking, guiding the development of tailored misinformation combat strategies worldwide.

A pivotal area for future investigation involves the execution of field experiments incorporating A/B testing methodologies with actual users on real social networking sites. This approach would allow for a more nuanced analysis of how different fact-checking methods influence user trust and belief systems in a naturalistic setting. By comparing the effectiveness of various fact-checking strategies in live environments, researchers can generate evidence-based recommendations for social media platforms seeking to mitigate the spread of misinformation among their user base.

Further exploration into the psychological mechanisms underlying trust in fact-checking interventions offers another promising research direction. Delving into why certain fact-checking approaches are more successful than others in enhancing trust could unveil critical insights into user psychology. Such an understanding would be instrumental in designing fact-checking mechanisms that are not only more persuasive but are also tailored to address the cognitive biases and heuristics that influence information processing on social media.

A comprehensive examination that assesses both the trust and effectiveness of fact-checking interventions in correcting false beliefs concurrently represents a significant gap in the current literature. A parallel study design would facilitate a holistic understanding of the dual objectives of fact-checking: building trust in the platform and

effectively correcting misinformation. By elucidating the relationship between these two outcomes, researchers can identify whether interventions that increase trust simultaneously contribute to a more informed and discerning user population.

Expanding the scope of fact-checking sources and the variety of misinformation topics presents another fertile ground for research. Future studies could investigate the impact of fact-checking interventions across a broader spectrum of sources, including both internal and external, and across different thematic areas. Such research would contribute to a more comprehensive understanding of the generalizability and specificity of fact-checking's effectiveness, providing insights into how different contexts and source types influence user perceptions and behavior.

Incorporating psychological traits and behavioral metrics into future studies could yield richer insights into the individual differences that affect the reception of fact-checking initiatives. Additionally, exploring the impact of media literacy education alongside fact-checking could reveal synergistic approaches to improving information discernment among social media users. Investigating algorithmic and technological solutions, such as AI-driven fact-checking tools, could further enhance the scalability and precision of efforts to counter misinformation.

By pursuing these future directions, researchers can significantly advance our understanding of the complex dynamics surrounding fact-checking on social media. Such efforts will be instrumental in devising more effective strategies to combat misinformation, ultimately contributing to the creation of a more informed and critically engaged digital public sphere.

This study sheds light on how fact-checking interventions affect trust in social media platforms, highlighting the complex relationship between the source of corrections, the type of correction, and users' pre-existing trust levels. Although the findings provide initial insights into the potential of fact-checking to improve platform trust, there are significant limitations, and further research is necessary. Future studies, using field experiments, psychological analyses, and longitudinal research, can deepen our understanding of effective fact-checking strategies. By expanding the range of fact-checking sources, contexts, and examining the long-term impacts on user trust and misinformation correction, subsequent research can offer valuable guidance for social media platforms that want to fight the pervasive problem of misinformation. Ultimately, improving the effectiveness of fact-checking interventions is essential for creating an informed, discerning, and resilient digital public sphere in a social media-dominated era.

Appendix A. Social Media Usage

How often do you use each of the following:

	Daily	A few times a week	Once a week	A few times a month	Once a month	Less frequently than once a month	never
Facebook							
X (formerly Twitter)							
YouTube							
TikTok							
Instagram							

How often do you share/post/upload content (video, photo, text) on each of the following:

	Daily	A few times a week	Once a week	A few times a month	Once a month	Less frequently than once a month	never
Facebook							
X (formerly Twitter)							
YouTube							
TikTok							
Instagram							

Appendix B. General Trust Scale

Using the following scale, please indicate how much you agree or disagree with the following statements:

1 **Strongly Disagree**	**2** **Disagree**	**3** **Neutral**	**4** **Agree**	**5** **Strongly Agree**

1.) Most people are basically honest.
2.) Most people are trustworthy
3.) Most people are basically good and kind.
4.) Most people are trustful of others
5.) I am trustful
6.) Most people will respond in kind when they are trusted by others

Appendix C. Media Trust Scale

When looking for general news and information, how much would you trust each type of source for general news and information?

