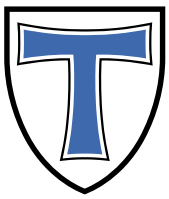
JUSTUS LIEBIG UNIVERSITY GIESSEN



DISSERTATION

DATA-DRIVEN PERSPECTIVES ON SOCIETAL ISSUES: ADDRESSING CHALLENGES IN THE DIGITAL AGE

Submitted in fulfilment of the requirements for the degree of

DOCTOR RERUM POLITICARUM (Dr. rer. pol.)

at the

Justus Liebig University Giessen
Department of Business Studies and Economics
Chair for Data Science & Digitization

by

Chiara Patricia Drolsbach
July 14, 2025

Justus-Liebig-Universität Gießen
Fachbereich Wirtschaftswissenschaften
Professur für Data Science & Digitalisierung
Licher Straße 62
35394 Gießen

Dekanin:	Prof. Dr. Corinna Ewelt-Knauer
Erstgutacher:	Prof. Dr. Nicolas Pröllochs
Zweitgutachter:	Prof. Dr. Peter Winker

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Dr. Nicolas Pröllochs, who gave me the opportunity to pursue my PhD and supported me with continuous feedback and guidance. I am very grateful to have had the chance to work and learn in such a stimulating research environment. Furthermore, I am also thankful to my co-supervisor, Prof. Dr. Peter Winker, for his support and helpful feedback over the past years.

I would also like to thank Prof. Dr. Georg Götz, at whose chair my academic journey began. It was there that I first gained insights into the field of research and was able to gather valuable experience already during my studies. Moreover, the strong sense of community at VWL I left a lasting impression on my time at Justus Liebig Universität Giessen.

My research journey was significantly shaped by collaboration with my colleagues and coauthors, including Max, Phil, and Kirill, who made major contributions to parts of my dissertation, as well as my colleagues from BWL XI: Michelle, Moritz, and Emma. I truly enjoyed collaborating with you in research and teaching. Not to forget all the other colleagues across the Department of Economics, with whom I had the privilege of spending many lunch and coffee breaks, having endless discussions, and some of whom have since become close friends. Without you, I would not be where I am today.

Last but not least, I would like to thank my family and friends, who supported me through difficult and demanding times and always lifted me up. Your unwavering belief in me gave me strength when I needed it most, and your encouragement reminded me of what truly matters beyond academic success.

Contents

Chapter 1

Introduction

1.1 Motivation

In the digital age, the ways in which people communicate, exchange information, conduct business, and generate knowledge have undergone a profound transformation. Digital technologies now influence nearly every aspect of society, reshaping social, political, and economic structures (Helbing, 2018; Jabłoński et al., 2020; Morrar et al., 2017). This transformation is inherently dual in nature: it enables participation, innovation, and efficiency, while also fostering market and power concentration, accelerating the rapid spread of misinformation, and granting a few platforms disproportionate control over information flows through opaque algorithms (Lazer et al., 2018; Tirole, 2023). These opposing dynamics have far-reaching implications, as digital systems not only amplify existing risks but also offer unprecedented tools for understanding and addressing them.

The *Global Risk Report 2024*, published by the World Economic Forum (2024), identifies mis- and disinformation, inflationary pressures, economic downturn, and societal polarization as dominant short- and long-term global risks. These risks reinforce one another and reflect a rapidly evolving and interconnected global landscape. The report further highlights additional concerns such as the cost-of-living crisis, calling on governments to support citizens with targeted and timely interventions. The 2025 report emphasizes the increasing role of AI-generated misinformation and the intensifying geopolitical and geoeconomic tensions (WEF, 2025). Together, these insights highlight the urgency of equipping policymakers with tools that can anticipate emerging risks and support timely, evidence-based interventions in a world where crises are no longer isolated, but tightly interwoven phenomena with global consequences.

To address the challanges, researchers increasingly rely on digital data sources that provide behavior-based insights into social dynamics, captured through everyday online and offline interactions (Blazquez & Domenech, 2018; Einav & Levin, 2014; Varian, 2014). The emergence of Big Data has introduced vast and complex datasets that capture high-frequency information. The widely recognized "5 Vs" framework (Volume, Velocity, Variety, Veracity, and Value) highlights the unprecedented scale, speed, heterogeneity, and complexity of these data, as well as their potential to generate economic and societal value (Bello-Orgaz et al., 2016). Recent estimates suggest that over 402.74 million terabytes of data are created, captured, copied, and consumed daily (Statista, 2025), with roughly 85% being unstructured content such as text, images, audio, and video content (Eberendu et al., 2016). In research, data collection has increasingly shifted toward author-collected data and the use of laboratory and field experiments, facilitating more targeted and credible empirical analyses (Hamermesh,

2013; Steelman et al., 2014). These developments underscore how Big Data is reshaping the empirical toolkit of researchers, offering new avenues to understand and respond to the societal impacts of rapid digital transformation.

With the rapid expansion in the volume and diversity of new data sources, novel analytical approaches are essential to harness their full potential. In this context, computational techniques such as machine learning (Jordan & Mitchell, 2015), natural language processing (Feuerriegel et al., 2025), and econometrics (Einav & Levin, 2014) are increasingly developed and applied to extract insights from complex, real-world data. This marks a shift from theory- or survey-based research toward large-scale behavioral analyses. In the past five decades, empirical methods have become central to economic research, alongside the rise of Computational Social Science, which uses computational techniques to analyze social behavior (Bao et al., 2025; Lazer et al., 2009; Prieto Gutiérrez et al., 2023). Crucially, these methodological innovations allow researchers to go beyond what individuals say they do and instead analyze actual behavior, providing a more robust foundation for evidence-based research.

This shift has sparked a wide array of interdisciplinary research. As global challenges grow more complex, so does the need for interdisciplinary approaches that integrate domain expertise with computational precision. Researchers increasingly work with large-scale datasets, from social media data and transaction records to election results and survey data, to better understand contemporary societal issues. These data sources have enabled empirical studies on topics such as political polarization, misinformation spread, public health behavior, and the evaluation of policy interventions (e. g., Bär et al., 2025; Bavel et al., 2020; Bonaccorsi et al., 2020; Briel et al., 2022; Geissler et al., 2022; Grebe et al., 2024; Robertson et al., 2023). These studies employ an increasingly diverse methodological toolkit combining econometric identification strategies with techniques from data science and predictive analytics. The result is a rapidly evolving research landscape that blends behavioral insight with computational rigor, enabling novel contributions to longstanding social and economic questions.

This dissertation examines three interrelated societal challenges emerging from the digital transformation and globalization: the spread of misinformation, the regulation of digital platforms, and the effectiveness of economic policy responses to global shocks. Each reflects a key pressure point where digital technologies intersect with societal outcomes:

- **Misinformation on Social Media**: The rise of digital platforms facilitates the rapid spread of misinformation by enabling low-cost, high-speed content dissemination across vast audiences, often without editorial oversight (Ecker et al., 2022; Shao et al., 2016; Vosoughi et al., 2018). Viral misinformation can shape opinions and harm society, especially during elections and crises (e. g., Allcott & Gentzkow, 2017; Bakshy et al., 2015; Bär et al., 2023; Del Vicario et al., 2016; Geissler et al., 2022). Scholars have highlighted the role of social media virality, algorithmic amplification, and filter bubbles, in exacerbating these trends (Allcott & Gentzkow, 2017; Lazer et al., 2018). From Chapter 2 to Chapter 5, this dissertation contributes empirical insights grounded in behavioral data and platform analytics to enhance understanding of misinformation diffusion mechanisms and the effectiveness of countermeasures such as fact-checking.
- **Regulation of Digital Platforms**: As digital platforms play an increasingly central role in shaping markets, public discourse, and access to information, governments are responding with new regulatory frameworks to address concerns over market dominance, content governance, and transparency. In the European Union, the Digital Services

Act (DSA) and Digital Markets Act (DMA) represent landmark efforts to establish clearer responsibilities for online intermediaries and gatekeepers. The DMA targets anti-competitive behavior; the DSA mandates content moderation, algorithmic transparency, and user protection (Chiarella, 2023; Laux et al., 2021). In Chapter 6, we perform an empirical analysis of how platforms implement these new requirements, focusing on content moderation practices under the DSA.

- **Evaluation of Policy Interventions**: Global crises, such as the COVID-19 pandemic, the Russian invasion of Ukraine and other geopolitical conflicts have triggered inflationary pressures not seen in decades (Habib & Kayani, 2024; Maurya et al., 2023; Vieira & da Silva, 2024). Among the main contributors to these price surges are energy and fuel prices, which are highly sensitive to external shocks and have played a central role in recent inflation dynamics. In this context, Chapter 7 of this dissertation examines how fuel prices responded to the fuel discount introduced in various European countries in 2022 and whether fuel stations passed on the reduced tax to consumers.

To address these issues, this dissertation argues that Big Data, when paired with appropriate analytical and computational methods, is essential for understanding and mitigating societal and economic risks in the 21st century. It combines diverse data sources with robust empirical strategies, drawing on publicly available datasets, web-scraped information, and original survey data to analyze real-world behavior in the digital age. The applied methods are tailored to each research question and include negative binomial regression, staggered difference-in-differences, and hierarchical modeling to account for nested structures. Additionally, natural language processing (NLP) techniques are used to extract topics, sentiment, and semantic patterns from unstructured text and media. Through this interdisciplinary, data-driven approach, the dissertation seeks to deepen understanding of real-world challenges and critically evaluate the effectiveness of relevant interventions and policy measures.

1.2 Research Framework and Thesis Structure

In this doctoral dissertation, I present a collection of six research papers, divided into three key topics: Misinformation on Social Media, Regulation of Digital Platforms, and Evaluation of Policy Interventions. Table 1.1 provides an overview of the six papers, co-authors, their publication status, and individual contributions. The aim of all papers is to provide an understanding of different societal and economic risks. To address the research questions, appropriate econometric and computational methods are employed. This section offers a brief overview of each topic and paper and outlines the structure of the subsequent chapters.

1.2.1 Misinformation on Social Media

In recent years, social media has become a primary source of news and information for billions of people worldwide (Lazer et al., 2009; Pew Research Center, 2016). Platforms like Facebook, X, Instagram, and TikTok have transformed how people consume information, interact with news, and engage in public discourse (Levy, 2021; Shore et al., 2018). In 2024, 54% of U.S. adults and 45% of German adults reported that they consume news on social media, a slight increase compared to previous years (Behre et al., 2024; Pew Research Center, 2024).

As social media becomes increasingly central to how people consume information, the volume and spread of misinformation have grown significantly. Unlike traditional news outlets that

Table 1.1: Overview of individual contributions to co-authored papers.