	1 low trust	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9 High trust**
Facebook									
Twitter									
TikTok									
Instagram									
YouTube									
Newspaper									
TV									
Search Engine									

Appendix D. Platform Trust Scale

Using the following scale, please indicate how much you agree or disagree with the following statements:

1 **Strongly Disagree**	**2** **Disagree**	**3** **Neutral**	**4** **Agree**	**5** **Strongly Agree**

1 The platform trustworthy
2 The platform shows valid information
3 The platform's content is unbiased
4 The platform's content covers important topics
5 The platform has effective measures to regulate false information
6 The platform is prone to censorship *(reverse-coded)*
7 The platform creates a safe space
8 The platform is reliable
9 The platform will keep its word
10 The platform might take advantage of its users *(reverse-coded)*
11 The platform might manipulate others *(reverse-coded)*
12 I trust the platform

References

Amazeen, M. A., Thorson, E., Muddiman, A., & Graves, L. (2018). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly*, *95*(1), 28–48. https://doi.org/10.1177/1077699016678186

Amran, M. A. (2016). Trust-repair discourse on Facebook's Cambridge Analytica scandal. *Elite Journal*, *Vol.1*(No.2). http://journal.fib.uho.ac.id/index.php/elite/article/view/1533/1281

Ayaburi, E. W., & Treku, D. N. (2020). Effect of penitence on social media trust and privacy concerns: The case of Facebook. *International Journal of Information Management*, *50*, 171–181. https://doi.org/10.1016/j.ijinfomgt.2019.05.014

Bachmann, I., & Valenzuela, S. (2023). Studying the downstream effects of fact-checking on social media: Experiments on correction formats, belief accuracy, and media trust. *Social Media + Society*, *9*(2), 20563051231179694. https://doi.org/10.1177/20563051231179694

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, *47*(2), 241–262. https://doi.org/10.1017/S0007123415000186

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE*, *10*(2), e0118093. https://doi.org/10.1371/journal.pone.0118093

Bode, L., & Vraga, E. K. (2018). See something, say something: Correction of global health misinformation on social media. *Health Communication*, *33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

Brenan, M. (2020, September 30). *Americans remain distrustful of mass media*. Gallup.Com. https://news.gallup.com/poll/321116/americans-remain-distrustful-mass-media.aspx

Bridgman, A., Merkley, E., Loewen, P. J., Owen, T., Ruths, D., Teichmann, L., & Zhilin, O. (2020). The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-028

Bright, L. F., Lim, H. S., & Logan, K. (2021). "Should I post or ghost?": Examining how privacy concerns impact social media engagement in us consumers. *Psychology & Marketing*, *38*(10), 1712–1722. https://doi.org/10.1002/mar.21499

Byrne, S., & Hart, P. S. (2009). The boomerang effect a synthesis of findings and a preliminary theoretical framework. *Annals of the International Communication Association*, *33*(1), 3–37. https://doi.org/10.1080/23808985.2009.11679083

Carson, A., Gibbons, A., Martin, A., & Phillips, J. B. (2022). Does third-party fact-checking increase trust in news stories? An Australian case study using the "sports rorts" affair. *Digital Journalism*, *10*(5), 801–822. https://doi.org/10.1080/21670811.2022.2031240

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Chou, W.-Y. S., Gaysynsky, A., & Vanderpool, R. C. (2021). The COVID-19 misinfodemic: Moving beyond fact-checking. *Health Education & Behavior*, *48*(1), 9–13. https://doi.org/10.1177/1090198120980675

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

Cummings, L. L., & Bromiley, P. (1996). The Organizational Trust Inventory (OTI): Development and validation. In R. M. Kramer & T. R. Tyler (Eds.), Trust in organizations: Frontiers of theory and research (pp. 302–330). Sage Publications, Inc. https://doi.org/10.4135/9781452243610.n15

Dimock, M. (2019, June 5). An update on our research into trust, facts and democracy. *Pew Research Center*. https://www.pewresearch.org/2019/06/05/an-update-on-our-research-into-trust-facts-and-democracy/

Edelman. (2023). *Edelman trust barometer—Navigating a polarized world*. https://www.edelman.com/trust/2023/trust-barometer

Einwiller, S. A., & Kamins, M. A. (2008). Rumor has it: The moderating effect of identification on rumor impact and the effectiveness of rumor refutation. *Journal of Applied Social Psychology*, *38*(9), 2248–2272. https://doi.org/10.1111/j.1559-1816.2008.00390.x

Fawzi, N. (2020). Objektive Informationsquelle, Watchdog und Sprachrohr der Bürger? Die Bewertung der gesellschaftlichen Leistungen von Medien durch die Bevölkerung. [Objective source of information, watchdog and mouthpiece for citizens? The population's assessment of the social performance of media.] *Publizistik*, *65*(2), 187–207. https://doi.org/10.1007/s11616-020-00572-w