Paper	**Diffusion of Community Fact-Checked Misinformation on Twitter**
Co-authors	Nicolas Pröllochs
Status	Published in *Proceedings of the ACM on Human-Computer Interaction, Volume 7, Issue CSCW2*, https://doi.org/10.1145/3610058
Contribution	75%
Paper	**Believability and Harmfulness Shape the Virality of Misleading Social Media Posts**
Co-authors	Nicolas Pröllochs
Status	Published in *Proceedings of the ACM Web Conference 2023 (WWW'23)*, https://doi.org/10.1145/3543507.3583857
Contribution	75%
Paper	**Community Notes Increase Trust in Fact-Checking on Social Media**
Co-authors	Kirill Solovev & Nicolas Pröllochs
Status	Published in *PNAS Nexus, Volume 3, Issue 7, July 2024, pgae217*, https://doi.org/10.1093/pnasnexus/pgae217
Contribution	40%
Paper	**Characterizing AI-Generated Misinformation on Social Media**
Co-authors	Nicolas Pröllochs
Status	*Working Paper* (submitted to *ICWSM*), https://doi.org/10.48550/arXiv.2505.10266
Contribution	75%
Paper	**Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database**
Co-authors	Nicolas Pröllochs
Status	Published in *Companion Proceedings of the ACM Web Conference 2024 (WWW'24 Companion)*, https://doi.org/10.1145/3589335.3651482
Contribution	75%
Paper	**Pass-Through of Temporary Fuel Tax Reductions: Evidence from Europe**
Co-authors	Phil-Adrian Klotz & Maximilian Maurice Gail
Status	Published in *Energy Policy, Volume 183, December 2023*, https://doi.org/10.1016/j.enpol.2023.113833
Contribution	33.33%

rely on rigorous editorial oversight, social media platforms largely lack such verification mechanisms, enabling the rapid dissemination of both accurate and misleading content (Lazer et al., 2018; Shore et al., 2018). This has raised serious concerns about the credibility of online information and users' ability to distinguish trustworthy sources from unverified claims. Recent studies reflect this urgency, with academic publications on social media misinformation rising from about 20 annually between 2014 and 2017 to roughly 160 per year between 2020 and 2022 (Pérez-Escolar et al., 2023). The architecture of social media platforms encourages habitual behavior that undermines critical evaluation of content (Ceylan et al., 2023). Alarmingly, just 15% of the most habitual sharers are responsible for around 40% of false news circulation (Ceylan et al., 2023), and during the 2020 U.S. presidential election, a small group of "supersharers" accounted for approximately 80% of the fake news spread (Baribi-Bartov et al., 2024).

Misinformation spreads differently than true content, often traveling faster and reaching

broader audiences (Del Vicario et al., 2016; Friggeri et al., 2014; Vosoughi et al., 2018). Content that contains emotional or sensational language, particularly when it evokes anger or moral condemnation, tends to go viral by triggering social cues such as high engagement metrics, which further incentivize user interaction and entrench misinformation in a feedback loop (Chuai & Zhao, 2022; Epstein et al., 2023; Pröllochs et al., 2021a, 2021b; Solovev & Pröllochs, 2022). This cycle is intensified by partisanship and confirmation bias, which lead users to preferentially engage with content from co-partisans that align with their preexisting beliefs, reinforcing echo chambers and increasing resistance to correction (Altay et al., 2023; Barberá et al., 2015; Ceylan et al., 2023; Moravec et al., 2019, 2020; Mosleh & Rand, 2022). In addition, users with lower cognitive ability are less capable of discerning misinformation (Pennycook & Rand, 2019a, 2019b; Roozenbeek et al., 2020), and are more likely to believe repeated falsehoods due to the familiarity effect (Ecker et al., 2022; Pennycook et al., 2018).

In response, social media platforms have increasingly explored proactive and reactive measures to reduce misinformation exposure and improve content discernment. Central to these efforts are countermeasures that help users to navigate misleading content more effectively. Common interventions include media literacy training (e. g., Brashier et al., 2021; Bruns et al., 2024; Guess et al., 2020; Wang et al., 2025), nudges (e. g., Pennycook et al., 2021), and fact-checking (e. g., Berger et al., 2025; DeVerna et al., 2024; Walter et al., 2020). Research on the effectiveness of these interventions does not clearly favor one over the others; instead, they are increasingly seen as complementary tools that support users in identifying and resisting misinformation (Bak-Coleman et al., 2022; Kozyreva et al., 2024).

Among these interventions, fact-checking, typically conducted by experts and implemented through visible warning labels or professional assessments, has emerged as one of the most prominent and extensively studied approaches. It has consistently been shown to reduce belief in and engagement with false content (Clayton et al., 2020; Martel et al., 2024a; Pennycook et al., 2020). However, expert fact-checking faces multiple challenges such as scalability due to the limited number of human experts (Martel et al., 2024b; Pennycook & Rand, 2019a), and distrust, particularly among Republican users (Pennycook & Rand, 2019a; Resnick et al., 2023). Surveys indicate that 70% of Republicans and half of all U.S. adults believe that fact-checkers are biased (Poynter, 2019). Therefore, even if a social media post is flagged, its influence on users may remain minimal because of insufficient trust in the source (Brandtzaeg & Følstad, 2017). Automated systems using NLP and machine learning can scale efficiently but struggle with accuracy and data limitations (Ducci et al., 2020; Godel et al., 2021; Wu et al., 2019). A promising alternative is community-based fact-checking, leveraging the "wisdom of crowds". Studies show that politically balanced, non-expert ratings, even from relatively small groups, correlate strongly with expert fact-checks (Allen et al., 2021; Pennycook & Rand, 2019a) and decentralized networks enhance their effectiveness (Frey & van de Rijt, 2021).

Building on the promise of crowdsourced fact-checking as a scalable and politically neutral alternative, X (formerly Twitter) has implemented one of the most prominent examples: Community Notes (X, 2021). Formerly launched as Birdwatch in January 2021, Community Notes allows users to collaboratively identify and annotate misleading content directly on the platform. Registered contributors can submit notes (i. e., brief textual explanations) on posts and identify them as misleading or not misleading. The notes are then rated by other contributors as "helpful" or "not helpful" (Pröllochs, 2022). For a note to be publicly displayed, it must receive at least five helpful ratings from contributors with diverse political and ideological perspectives, a mechanism designed to minimize partisan bias and promote

factual integrity (Wojcik et al., 2022; X, 2024). Importantly, the data and all underlying algorithms are publicly available, ensuring transparency and enabling independent analysis. As of June 2025 more than one million contributors have written almost two million notes for more than 1.2 million posts, making it the most extensive real-world deployment of crowd-sourced fact-checking. The initiative is being continuously developed to improve its quality, visibility and reach (X, 2025).

From a research perspective, the implementation of Community Notes on X presents substantial opportunities for exploring a variety of scientific questions. In recent years, a wide range of studies have demonstrated that Community Notes can be an effective tool to combat misinformation under certain conditions. Pröllochs (2022) showed early on that notes tend to be written for posts considered misleading and are disproportionately directed at content with high engagement or authored by socially influential users. However, these notes often elicit lower consensus and are more likely to be perceived as argumentative (Pilarski et al., 2024). Bobek and Pröllochs, 2025 show that being fact-checked by Community Notes does not effect the authors follower base significantly. Notably, Allen et al. (2021) and Wojcik et al. (2022) provide evidence that contributor partisanship influences both note-writing and note-rating behavior, with users more likely to rate notes from counterpartisans as unhelpful. More recent research shows that the most important factors of helpfulness are whether notes are supported by external sources, and whether these sources are politically biased or not, with politically neutral sources increasing helpfulness ratios the most (Solovev & Pröllochs, 2025). At the linguistic level, Pröllochs (2022) and Phillips et al. (2024) find that notes written in a neutral tone receive the highest helpfulness scores, while those with emotionally charged or highly negative language are rated as less helpful. Addressing the trust issues that expert fact-checkers face, we were able to show that Community Notes are perceived as more trustworthy than traditional misinformation flags and enhance users' ability to detect misleading content across the political spectrum (see Chapter 4).

The impact of Community Notes on user engagement has also been a key focus of research. While we show that misleading posts with notes are less viral than non-misleading posts (see Chapter 2) and that easily believable and not particularly harmful noted posts are significantly more viral (see Chapter 3), Chuai, Tian, et al. (2024) argue that the system remains too slow to counteract the initial virality of such content. However, Chuai, Pilarski, et al. (2024) show that over time, notes can reduce reposts by approximately 61.4%. Further analyses indicate that Community Notes are especially effective in reducing engagement with posts by less influential users, those with neutral or mildly negative sentiment, and those discussing health-related topics (Chuai, Pilarski, et al., 2024). Furthermore, recent evidence suggests that the display of community fact-checks significantly increases expressions of negativity, anger, disgust, and moral outrage in replies, indicating that users view misinformation as a violation of social norms and react emotionally when falsehood is exposed (Chuai et al., 2025). In Chapter 5, we extend this line of research by using Community Notes to detect and characterize AI-generated misinformation, revealing systematic differences in content, authorship, and virality compared to human-generated misleading posts.

Taken together, these findings suggest that while Community Notes may not fully prevent the initial spread of viral misinformation, they hold considerable potential to reduce long-term engagement with misleading content. As such, Community Notes represent a promising and scalable strategy for mitigating misinformation online; one that warrants continued empirical investigation, particularly regarding which types of notes and source posts drive the strongest

effects. Based on the strong scientific support of the concept, other social media platforms have begun piloting similar initiatives in the U.S., namely "Footnotes" on TikTok (TikTok, 2025), "Community Notes" on Facebook and Instagram (Meta, 2025), and "Notes" on YouTube (YouTube, 2024).

As outlined in the preceding sections, papers I have co-authored contribute to the expanding body of research on misinformation and community-based fact-checking, which has received growing academic attention in recent years. In the following, I provide a more detailed overview of the individual papers. Each contribution is revisited to outline its specific research focus, methodology, and key findings, thereby highlighting its relevance and implications within the broader discourse on combating misinformation online.

Chapter 2: Diffusion of Community Fact-Checked Misinformation on Twitter

In Chapter 2, we examine the diffusion of community fact-checked misinformation that has been identified via Community Notes (called Birdwatch at the time of the study) during the projects' pilot phase. During this phase, a limited number of contributors in the US were able to annotate posts as misleading or not misleading, add a textual explanation that provides context to the tweet, and answer a set of checkbox questions characterizing the tweet (e. g., types of misinformation). We aim to analyze differences in the diffusion of posts identified as misleading and not misleading. To this end, we draw upon a large dataset of more than 15,000 posts that have been annotated by Community Notes contributors between January 2021 and February 2022.

This research paper has been published in the following conference proceedings:

> **Drolsbach, C.P.** & Nicolas Pröllochs (2023). Diffusion of Community Fact-Checked Misinformation on Twitter. *Proceedings of the ACM on Human-Computer Interaction, Volume 7, Issue CSCW2*

We quantify the virality (i. e., Retweet Count) of misleading vs. not misleading posts using an explanatory regression model, while controlling for the time elapsed between the publication of the post and the fact-check, the sentiment of the post, and the social influence of its author (i. e., follower count, followee count, account age, verification status). In contrast to previous research, which focused primarily on misinformation that has been fact-checked by experts, our results show that misleading posts are expected to receive 35.85% fewer retweets than not misleading posts. A comparison of characteristics of our dataset with the rumors analyzed by Vosoughi et al. (2018) reveals that differences are partially explainable by the fact that different types of fact-checkers focus on different targets. Community Notes contributors tend to fact-check posts from larger accounts with greater social influence, while expert fact-checkers tend to focus on rumors that are shared by smaller accounts. Further, we find significant differences between different types of misinformation: misleading posts categorized as "Manipulated Media" or "Satire" receive more retweets than not misleading posts, while those labeled "Factual Error", "Missing Important Context", "Unverified Claim as Fact" and "Other" receive fewer retweets. Contributing to the discourse on whether social media users tend to exploit a tool like Community Notes, we show through a user study that the majority of fact-checks is perceived as reliable, while only a relatively small share is seen as purposely deceptive.

Our results underline the potential of community-based fact-checking as a scalable and trustworthy tool to combat misinformation on social media, particularly by targeting content that is relevant to social media users but may be overlooked by expert fact-checkers. This highlights

that expert- and community-based approaches could complement each other well, given their differing selection criteria and focus areas. However, our findings also suggest that differences in sample selection play a key role in explaining divergent results across studies and must be carefully considered in future research. Notably, our study is limited by its observational design, its focus on X's Community Notes pilot (which was not visible to most users), and the potential non-representativeness of the contributor base, all of which constrain the generalizability of our findings and point to the need for future controlled experiments and field validations.

Chapter 3: Believability and Harmfulness Shape the Virality of Misleading Social Media Posts

In Chapter 3, we dive deeper into the underlying characteristics of misinformation that influence their virality, namely their perceived harmfulness and believability. We utilize the same dataset as in Chapter 3, but focus on posts identified as being misleading. This research paper has been published in the following conference proceedings:

> **Drolsbach, C.P.** & Nicolas Pröllochs (2023). Believability and Harmfulness Shape the Virality of Misleading Social Media Posts. *Proceedings of the ACM Web Conference (WWW'23)*

A unique feature of X's Community Notes initiative is that, besides identifying misleading posts and writing a textual note, contributors can also rate the perceived believability and harmfulness of the posts (this feature was available until October 2023). Building on this, our study not only analyzes the diffusion of crowd fact-checked misinformation, but also examines how underlying characteristics of misinformation are associated with its virality.

To this end, we specify an explanatory regression model that explains the virality of misleading posts based on their preceived believability and harmfulness, while controlling for variables of social influence. Our results show that posts that are perceived as *believable* receive 217.09% more retweets, while posts that are perceived as *harmful* receive 41.32% fewer retweets. In order to assess the accuracy of the contributor-assigned characteristics, we conducted a user study with seven U.S.-based, English-speaking participants who rated the believability and harmfulness of 150 misleading tweets. The results show that posts labeled as believable or harmful via Community Notes were also perceived as significantly more believable and harmful by independent annotators, supporting the validity of these classifications.

Altogether, our results indicate that post characteristics play a key role in the virality of misinformation. Specifically, misleading posts that are easily believable and not particularly harmful are significantly more viral. From a theoretical perspective, this may be explained by the hedonic mindset of social media users, which describes that if users do not believe the content or perceive it as harmful, sharing it may be less enjoyable. This insight has practical implications for the design of fact-checking and moderation strategies, as platforms may prioritize expert review of content that is both believable and potentially harmful. However, our findings are observational and limited to the Community Notes pilot phase, and future work should explore whether these patterns generalize across platforms, populations, and more mature community fact-checking environments.

Chapter 4: Community notes increase trust in fact-checking on social media

In Chapter 4, we tackle another challenge in fact-checking: trust. Surveys show that more than 50% of Americans distrust expert fact-checkers and accuse them of biased ratings (Poynter, 2019), although research shows that fact-checkers do not disproportionately fact-check politicians from one political side (Greene et al., 2025). Focusing on crowd-based fact-checking, we aim to analyze whether this mistrust can be resolved. This research paper has been published in the following journal article:

> **Drolsbach, C.P.** & Nicolas Pröllochs (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus, Volume 3, Issue 7, pgae217*

We conducted a survey with more than 1 800 U.S. participants who were presented with different types of fact-checks and asked to what extent they trusted these fact-checks. As extensions, we also analyzed participants' ability to discern misleading and not misleading posts, as well as their sharing behavior. By fitting a hierarchical logistic regression model to predict whether a fact-check was rated as trustworthy, we found that Community Notes were perceived as significantly more trustworthy than simple misinformation flags across both sides of the political spectrum. Our findings demonstrate that context plays a critical role in the effectiveness of fact-checking interventions. Participants consistently rated text-based community notes, which explain why a post is misleading (i. e., explain the context of a post), as significantly more trustworthy than simple, context-free misinformation flags. Therefore, this trust boost was not due to the source of the fact-check (community vs. experts), but rather the presence of explanatory context. We also observed that this effect is not uniform across political groups. Trust in fact-checks, and the impact of community notes, varied depending on participants' political alignment and the ideological congruence of the misinformation. Specifically, community notes improved the discernment of politically concordant misinformation, but had less impact on politically discordant content, which users likely already perceived as clearly false.

In addition, our results reinforce the notion that partisanship shapes users' reactions to misinformation interventions. Participants who preferred Trump over Biden were significantly less likely to trust any form of fact-checking and showed lower accuracy in distinguishing misleading from accurate posts. They were also more likely to re-share misinformation. These findings are consistent with prior research showing that right-leaning users are more skeptical of fact-checks and tend to place more trust in content aligned with their preexisting beliefs. While community notes did enhance overall post discernment, they did not consistently reduce users' intention to share misleading content compared to basic misinformation flags. This highlights the persistent gap between recognizing misinformation and acting upon that knowledge. Altogether, our results underscore the importance of political identity and cognitive bias in shaping the effectiveness of fact-checking strategies and suggest that context-rich, transparent interventions like community notes offer a promising, though not complete, solution to the misinformation challenge.

While our study offers valuable insights, it also has some limitations. The fact-checks were applied only to verifiably misleading posts, leaving out false positives and untagged misinformation that frequently appear in real-world scenarios. Additionally, since our sample consisted solely of U.S. participants, the results may not extend to different cultural or political contexts. Moreover, the controlled experimental setting cannot fully replicate the complex environment

of social media, where factors like peer influence and platform algorithms heavily shape user behavior.

Chapter 5: Characterizing AI-Generated Misinformation on Social Media

Artificial intelligence (AI) technologies, especially generative tools like deepfakes, are transforming the social media landscape by enabling the creation of highly realistic synthetic content (Feuerriegel et al., 2023; Westerlund, 2019). These advancements blur the line between real and fabricated media, complicating efforts to detect misinformation and posing new challenges for platform moderation. Despite growing concern over the societal risks of AI-generated misinformation, including the erosion of trust in media and democratic institutions, its actual spread and characteristics on social media remain poorly understood.

In Chapter 5, we address that gap by conducting a large-scale empirical analysis of AI-generated misinformation on X. Using a dataset of num91452 posts that have been identified as misleading via Community Notes between January 2023 and January 2025, we examine how AI-generated misinformation differs from traditional misinformation in terms of content and author characteristics, virality, and perceived harmfulness and believability. This research paper is currently a working paper and has been submitted to be published at ICWSM 2026:

> **Drolsbach, C.P.** & Nicolas Pröllochs (2025). Characterizing AI-Generated Misinformation on Social Media. *Working Paper. Submitted to ICWSM*

Prior studies show that the increased scalability, multilingualism, and multimodality of AI-generated content complicate its detection and mitigation (Feuerriegel et al., 2023; Timmerman et al., 2023). Human-centered methods that focus on individual classification often struggle to accurately identify AI-generated misinformation (Bray et al., 2023; Cooke et al., 2024; Diel et al., 2024; Groh et al., 2024; Somoray & Miller, 2023) and lack scalability (Groh et al., 2024). Machine learning-based methods offer a scalable solution, but so far struggle with accuracy (Almars, 2021; Feuerriegel et al., 2023; Zhou et al., 2023).

In contrast to prior approaches that attempt to detect AI-generated content by analyzing the media itself, we propose an LLM-based identification method that leverages human-written fact-checks. Specifically, we build an OpenAI Assistant that interprets the textual explanation attached to Community Notes to determine whether it refers to AI-generated content. This allows us to identify 4 577 (i. e., 5.06%) misleading posts that contain AI-generated content. We also use an LLM-based approach to annotate a subset of posts with a wide variety of content characteristics, such as the sentiment, topic, harmfulness and believability.

Our analysis of AI-generated versus non-AI-generated misleading posts reveals notable differences in content characteristics, author profiles, and engagement dynamics. AI-generated posts are significantly more likely to include visual media (images or videos) and are often more positively toned, with a higher share of positive sentiment and a focus on entertainment-related topics. This contrasts with non-AI-generated posts, which more frequently cover serious issues like health and exhibit a more negative sentiment. AI-generated content also tends to be shared by users with larger follower bases, older accounts, and fewer followees, suggesting these posts originate from more influential or established accounts. Furthermore, accounts sharing AI-generated misinformation have slightly more conservative partisanship scores, although both groups lean conservative overall.

In terms of virality, AI-generated misinformation significantly outperforms non-AI content, receiving more retweets, likes, and impressions even after controlling for media type, account characteristics, and content topics (all $p < 0.01$). Regression results show that media presence strongly boosts engagement, and verified accounts enjoy higher visibility but not necessarily more user interaction. Interestingly, while AI-generated posts are more engaging, they are slightly less likely to be seen as harmful or believable compared to non-AI-generated ones. While the effects reach statistical significance, their magnitude is small. AI-generated posts are more viral and visually engaging, but users rate their believability and potential harm as similar to, or marginally lower than, those of human-produced misinformation.