Flintham, M., Karner, C., Bachour, K., Creswick, H., Gupta, N., & Moran, S. (2018). Falling for fake news: Investigating the consumption of news via social media. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3173574.3173950

Gallup and Knight Foundation. (2020). *American views 2020: Trust, media and democracy*. Gallup and Knight Foundation. https://knightfoundation.org/reports/american-views-2020-trust-media-and-democracy/

Garrett, B. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Harvard University Press. https://doi.org/10.4159/harvard.9780674060982

Garrett, R. K., & Poulsen, S. (2019). Flagging Facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings. *Journal of Computer-Mediated Communication*, *24*(5), 240–258. https://doi.org/10.1093/jcmc/zmz012

Goodman, J., & Carmichael, F. (2020, May 29). Coronavirus: Bill Gates 'microchip' conspiracy theory and other vaccine claims fact-checked. *BBC News*. https://www.bbc.com/news/52847648

Hanitzsch, T., & Vos, T. P. (2018). Journalism beyond democracy: A new look into journalistic roles in political and everyday life. *Journalism*, *19*(2), 146–164. https://doi.org/10.1177/1464884916673386

Harari, Y. N. (2018). *21 lessons for the 21st century* (First edition). Spiegel & Grau.

Ipsos. (2023). *Ipsos global trustworthiness monitor: Stability in an unstable world*. https://www.ipsos.com/sites/default/files/ct/publication/documents/2023-01/ipsos-global-trustworthiness-monitor-stability-in-an-unstable-world.pdf

Jasielska, D., Rogoza, R., Zajenkowska, A., & Russa, M. B. (2021). General trust scale: Validation in cross-cultural settings. *Current Psychology*, *40*(10), 5019–5029. https://doi.org/10.1007/s12144-019-00435-2

Jenkins, M. (2018, October 12). *Understanding disinformation is impossible without understanding your audience*. Medium. https://medium.com/@mjenkins/understanding-disinformation-is-impossible-without-understanding-your-audience-7c64d3280494

Jones, D. A. (2004). Why Americans don't trust the media: A preliminary analysis. *Harvard International Journal of Press/Politics*, *9*(2), 60–75. https://doi.org/10.1177/1081180X04263461

Karduni, A., Wesslen, R., Santhanam, S., Cho, I., Volkova, S., Arendt, D., Shaikh, S., & Dou, W. (2018). Can you verifi this? Studying uncertainty and decision-making about misinformation using visual analytics. *Proceedings of the International*

AAAI Conference on Web and Social Media, *12*(1). https://doi.org/10.1609/icwsm.v12i1.15014

Kim, A., & Dennis, A. R. (2019). Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, *43*(3), 1025–1039. https://doi.org/10.25300/MISQ/2019/15188

Lan, D. H., & Tung, T. M. (2024). Exploring fake news awareness and trust in the age of social media among university student TikTok users. *Cogent Social Sciences*, *10*(1), 2302216. https://doi.org/10.1080/23311886.2024.2302216

Larson, H. J., de Figueiredo, A., Xiahong, Z., Schulz, W. S., Verger, P., Johnston, I. G., Cook, A. R., & Jones, N. S. (2016). The state of vaccine confidence 2016: Global insights through a 67-country survey. *EBioMedicine*, *12*, 295–301. https://doi.org/10.1016/j.ebiom.2016.08.042

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Majerczak, P., & Strzelecki, A. (2022). Trust, media credibility, social ties, and the intention to share towards information verification in an age of fake news. *Behavioral Sciences*, *12*(2), 51. https://doi.org/10.3390/bs12020051

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, *34*(4), 754–772. https://doi.org/10.1016/j.clsr.2018.05.017

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, *35*(2), 196–219. https://doi.org/10.1080/10584609.2017.1334018

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2), 1–25. https://doi.org/10.1145/1985347.1985353

Melki, J., Tamim, H., Hadid, D., Makki, M., El Amine, J., & Hitti, E. (2021). Mitigating infodemics: The relationship between news exposure and trust and belief in COVID-19 fake news and social media spreading. *PLOS ONE*, *16*(6), e0252830. https://doi.org/10.1371/journal.pone.0252830

Menczer, F., & Hills, T. (2020). The attention economy. *Scientific American*, *323*(6), 54–61. https://doi.org/doi:10.1038/scientificamerican1220-54

Nassetta, J., & Gross, K. (2020). State media warning labels can counteract the effects of foreign disinformation. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-45