Our results implicate that AI-generated content should be treated as a distinct dimension in misinformation research, prompting the need for updated models of detection, spread, and engagement that go beyond traditional account-based or content-based heuristics. This has implications for platform moderation strategies, which must adapt to the unique virality and persuasive qualities of AI-generated misinformation, potentially through community-driven tools and cross-platform coordination. However, these findings are subject to several limitations: the study is observational and cannot infer causality, focuses solely on platform X, and relies on user-flagged misinformation, which may overlook less visible cases. Future research should address these gaps by expanding to other platforms, investigating causal pathways, and examining the effectiveness of fact-checking across different types of misinformation.

1.2.2 Regulation of Digital Platforms

Given the growing relevance and importance of online platforms and digital markets, the European Union (EU) introduced specific rules with the aim of creating "a safer and more open digital space".[1] While the Digital Markets Act (DMA) is concerned with economic imbalances, unfair business practices by "gatekeepers" and their negative consequences on competition, the Digital Services Act (DSA) focuses on the dissemination of illegal content on platforms. It addresses issues such as the liability of online intermediaries for third party content, safety of users or asymmetric due diligence obligations for different providers or information society services (Chiarella, 2023; Laux et al., 2021). Both regulations entered into force in November 2022, followed by varying applications dates for different markets and platforms.

Under the DSA, online platforms must take responsibility for user safety, illegal content, and transparency. Specifically, they must provide clear reporting mechanisms for illegal content, such as yearly *Transparency Reports* (Article 15, DSA) and inform users about content moderation decisions by submitting so-called *Statements of Reasons* (SoRs) for each moderation activity (Article 17, DSA). In line with Article 24(5) of the DSA, all submitted SoRs must be collected and published on the DSA Transparency Database (DSA-TDB). Platforms must also ensure transparency in advertising and how their algorithms work. Larger platforms (i. e., Very Large Online Platforms (VLOPs)) face stricter obligations, including risk assessments, external audits, and cooperation with EU authorities.

From the platform's perspective, content moderation involves a range of tools and processes designed to prevent the spread of illegal or harmful content within online communities (Grimmelmann, 2015; Roberts, 2020). To enforce community guidelines and reduce harmful

[1] https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package, accessed: 15.04.2025

behavior, social media providers use measures such as content removal, visibility reduction, labeling, and account suspensions (Jiang et al., 2023). While effective moderation is essential for maintaining safe and civil spaces (Gillespie, 2018), it can sometimes have unintended consequences, including increased rule-breaking or the migration of harmful content to other platforms (Ali et al., 2021; Chang & Danescu-Niculescu-Mizil, 2019; Mitts et al., 2022; Russo et al., 2023). Over the last years, moderation has evolved from small manual review teams to sophisticated systems combining automated detection with human moderators (Meta, 2023; TikTok, 2023; X, 2023; YouTube, 2023), both paid (Roberts, 2020) and voluntary (Matias, 2016). Despite these advancements, platforms generally keep the specifics of their moderation practices confidential (Gillespie, 2018; Jhaver et al., 2019).

Chapter 6: Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database

While the DSA clearly describes the obligations online platforms need to fulfill, it remains unclear how platforms act. In Chapter 6, we provide a holistic early look at content moderation decisions of social media platforms in the EU. By analyzing more than 156 million *Statements of Reason*, that were submitted in the first two months to the DSA-TDB, we aim to reveal differences in content moderation practices and how social media platforms implement their obligations under the DSA. Due to the unique dataset of the EU's DSA-TDB we are able to provide the first empirical analysis of real-world content moderation in the EU across platforms. This research paper has been published in the following conference proceedings:

> **Drolsbach, C.P.** & Nicolas Pröllochs (2024). Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database. *Companion Proceedings of the ACM Web Conference (WWW'24)*

Our analysis covers a broad set of factors to understand which content types are typically targeted (RQ1), the specific reasons driving moderation decisions (RQ2), the role of automation in these processes (RQ3), and the kinds of moderation actions platforms apply (RQ4).

Our results reveal significant variation in the moderation volume, disproportionate to the number of monthly active users in the EU. TikTok accounted for the majority of submissions (64%), followed by Facebook (22%) and Pinterest (8%), while other platforms such as X, Snapchat and LinkedIn each account for less than 0.5%. Further, most moderated content was categorized as text or video, though many platforms used a vague "Other" category, reflecting the limitations of a one-dimensional content classification system. Content was mostly flagged as *incompatible* rather than *illegal*, except on X, which focused on illegal content, especially violence and pornography. Moderation actions also varied: most platforms favored content removal or demotion, while Snapchat emphasized disabling content and X failed to clearly distinguish between demotion and removal.

We observe that moderation was largely automated: 60.67% of decisions were fully automated, and 99% of these involved automatically detected violations. TikTok mainly employs fully automated moderation, while Facebook, Pinterest and Instagram combine automated means and human intervention. The other platforms (i. e., YouTube, Snapchat, X, and LinkedIn), however, performed the majority of their interventions manually. A regression analysis shows that automatically detected content is vastly more likely to be automatically moderated, particularly for text. Regarding content moderation actions (RQ4), most platforms primarily removed content (55.15%), followed by demotion (25.15%), and account suspension or

termination (14.96%). Pinterest emphasized demotion, while Snapchat relied heavily on disabling content. X, notably, did not clearly differentiate between demotion and removal, often labeling content as "not suitable for work."

Overall, our findings reveal inconsistent content moderation practices across platforms, suggesting varied interpretations of DSA obligations. These differences in volume, focus, and automation raise questions about enforcement consistency and transparency. This highlights the need for clearer regulatory guidance and more standardized reporting to ensure uniform handling of rule-breaking content.

1.2.3 Evaluation of Policy Interventions

In addition to developments in the digital world, today's global landscape is increasingly shaped by major crises and conflicts, such as the COVID-19 pandemic and the Russian invasion of Ukraine. These events have contributed to rising consumer prices (Habib & Kayani, 2024; Maurya et al., 2023; Vieira & da Silva, 2024). Energy prices, in particular, are highly sensitive to global disruptions and have played a significant role in recent inflation dynamics (Kpodar & Abdallah, 2017). Their volatility stems from demand shocks and geopolitical uncertainty, often leading to immediate market reactions (Baumeister & Kilian, 2016; Kilian & Vega, 2011). Within the energy sector, fuel prices are especially reactive and impactful. They respond quickly to global conflicts and supply risks, with even anticipated disruptions causing sharp price swings (Baumeister & Kilian, 2016; Kpodar & Abdallah, 2017). This effect is reinforced by inelastic demand and limited short-term alternatives (Alberini et al., 2022; Frondel & Vance, 2010). As a result, fuel prices directly influence transportation costs, household budgets, and the broader price level. They also affect economic behavior, from travel patterns to interest in alternative energy. Thus, energy and fuel markets serve as key transmission channels through which global crises translate into everyday economic pressure.

Chapter 7: Pass-through of Temporary Fuel Tax Reductions: Evidence from Europe

Given these circumstances, Chapter 7 investigates whether government interventions can effectively relieve consumers from rising fuel costs. In early 2022, several European countries implemented temporary fuel tax reductions in response to increasing energy prices. However, given the oligopolistic structure of fuel retail markets, it remains uncertain to what extent these tax cuts were actually passed through to end consumers. This research paper has been published in the following journal article:

> **Drolsbach, C.P.**, Maximilian Maurice Gail & Phil-Adrian Klotz. (2023). Pass-through of Temporary Fuel Tax Reductions: Evidence from Europe. *Energy Policy, Volume 183*

To evaluate our research question, we employ a staggered Difference-in-Differences (DiD) design that exploits the variation in the timing and size of tax reductions across countries. The empirical strategy compares fuel price changes in treatment countries, namely France, Germany, and Italy, to a control group consisting of Austria, Estonia, Lithuania, and Latvia, which did not implement similar tax policies during the same period. This quasi-experimental design allows for causal inference under the parallel trends assumption, which is assessed through event study specifications and placebo tests. A key advantage of this method is its ability to isolate the effect of tax interventions from confounding market dynamics and to leverage policy heterogeneity across countries for robust identification.

The dataset includes daily fuel prices disagreggated by fuel type (gasoline vs. diesel), for 52 unique country-service-station-chain combinations over 241 days, spanning the period from January 3 to August 31, 2022, allowing for a granular analysis of pricing behavior. To calculate margins, we deduct fuel taxes and the crude oil cost component from the consumer prices. Additionally, we control for country-level indicators capturing characteristics of both the upstream and downstream fuel markets (i. e., global oil prices, refinery output, number of gas stations, total imports of oil and petroleum products) and include time-, country-, and chain-fixed effects.

The results suggest heterogeneous pass-through rates over time depending on country and fuel type. However, we conclude with two key findings: (1) The average pass-through rates are very high, which is why we can assume a full-shifting of the temporary tax reductions, and (2) the estimated average pass-through rates are higher for gasoline than for diesel. This may be due to specific conditions in the 2022 European energy market—particularly, high demand for diesel as a heating substitute and fuel for industrial use following the Russian invasion of Ukraine. Additionally, the timing and extent of tax pass-through to consumers differ across countries and fuel types, with full pass-through occurring at varying speeds and occasional periods of over-shifting. While the mostly one-to-one pass-throughs meant that average retail margins remained largely unchanged, petroleum companies initially earned some positive margins shortly after the tax reductions, before the full pass-through was completed.

Our results indicate that temporary fuel tax reductions seem to be an effective short-term measure to lower consumer prices. However, policy makers must carefully consider distributional concerns and climate-related consequences. For instance, the tax cuts may disproportionately benefit wealthier households with higher fuel consumption, and they may counteract long-term decarbonization efforts. In addition, the fiscal cost of such policies raises questions about their sustainability and efficiency relative to alternative interventions.

Bibliography

Alberini, A., Horvath, M., & Vance, C. (2022). Drive less, drive better, or both? behavioral adjustments to fuel price changes in germany. *Resource and Energy Economics*, *68*, 101292.

Ali, S., Saeed, M. H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Understanding the effect of deplatforming on social networks. *WebSci*.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393.

Almars, A. M. (2021). Deepfakes detection techniques using deep learning: A survey. *Journal of Computer and Communications*, *9*, 20–35.

Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *HKS Misinformation Review*, *4*(4).

Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, *6*(10), 1372–1380.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Bao, H., Zhang, J., Cao, M., & Evans, J. A. (2025). From division to unity: A large-scale study on the emergence of computational social science, 1990-2021. *Companion Proceedings of the ACM on Web Conference 2025*, 859–863.

Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). New threats to society from free-speech social media platforms. *Communications of the ACM*, *66*(10), 37–40.

Bär, D., Pröllochs, N., & Feuerriegel, S. (2025). The role of social media ads for election outcomes: Evidence from the 2021 german election. *PNAS nexus*, *4*(3), pgaf073.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542.

Baribi-Bartov, S., Swire-Thompson, B., & Grinberg, N. (2024). Supersharers of fake news on twitter. *Science*, *384*(6699), 979–982.

Baumeister, C., & Kilian, L. (2016). Forty years of oil price fluctuations: Why the price of oil may still surprise us. *Journal of Economic Perspectives*, *30*(1), 139–160.

Bavel, J. J. V., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., et al. (2020). Using social and behavioural science to support covid-19 pandemic response. *Nature human behaviour*, *4*(5), 460–471.

Behre, J., Hölig, S., & Möller, J. (2024). Reuters institute digital news report 2024: Ergebnisse für deutschland.

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*, 45–59.

Berger, L. M., Kerkhof, A., Mindl, F., & Münster, J. (2025). Debunking "fake news" on social media: Immediate and short-term effects of fact-checking and media literacy interventions. *Journal of Public Economics*, *245*, 105345.

Blazquez, D., & Domenech, J. (2018). Big data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*, 99–113.

Bobek, M., & Pröllochs, N. (2025). Community fact-checks do not break follower loyalty. *arXiv preprint arXiv:2505.10254*.

Bonaccorsi, G., Pierri, F., Cinelli, M., Flori, A., Galeazzi, A., Porcelli, F., Schmidt, A. L., Valensise, C. M., Scala, A., Quattrociocchi, W., et al. (2020). Economic and social consequences of human mobility restrictions under covid-19. *Proceedings of the national academy of sciences*, *117*(27), 15530–15535.

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

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5), e2020043118.

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, *9*(1), tyad011.

Briel, S., Osikominu, A., Pfeifer, G., Reutter, M., & Satlukal, S. (2022). Gender differences in wage expectations: The role of biased beliefs. *Empirical Economics*, 1–26.

Bruns, H., Dessart, F. J., Krawczyk, M., Lewandowsky, S., Pantazi, M., Pennycook, G., Schmid, P., & Smillie, L. (2024). Investigating the role of source and source trust in prebunks and debunks of misinformation in online experiments across four eu countries. *Scientific Reports*, *14*(1), 20723.

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), e2216614120.

Chang, J., & Danescu-Niculescu-Mizil, C. (2019). Trajectories of blocked community members: Redemption, recidivism and departure. *WWW*.

Chiarella, M. L. (2023). Digital markets act (dma) and digital services act (dsa): New rules for the eu digital environment. *Athens JL*, *9*, 33.

Chuai, Y., Pilarski, M., Renault, T., Restrepo-Amariles, D., Troussel-Clément, A., Lenzini, G., & Pröllochs, N. (2024). Community-based fact-checking reduces the spread of misleading posts on social media. *arXiv preprint arXiv:2409.08781*.

Chuai, Y., Sergeeva, A., Lenzini, G., & Pröllochs, N. (2025). Community fact-checks trigger moral outrage in replies to misleading posts on social media. *CHI*.

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the roll-out of community notes reduce engagement with misinformation on x/twitter? *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1–52.

Chuai, Y., & Zhao, J. (2022). Anger can make fake news viral online. *Frontiers in Physics*, *10*, 970174.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., et al. (2020). Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*, 1073–1095.

Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2024). As good as a coin toss: Human detection of ai-generated images, videos, audio, and audiovisual stimuli. *arXiv preprint arXiv:2403.16760*.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *PNAS*, *113*(3), 554–559.

DeVerna, M. R., Yan, H. Y., Yang, K.-C., & Menczer, F. (2024). Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, *121*(50), e2322823121.

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, *16*, 100538.

Ducci, F., Kraus, M., & Feuerriegel, S. (2020). Cascade-lstm: A tree-structured neural classifier for detecting misinformation cascades. *KDD*.

Eberendu, A. C., et al. (2016). Unstructured data: An overview of the data of big data. *International Journal of Computer Trends and Technology*, *38*(1), 46–50.

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

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*(6210), 1243089.

Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G., & Rand, D. (2023). The social media context interferes with truth discernment. *Science Advances*, *9*(9), eabo6169.

Feuerriegel, S., DiResta, R., Goldstein, J. A., Kumar, S., Lorenz-Spreen, P., Tomz, M., & Pröllochs, N. (2023). Research can help to tackle ai-generated disinformation. *Nature Human Behaviour*, *7*, 1818–1821.

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., et al. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, *4*, 96–111.

Frey, V., & van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, *67*(7), 4273–4286.

Friggeri, A., Adamic, L. A., Eckles, D., & Cheng, J. (2014). Rumor cascades. *ICWSM*.

Frondel, M., & Vance, C. (2010). Driving for fun? comparing the effect of fuel prices on weekday and weekend fuel consumption. *Energy Economics*, *32*(1), 102–109.

Geissler, D., Bär, D., Pröllochs, N., & Feuerriegel, S. (2022). Russian propaganda on social media during the 2022 invasion of ukraine. *arXiv:2211.04154*.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. (2021). Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, *1*(1), 1–36.

Grebe, M., Kandemir, S., & Tillmann, P. (2024). Uncertainty about the war in ukraine: Measurement and effects on the german economy. *Journal of Economic Behavior & Organization*, *217*, 493–506.

Greene, K. T., Pisharody, N., Carroll, F., & Shapiro, J. N. (2025). Fact-checks focus on famous politicians, not partisans. *PNAS Nexus*, *4*(1), pgae567.

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology*, *17*(1).

Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. (2024). Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications*, *15*(1), 7629.

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour*, *4*(5), 472–480.

Habib, A. M., & Kayani, U. N. (2024). Price reaction of global economic indicators: Evidence from the covid-19 pandemic and the russia–ukraine conflict. *SN Business & Economics*, *4*(1), 19.

Hamermesh, D. S. (2013). Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, *51*(1), 162–172.

Helbing, D. (2018). Societal, economic, ethical and legal challenges of the digital revolution: From big data to deep learning, artificial intelligence, and manipulative technologies. In *Towards digital enlightenment: Essays on the dark and light sides of the digital revolution* (pp. 47–72). Springer.

Jabłoński, A., Jabłoński, M., Jabłoński, A., & Jabłoński, M. (2020). The impact of the digital technology revolution on creating new markets and people's behavior. *Social Business Models in the Digital Economy: New Concepts and Contemporary Challenges*, 25–49.

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction*, *26*(5), 1–35.

Jiang, J. A., Nie, P., Brubaker, J. R., & Fiesler, C. (2023). A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction*, *30*(1).

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kilian, L., & Vega, C. (2011). Do energy prices respond to us macroeconomic news? a test of the hypothesis of predetermined energy prices. *Review of Economics and Statistics*, *93*(2), 660–671.

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J. B., Barzilai, S., Basol, M., berinsky adam, a., Betsch, C., Cook, J., Fazio, L., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., … Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *Forthcoming*.

Kpodar, K., & Abdallah, C. (2017). Dynamic fuel price pass-through: Evidence from a new global retail fuel price database. *Energy Economics*, *66*, 303–312.

Laux, J., Wachter, S., & Mittelstadt, B. (2021). Taming the few: Platform regulation, independent audits, and the risks of capture created by the dma and dsa. *Computer Law & Security Review*, *43*, 105613.

Lazer, D., 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.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, *111*(3), 831–70.

Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2024a). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, *19*(2), 477–488.

Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2024b). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, *19*(2), 477–488.

Matias, J. N. (2016). The civic labor of online moderators. *Internet Politics and Policy Conference.*

Maurya, P. K., Bansal, R., & Mishra, A. K. (2023). Russia–ukraine conflict and its impact on global inflation: An event study-based approach. *Journal of Economic Studies*, *50*(8), 1824–1846.

Meta. (2023). How enforcement technology works [Accessed: 2023-11-23].

Meta. (2025). Introducing community notes [Accessed: 2025-06-03].

Mitts, T., Pisharody, N., & Shapiro, J. (2022). Removal of anti-vaccine content impacts social media discourse. *WebSci*.

Moravec, P. L., Kim, A., & Dennis, A. R. (2020). Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research*, *31*(3), 987–1006.

Moravec, P. L., Minas, R. K., & Dennis, A. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly*, *43*(4), 1343–1360.

Morrar, R., Arman, H., & Mousa, S. (2017). The fourth industrial revolution (industry 4.0): A social innovation perspective. *Technology innovation management review*, *7*(11), 12–20.

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, *13*(1), 7144.

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

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.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., & Rand, D. G. (2019a). Fighting misinformation on social media using crowdsourced judgments of news source quality. *PNAS*, *116*(7), 2521–2526.

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

Pérez-Escolar, M., Lilleker, D., & Tapia-Frade, A. (2023). A systematic literature review of the phenomenon of disinformation and misinformation. *Media and communication*, *11*(2), 76–87.

Pew Research Center. (2016). News use across social media platforms 2016 [Accessed: 2022-04-02].

Pew Research Center. (2024). Social media and news fact sheet [Accessed: 2025-04-11].

Phillips, S., Wang, S. Y. N., Carley, K. M., Rand, D., & Pennycook, G. (2024). Emotional language reduces belief in false claims. *OSF*, jn23a.

Pilarski, M., Solovev, K., & Pröllochs, N. (2024). Community notes vs. snoping: How the crowd selects fact-checking targets on social media. *ICWSM*.

Poynter. (2019). Most republicans don't trust fact-checkers, and most Americans don't trust the media [Accessed: 2022-04-02].

Prieto Gutiérrez, J. J., Segado Boj, F. J., & Da Silva França, F. (2023). Artificial intelligence in social science: A study based on bibliometrics analysis.

Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. *ICWSM*.

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021a). Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports*, *11*, 22721.

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021b). Emotions in online rumor diffusion. *EPJ Data Science*, *10*(1), 51.

Resnick, P., Alfayez, A., Im, J., & Gilbert, E. (2023). Searching for or reviewing evidence improves crowdworkers' misinformation judgments and reduces partisan bias. *Collective Intelligence*, *2*(2), 26339137231173407.

Roberts, S. T. (2020). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature human behaviour*, *7*(5), 812–822.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., & Van Der Linden, S. (2020). Susceptibility to misinformation about Covid-19 around the world. *Royal Society Open Science*, *7*(10), 201199.

Russo, G., Horta Ribeiro, M., Casiraghi, G., & Verginer, L. (2023). Understanding online migration decisions following the banning of radical communities. *WebSci*.

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. *WWW Companion*.

Shore, J., Baek, J., & Dellarocas, C. (2018). Network structure and patterns of information diversity on twitter. *MIS Quarterly*, *42*(3), 849–972.

Solovev, K., & Pröllochs, N. (2022). Moral emotions shape the virality of covid-19 misinformation on social media. *WWW*.

Solovev, K., & Pröllochs, N. (2025). References to unbiased sources increase the helpfulness of community fact-checks. *arXiv preprint arXiv:2503.10560*.