Nieminen, S., & Rapeli, L. (2019). Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political Studies Review*, *17*(3), 296–309. https://doi.org/10.1177/1478929918786852

Nisbet, E. C., Cooper, K. E., & Ellithorpe, M. (2015). Ignorance or bias? Evaluating the ideological and informational drivers of communication gaps about climate change. *Public Understanding of Science*, *24*(3), 285–301. https://doi.org/10.1177/0963662514545909

Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-024

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., & Rand, D. G. (2017). Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3035384

Pingree, R. J., Brossard, D., & McLeod, D. M. (2014). Effects of journalistic adjudication on factual beliefs, news evaluations, information seeking, and epistemic political efficacy. *Mass Communication and Society*, *17*(5), 615–638. https://doi.org/10.1080/15205436.2013.821491

Pingree, R. J., Watson, B., Sui, M., Searles, K., Kalmoe, N. P., Darr, J. P., Santia, M., & Bryanov, K. (2018). Checking facts and fighting back: Why journalists should defend their profession. *PLOS ONE*, *13*(12), e0208600. https://doi.org/10.1371/journal.pone.0208600

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), e2104235118. https://doi.org/10.1073/pnas.2104235118

Primig, F. (2022). The influence of media trust and normative role expectations on the credibility of fact checkers. *Journalism Practice*, 1–21. https://doi.org/10.1080/17512786.2022.2080102

Schwarz, N., & Newman. (2017, August). *How does the gut know truth?* Https://Www.Apa.Org. https://www.apa.org/science/about/psa/2017/08/gut-truth

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy*, *2*(1), 85–95. https://doi.org/10.1353/bsp.2016.0009

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. In *Advances in Experimental Social Psychology* (Vol. 39, pp. 127–161). Elsevier. https://doi.org/10.1016/S0065-2601(06)39003-X

Shearer, E., & Gottfried, J. (2017, September 7). News use across social media platforms 2017. *Pew Research Center's Journalism Project*. https://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media: Sharing fact-checking messages on social media. *Journal of Communication*, *67*(2), 233–255. https://doi.org/10.1111/jcom.12284

Stewart, E. (2021). Detecting fake news: Two problems for content moderation. *Philosophy & Technology*, *34*(4), 923–940. https://doi.org/10.1007/s13347-021-00442-x

Tandoc Jr., E. C., Duffy, A., Jones-Jang, S. M., & Wen Pin, W. G. (2021). Poisoning the information well?: The impact of fake news on news media credibility. *Journal of Language and Politics*, *20*(5), 783–802. https://doi.org/10.1075/jlp.21029.tan

Tanzer, M., Campbell, C., Saunders, R., Luyten, P., Booker, T., & Fonagy, P. (2021). *Acquiring knowledge: Epistemic trust in the age of fake news* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/g2b6k

TikTok. (2023). Beyond brand safety: Building for a new era of safety, privacy and security. *Brand Safety*. https://www.tiktok.com/business/en-US/blog/beyond-brand-safety-tiktok?redirected=1

TTC Labs. (n.d.). About trust, transparency, and control labs (TTC Labs). *TTC Labs*. https://www.ttclabs.net/about

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008. https://doi.org/10.1002/gch2.201600008

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wang, Y., Min, Q., & Han, S. (2016). Understanding the effects of trust and risk on individual behavior toward social media platforms: A meta-analysis of the empirical evidence. *Computers in Human Behavior*, *56*, 34–44. https://doi.org/10.1016/j.chb.2015.11.011

Warner-Søderholm, G., Bertsch, A., Sawe, E., Lee, D., Wolfe, T., Meyer, J., Engel, J., & Fatilua, U. N. (2018). Who trusts social media? *Computers in Human Behavior*, *81*, 303–315. https://doi.org/10.1016/j.chb.2017.12.026

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion*, *18*(2), 129–166. https://doi.org/10.1007/BF02249397

Young, D. G., Jamieson, K. H., Poulsen, S., & Goldring, A. (2018). Fact-checking effectiveness as a function of format and tone: Evaluating factcheck.org and flackcheck.org. *Journalism & Mass Communication Quarterly*, *95*(1), 49–75. https://doi.org/10.1177/1077699017710453

Zhang, J., Featherstone, J. D., Calabrese, C., & Wojcieszak, M. (2021). Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. *Preventive Medicine*, *145*, 106408. https://doi.org/10.1016/j.ypmed.2020.106408