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior*, *149*, 107917.

Statista. (2025). Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028 [Accessed: 2025-06-03].

Steelman, Z. R., Hammer, B. I., & Limayem, M. (2014). Data collection in the digital age. *MIS quarterly*, *38*(2), 355–378.

TikTok. (2023). Community guidelines enforcement report [Accessed: 2023-11-23].

TikTok. (2025). Testing a new feature to enhance content on tiktok [Accessed: 2025-06-03].

Timmerman, B., Mehta, P., Deb, P., Gallagher, K., Dolan-Gavitt, B., Garg, S., & Greenstadt, R. (2023). Studying the online deepfake community. *Journal of Online Trust and Safety*, *2*(1).

Tirole, J. (2023). Competition and the industrial challenge for the digital age. *Annual Review of Economics*, *15*(1), 573–605.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of economic perspectives*, *28*(2), 3–28.

Vieira, F. V., & da Silva, C. G. (2024). Global inflation before and after the covid-19 pandemic: A panel data approach. *Economics Bulletin*, *44*(3), 889–903.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375.

Wang, S. Y. N., Phillips, S. C., Carley, K. M., Lin, H., & Pennycook, G. (2025). Limited effectiveness of psychological inoculation against misinformation in a social media feed. *PNAS nexus*, *4*(6), pgaf172.

World Economic Forum. (2025). Global risks report.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*, 40–53.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M., Coleman, K., & Baxter, J. (2022). Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv*. https://arxiv.org/abs/2210.15723

World Economic Forum. (2024). Global risks report.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explorations Newsletter*, *21*(2), 80–90.

X. (2021). Introducing Birdwatch, a community-based approach to misinformation.

X. (2023). X is commited to the open exchange of information [Accessed: 2023-11-23].

X. (2024). Diversity of perspectives [Accessed: 2024-11-23].

X. (2025). About community notes on x [Accessed: 2025-06-03].

YouTube. (2023). Government requests report [Accessed: 2023-12-04].

YouTube. (2024). Testing new ways to offer viewers more context and information on videos.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. *CHI*.

Chapter 2

Diffusion of Community Fact-Checked Misinformation on Twitter

Abstract

The spread of misinformation on social media is a pressing societal problem that platforms, policymakers, and researchers continue to grapple with. As a countermeasure, recent works have proposed to employ non-expert fact-checkers in the crowd to fact-check social media content. While experimental studies suggest that crowds might be able to accurately assess the veracity of social media content, an understanding of how crowd fact-checked (mis-) information spreads is missing. In this work, we empirically analyze the spread of misleading vs. not misleading community fact-checked posts on social media. For this purpose, we employ a dataset of community-created fact-checks from Twitter's "Birdwatch" pilot and map them to resharing cascades on Twitter. Different from earlier studies analyzing the spread of misinformation listed on third-party fact-checking websites (e. g., snopes.com), we find that community fact-checked misinformation is less viral. Specifically, misleading posts are estimated to receive 36.62% partial explanation may lie in differences in the fact-checking targets: community fact-checkers tend to fact-check posts from influential user accounts with many followers, while expert fact-checks tend to target posts that are shared by less influential users. We further find that there are significant differences in virality across different sub-types of misinformation (e. g., factual errors, missing context, manipulated media). Moreover, we conduct a user study to assess the perceived reliability of (real-world) community-created fact-checks. Here, we find that users, to a large extent, agree with community-created fact-checks. Altogether, our findings offer insights into how misleading vs. not misleading posts spread and highlight the crucial role of sample selection when studying misinformation on social media.

Keywords: social media, misinformation, fact-checking, crowd wisdom, information diffusion

2.1 Introduction

There are widespread concerns that misinformation on social media is damaging societies and democratic institutions (Lazer et al., 2018). In recent years, viral misinformation on social media has been observed repeatedly, especially during elections and crisis situations (Allcott & Gentzkow, 2017; Bakshy et al., 2015; Geissler et al., 2022; Oh et al., 2013; Pennycook, Bear, et al., 2020). In order to identify and eventually curb the spread of misinformation, expert fact-checkers of various third-party fact-checking organizations (e. g., snopes.com, politifact.com, factcheck.org) regularly investigate the veracity of social media rumors (Vosoughi et al., 2018; Wu et al., 2019). However, due to the limited amount of fact-checks that can be performed by these organizations, they are unable to accommodate the amount and speed of content creation on social media. Misinformation thus often continues to circulate and may only be detected when a tremendous amount of attention is paid to it (Epstein et al., 2020). Furthermore, about 50% of all Americans have concerns regarding the independence of the experts' assessment, i. e., distrust professional fact-checkers (Poynter, 2019). Given these challenges, the real-world impact of fact-checks from third-party fact-checking organizations may be limited.

In order to address the drawbacks of the expert verification approach, recent research has proposed to employ non-expert fact-checkers in the crowd to verify social media content (Allen et al., 2020, 2021; Bhuiyan et al., 2020; Drolsbach & Pröllochs, 2023; Epstein et al., 2020; Godel et al., 2021; Micallef et al., 2020; Pennycook & Rand, 2019). The rationale is that the "wisdom of crowds" (i. e., the aggregated assessments of non-expert fact-checkers) could result in an accuracy that is similar to that of experts (Frey & van de Rijt, 2021). Compared to the expert verification approach, harnessing the crowd for fact-checking would enable large numbers of fact-checks that could be carried out at higher frequency and lower cost (Allen et al., 2021; Pennycook & Rand, 2019). Furthermore, crowd-based fact-checking has the potential to remedy the problem of distrust in expert fact-checkers (Allen et al., 2021). Recent experimental studies indeed yielded promising results – suggesting that even relatively small crowds achieve an accuracy comparable to that of experts when fact-checking social media content (Bhuiyan et al., 2020; Epstein et al., 2020; Pennycook & Rand, 2019).

While community-based fact-checking systems might be able to produce accurate fact-checks at scale, an understanding of how (mis-)information diffuses through social networks is still in its infancy. Prior works have analyzed the spread of rumors that have been fact-checked by third-party fact-checking organizations (Friggeri et al., 2014; Pröllochs et al., 2021a; Solovev & Pröllochs, 2022b; Vosoughi et al., 2018). For instance, several studies have compared characteristics of resharing cascades (e. g., how often a social media post is shared) across true vs. false rumors, finding that falsehood is more viral than the truth (Pröllochs et al., 2021a; Solovev & Pröllochs, 2022b; Vosoughi et al., 2018). However, third-party fact-checking organizations tend to fact-check rumors on topics that are deemed to be of interest to a broad public and/or particularly concerning from the perspective of experts, while other misinformation remains unnoticed. In contrast, community fact-checked posts represent social media content that has been deemed worth fact-checking by actual social media users. Analyzing their diffusion would shed new light on the question of whether misinformation is more viral than the truth – or rather a result of sample selection. However, we are not aware of any previous research analyzing the diffusion of crowd fact-checked posts on social media. Moreover, little is known about which social media posts are picked up in community-based fact-checking and how the spread varies across different types of misinformation (e. g., factual errors, missing context). Answering these questions is the goal of this study.

Research questions: In this work, we empirically analyze the diffusion of misleading vs. not misleading social media posts that have been fact-checked by the crowd. Specifically, we address the following research questions:

- **(RQ1)** How do community fact-checked posts spread on social media? Are misleading posts more viral than not misleading posts?
- **(RQ2)** Are there differences in virality across different sub-types of community fact-checked misinformation (e. g., factual errors, missing context, manipulated media)?
- **(RQ3)** How do the fact-checking targets differ between community fact-checkers and expert fact-checkers?
- **(RQ4)** To what extent are (real-world) community-created fact-checks perceived as reliable?

Data & methodology: We collect a comprehensive dataset consisting of community-created fact-checks from Twitter's Birdwatch platform. We then map the fact-checks to the fact-checked tweet using Twitter's historical API. This allows us to calculate the size of the resharing cascades (i. e., the number of retweets) in order to measure the virality of the fact-checked post. Subsequently, we implement an empirical regression model and link the fact-checking label to the number of retweets. We further control for the sentiment of the post and the social influence of its author (e. g., number of followers, account age, etc.). We then perform hypothesis testing to analyze whether posts categorized as being misleading are more viral than not misleading posts.

Contributions: This study is the first to analyze the spread of crowd fact-checked misinformation on social media. We show that crowd fact-checked misleading posts are *less* viral than not misleading posts. Specifically, misleading posts are estimated to receive 36.85 % fewer retweets than not misleading posts. Notably, this finding differs from earlier work (Vosoughi et al., 2018), which has analyzed the diffusion of misinformation that has been fact-checked by third-party fact-checking organizations. We find that a partial explanation may lie in differences in the fact-checking targets: our findings suggest that community fact-checkers tend to fact-check posts from influential user accounts with many followers, while expert fact-checks tend to target rumors that are shared by less influential accounts. Our results further imply that there are significant differences in virality across different sub-types of misinformation (e. g., factual errors, missing context, manipulated media).

As an additional contribution, we conduct a user study to assess the perceived reliability of (real-world) community-created fact-checks. Here, we observe that users agree with a large share of community-created fact-checks, whereas only a relatively small share is perceived as being purposely deceptive (e.g., due to motivated reasoning).

2.2 Background

2.2.1 Misinformation on Social Media

Over the last decade, the importance of social media (e. g., Twitter, Facebook) as an information platform for large parts of the society has been subject to considerable growth (Lazer et al., 2018; Pew Research Center, 2016). On social media, any user can share information with his/her follower base (Shore et al., 2018). Compared to traditional media, there is little control

authority or oversight regarding the contents. For this reason, social media is highly vulnerable to the spread of misinformation. In fact, previous research suggests that social media platforms have become primary enablers of misinformation (e. g., Lazer et al., 2018). Online exposure to misinformation can affect how opinions are formed and causes detrimental societal effects (e. g., Allcott & Gentzkow, 2017; Bär et al., 2023b; Del Vicario et al., 2016). The latter has been repeatedly observed, especially during elections (e. g., Allcott & Gentzkow, 2017; Bakshy et al., 2015) and crisis situations (e. g., Geissler et al., 2022; Oh et al., 2010, 2013; Pennycook, McPhetres, et al., 2020; Solovev & Pröllochs, 2022b; Starbird et al., 2014).

A key feature of modern social media platforms is that users can also share others' content to increase its reach (e. g., "retweeting" on Twitter). This can result in misinformation cascades going "viral." While previous research has mainly focused on characteristics and (negative) consequences of misinformation on social media, studies analyzing differences in the virality across misleading vs. not misleading posts are relatively scant. Existing works in this direction have analyzed the diffusion of posts that have been fact-checked by third-party fact-checking organizations (Friggeri et al., 2014; Pröllochs & Feuerriegel, 2023; Pröllochs et al., 2021a; Solovev & Pröllochs, 2022b; Vosoughi et al., 2018). These studies found that misinformation diffuses significantly more virally than the truth. We are not aware of any previous study analyzing the spread of misleading vs. not misleading social media posts that have been fact-checked by the crowd.

2.2.2 Fact-Checking on Social Media

Reliable fact-checking strategies are a crucial necessity to limit the spread of misinformation on social media. Currently, there are two predominant strategies. First, expert assessment in the form of human experts can check the veracity of content; e. g., via third-party fact-checking platforms (e. g., snopes.com, politifact.com, factcheck.org). Second, machine learning models can be trained to automatically classify misinformation (Ma et al., 2016; Qazvinian et al., 2011). For this purpose, content-based features (e. g., text, images, video), context-based features (e. g., time, location), or propagation patterns (i. e., how misinformation circulates among users) can be used. However, both methods suffer from several drawbacks. While experts classify misinformation fairly accurately, this strategy is difficult to scale due to the limited number of available humans experts (Micallef et al., 2022; Pennycook & Rand, 2019). Besides, a large proportion of social media users do not trust the independence of expert fact-checkers (Poynter, 2019). In contrast, machine learning-based approaches are straightforward to scale, but typically show comparatively low accuracy (Wu et al., 2019).

Given the trade-off between scalability and accuracy of existing approaches, recent works have proposed to outsource fact-checking of social media content to non-expert fact-checkers in the crowd (Allen et al., 2020, 2021; Bhuiyan et al., 2020; Epstein et al., 2020; Godel et al., 2021; Micallef et al., 2020; Pennycook & Rand, 2019). The rationale is that the "wisdom of crowds" (i. e., the aggregated assessments of non-expert fact-checkers) could result in an accuracy that is comparable to that of experts (Frey & van de Rijt, 2021; Woolley et al., 2010). The ability of crowds to ensure relatively trustworthy and high-quality accumulation of knowledge has been observed in various other online settings, such as on platforms like Wikipedia and Stack Overflow (e. g., Dissanayake et al., 2019; Han et al., 2021; Okoli et al., 2014). Applying a crowd-based approach to fact-check social media posts might have several benefits (Pennycook & Rand, 2019). First, compared to expert assessments, significantly larger quantities of posts could be fact-checked. Second, trust issues with expert fact-checkers

could, at least partially, be mitigated. Experimental studies suggest that, while the assessment of individuals might be noisy and ineffective (Woolley et al., 2010), the crowd can be quite accurate in identifying misleading social media content. Here the assessment of even relatively small crowds has been found to be comparable to those of experts (Bhuiyan et al., 2020; Epstein et al., 2020; Pennycook & Rand, 2019). Despite challenges with politically motivated reasoning (Allen et al., 2022; Pröllochs, 2022), recent research further shows that users, to a large extent, perceive community-created fact-checks for social media posts as being informative and helpful (Pröllochs, 2022).

2.3 Data

2.3.1 Data Source: Community Fact-Checked Tweets from Birdwatch

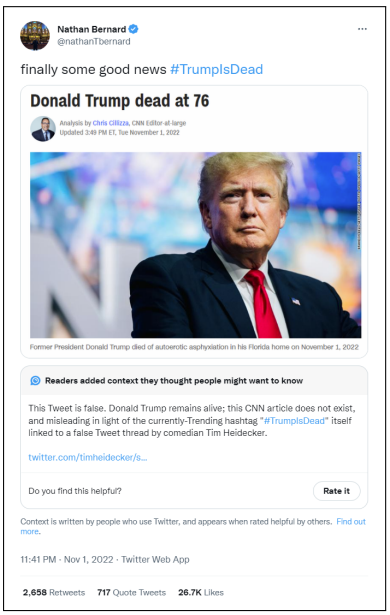
We analyze the spread of social media posts that have been community fact-checked on Twitter's Birdwatch pilot (X, 2021). On January 23, 2021, Twitter launched Birdwatch as a new approach to address misinformation on their platform (X, 2021). The goal is to fact-check social media content by harnessing the "wisdom of crowds." Birdwatch allows users to identify tweets they believe are misleading or not misleading and write notes that provide context to the tweet (so-called "Birdwatch notes"). Users can fact-check *any* tweet they come across on Twitter – directly when browsing Twitter (see examples in Fig. 3.1). Community fact-checking on Birdwatch comprises (1) checkbox questions that allow users to state whether a tweet might or might not be misleading (*Fact-Checking Label*); (2) an open text field (max 280 characters) where users can explain their judgment (*Text Explanation*), and (3) checkbox questions in which users can characterize the tweet and select reasons *why* they perceive the tweet as being misleading (*Misinformation Type*). For the latter, Birdwatch users can select one (or multiple) of the following answer options: (i) "Factual Error," (ii) "Missing Important Context," (iii) "Unverified Claim as Fact," (iv) "Outdated Information," (v) "Manipulated Media," (vi) "Satire," and (vii) "Other."

After a Birdwatch note is submitted, the fact-check is publicly available for other users to read. Birdwatch also features a rating system, which allows users to rate the helpfulness of the community-created fact-checks. These ratings are supposed to help identify which notes are most helpful and raise their visibility. Specifically, Birdwatch notes are shown directly on the fact-checked tweet if (i) the tweet is classified as misleading and (ii) it is rated by the community to be particularly helpful (see Fig. 3.1).

Importantly, the data for this study originates from Birdwatch's pilot phase in the US. During this pilot phase, interested users were required to actively sign up to join Birdwatch. Any Twitter user could apply to become a Birdwatch contributor. Users that had signed up on Birdwatch could see Birdwatch notes directly when browsing Twitter next to the fact-checked tweet. Non-participating users could access Birdwatch notes via a separate Birdwatch website. In early 2022, Birdwatch had approximately 3250 contributors, compared to 41.5 million daily active Twitter users in the US (Statista, 2022). Hence, during Birdwatch's pilot phase, community fact-checks from Birdwatch were practically not visible to the vast majority of

social media users and, thus, were unlikely to directly influence the diffusion of the fact-checked tweets.[1] Furthermore, Birdwatch notes did not have an effect on the way people see tweets or other system recommendations (X, 2021). The Birdwatch pilot phase thus provides a unique opportunity to study the spread of community fact-checked posts with little confounding factors.

A Misleading



B Not misleading

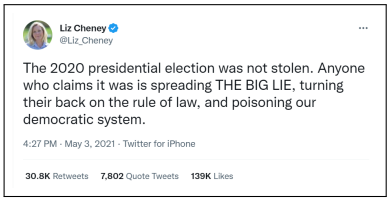


Figure 2.1: Examples of community fact-checked tweets. Only Birdwatch notes for misleading tweets are eligible to be directly shown on tweets. However, during our study period, community fact-checks from Birdwatch were practically not visible to the vast majority of social media users (i. e., only to pilot participants) and, thus, were unlikely to directly influence the diffusion of the fact-checked tweets. (a) Example of a tweet classified as misleading (*Fact-Checking Label*) and the *Text Explanation* of the corresponding Birdwatch note. The contributor selected "Factual Error," "Manipulated Media," and "Satire" as reasons for his/her classification (*Misinformation Type*). (b) Example of a tweet classified as not misleading.

2.3.2 Data Collection

We downloaded *all* Birdwatch notes between the introduction of the feature on January 23, 2021, and the end of February 2022 from the Birdwatch website, i. e., for an observation period of more than one year. The dataset contains a total number of 20 218 Birdwatch notes (i. e., community-created fact-checks) from 3 257 different contributors. We used the Twitter historical API to map the *tweetID* referenced in each Birdwatch note to the source tweet. This approach allowed us to collect the following information about each source tweet and the account of its authors: (i) the number of retweets, (ii) the number of followers, (iii) the number of followees, (iv) the account age, and (v) whether the user has been verified by Twitter.

Notably, multiple Birdwatch users can write Birdwatch notes for the same tweet. Therefore, the data sometimes includes multiple fact-checks for the same post. The average number of

[1] In early October 2022 (i. e., after our observation period), Twitter started to expand the Birdwatch program, allowing more Twitter users to view fact-checks directly on Twitter. Furthermore, Twitter rebranded Birdwatch to "Community Notes."

Birdwatch notes per tweet is 1.33, with few tweets having many notes and most tweets having few. Only 18.79% of the fact-checked tweets received more than one Birdwatch note. To avoid distortions due to multiple fact-checked tweets, we focus our analysis on the temporally first fact-check after the tweet has been posted. This filtering step resulted in a dataset consisting of 15 256 unique fact-checks (for 15 256 unique tweets). As part of our robustness checks, we also tested alternative approaches for handling multiple fact-checks (e. g., using Birdwatch's rating system, majority vote). Here we obtained qualitatively identical results.

2.3.3 Variable Description

Our dataset contains variables from two sources: (i) variables that are provided by the community-created fact-checks (i. e., the Birdwatch notes); and (ii) variables that represent information about the source tweet (e. g., the social influence of the author of the fact-checked tweet).

Fact-checks: The Birdwatch notes provide us with the following variables:

- *Misleading:* A binary indicator of whether a tweet has been reported as being misleading by the author of the Birdwatch note ($= 1$; otherwise $= 0$).
- *Delay:* A numeric variable measuring the number of days elapsed between the posting date of the source tweet and the fact-check.
- *Misinformation Type:* Seven dummy variables indicating reasons why a tweet has been reported as being misleading ("Factual Error," "Missing Important Context," "Unverified Claim as Fact," "Outdated Information," "Manipulated Media," "Satire," and "Other").

Source tweet: We used the Twitter historical API to map the *tweetID* referenced in each Birdwatch note to the source tweet and collected the following information about each source tweet:

- *Retweet Count:* A numeric variable denoting the number of retweets a single tweet receives on Twitter. The retweet count is a common measure for the virality of a resharing cascade (e. g., Brady et al., 2017; Solovev & Pröllochs, 2022b).
- *Followers:* The number of followers, i. e., the number of accounts that follow the author of the source tweet on Twitter.
- *Followees:* The number of followees, i. e., the number of accounts whom the author of the source tweet follows on Twitter.
- *Account Age:* The age of the author of the source tweet's account (in years).
- *Verified:* A binary dummy indicating whether the account of the source tweet has been officially verified by Twitter ($= 1$; otherwise $= 0$).
- *Sentiment:* We calculate a sentiment score measuring the positivity/negativity of the source tweet. Here we use a dictionary-based approach analogous to earlier research (e. g., Bär et al., 2023a; Jakubik et al., 2023; Rho & Mazmanian, 2020; Robertson et al., 2023; Vosoughi et al., 2018). We first remove stopwords, punctuation, special characters (e. g., hashtags), and URLs in each source tweet. Subsequently, we employ the NRC lexicon (Mohammad & Turney, 2013), which categorizes English words into positive and negative words. Following previous work (e. g., Rho & Mazmanian, 2020;

Solovev & Pröllochs, 2022a), the sentiment scores are then measured by calculating the difference between positive and negative words relative to the tweet length. For our sentiment analysis, we use the default implementation of the `sentimentr` package (with the built-in NRC lexicon) that also accounts for negations and valence shifters (see Rinker, 2019 for details).

2.4 Empirical Analysis

2.4.1 Diffusion of Misleading vs. Not Misleading Posts (RQ1)

We now empirically analyze the diffusion of misleading vs. not misleading posts that have been fact-checked on Twitter's Birdwatch platform. For this purpose, we first compare summary statistics. Note, however, that summary statistics should be interpreted with caution as the virality of social media posts strongly depends on the social influence of the author. To account for such confounding effects, we subsequently implement an empirical regression model with control variables that links the fact-checking label to the number of retweets. We then perform hypothesis testing to analyze whether posts categorized as being misleading are more viral than not misleading posts.

Summary statistics: Birdwatch users are vastly more likely to report misleading tweets than not misleading tweets. Out of 15 256 community fact-checked tweets, 14 384 (94.28 %) are classified as misleading and 872 (5.72 %) are classified as not misleading. In total, the fact-checked tweets in our dataset have been retweeted 29.45 million times. However, the retweet volume is higher for not misleading tweets than for misleading tweets. Specifically, the average retweets count amounts to 2 478 for not misleading tweets and to 1 478 for misleading tweets. A two-sided t-test confirms that the difference in means are statistically significant ($p < 0.01$). Misleading vs. not misleading tweets also exhibit considerable heterogeneity with regards to sentiment and the social influence of the author. The sentiment tends to be significantly more positive in not misleading tweets (mean sentiment of 0.022) than in misleading tweets (mean sentiment of -0.004). Misleading tweets are posted by users that have, on average, 41.17 % fewer followers. Also here, two-sided t-tests confirm that the difference in means are statistically significant ($p < 0.01$). We find only small differences in means for the variables *Followees*, *Account Age* and, *Verified*, which are not statistically significant at common significance thresholds. Fig. 2.2 further visualizes the complementary cumulative distribution functions (CCDFs). Kolmogorov-Smirnov (KS) tests show that, with the exception of *Account Age*, the differences in the distributions between misleading and not misleading tweets are statistically significant ($p < 0.01$).

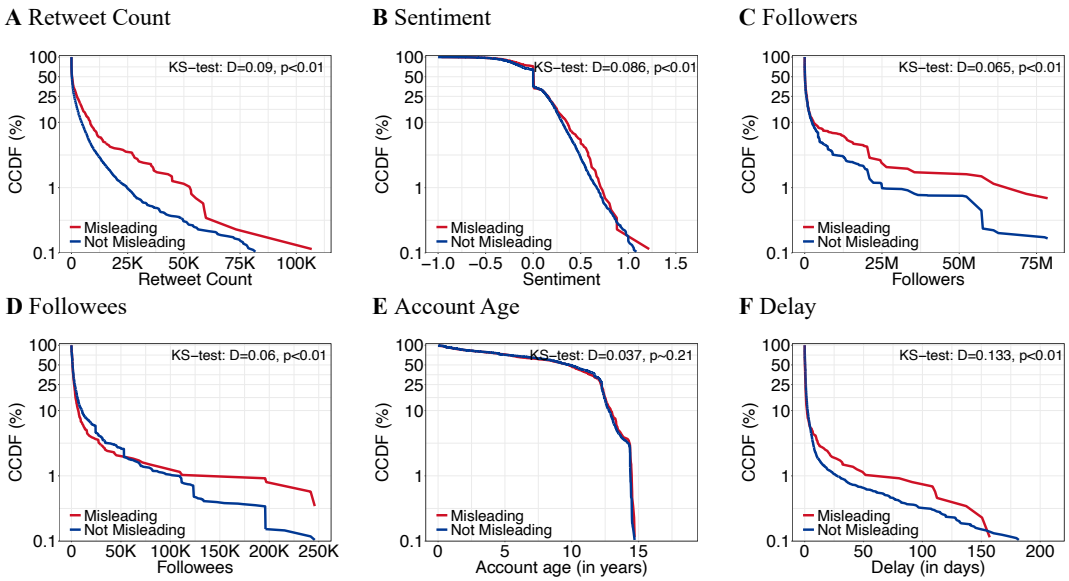


Figure 2.2: Complementary cumulative distribution functions (CCDFs) for (a) *Retweet Count*, (b) *Sentiment*, (c) *Followers*, (d) *Followees*, (e) *Account Age*, and (f) *Delay*.

Regression model: We implement explanatory regression analysis to better understand the diffusion of misleading vs. not misleading crowd fact-checked posts. In contrast to summary statistics, this allows us to estimate effect sizes after controlling for confounding effects. The dependent variable in our regression analysis is given by $RetweetCount_i$, that is, the number of retweets for a fact-checked tweet i. The retweet count is a non-negative count variable, and its variance is larger than the mean. To adjust for overdispersion, we draw upon a negative binomial regression to model the retweets count (Pröllochs et al., 2021a; Solovev & Pröllochs, 2022b). The key explanatory variable is $Misleading_i$, i. e., whether the tweet has been classified as misleading by Birdwatch users (i. e., $= 1$ if true, otherwise $= 0$). Additionally, we include the elapsed time between the publication of the tweet and the fact-check ($Delay_i$). Furthermore, we must control for the social influence of the source tweet and its author. Therefore, we adjust for variables known to affect the retweet rate (Brady et al., 2017; Pröllochs et al., 2021b; Solovev & Pröllochs, 2022b; Stieglitz & Dang-Xuan, 2013; Vosoughi et al., 2018), which includes the number of followers ($Followers_i$) and followees ($Followees_i$), the account age ($AccountAge_i$), and whether the account was verified by Twitter ($Verified_i$). In addition, we control for the sentiment of the source tweet ($Sentiment_i$). The resulting model is

with intercept β_0 and month-year fixed effects u_i to adjust for differences in the start date and age of the resharing cascades. For the sake of interpretability, we z-standardize all continuous variables. This allows us to compare the effects of regression coefficients on the dependent variable measured in standard deviations. Note that since we apply a negative binomial regression, the interpretation of the effect sizes requires an exponential transformation of the coefficients.

Coefficient estimates: The coefficient estimates for the regression model are reported in Fig. 6.4. We find that misleading tweets are significantly less viral than not misleading tweets. Specifically, the coefficient for *Misleading* is -0.456 ($p < 0.01$), which implies that misleading tweets are expected to receive $e^{-0.459} - 1 \approx 36.62\,\%$ fewer retweets. Furthermore, we observe that the coefficient estimate for *Delay* is small in magnitude and not statistically

significant at common significance threshold. This implies that differences in the fact-checking speed are not significantly associated with differences in virality of crowd fact-checked posts.

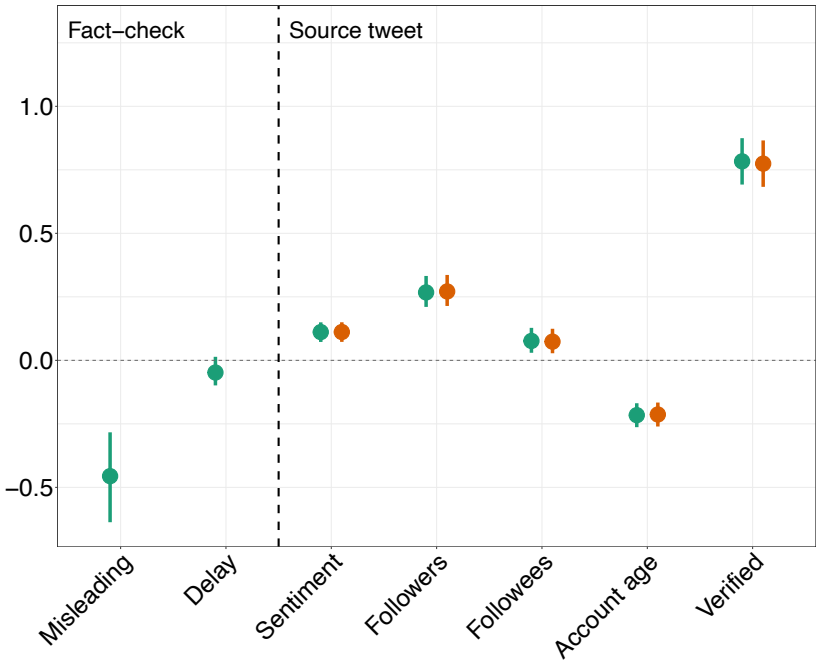


Figure 2.3: Coefficient estimates for negative binomial regression with the retweet count as dependent variable. Model (a) includes all variables given by the source tweet (orange). Model (b) additionally includes variables concerning the fact-check (green). The vertical bars represent 99 % confidence intervals. Month-year fixed effects are included.

Concordant with the literature (Solovev & Pröllochs, 2022b; Stieglitz & Dang-Xuan, 2013; Vosoughi et al., 2018), we observe statistically significant estimates for the variables characterizing the social influence of the author of the source tweet. The number of followers has a large positive effect on the number of retweets (coef: 0.267; $p < 0.01$), while the number of followees has a smaller positive effect (coef: 0.076; $p < 0.01$). A higher account age decreases the expected number of retweets (coef: -0.216; $p < 0.01$), while posts from verified accounts are expected to receive more retweets (coef: 0.783; $p < 0.01$). Similar to earlier work (Pröllochs et al., 2021a), we also find that more positive sentiment is associated with more retweets (coef: 0.111; $p < 0.01$).

2.4.2 Diffusion of Different Types of Misinformation (RQ2)

If fact-checkers on Birdwatch have classified a tweet as being misleading, they additionally need to answer checkbox questions on the reasons *why* they perceive it as such. As aforementioned, Birdwatch users can select one (or multiple) of the following answer options: (i) "Factual Error," (ii) "Missing Important Context," (iii) "Unverified Claim as Fact," (iv) "Outdated Information," (v) "Manipulated Media," (vi) "Satire," and (i) "Other." Fig. 2.5 shows that the vast majority of tweets have been categorized as misleading because of factual errors (62.13 %), missing context (61.38 %), or because they treat unverified claims as fact (49.99 %). The other categories are relatively rare.

We repeat our regression analysis with dummy variables referring to the different types of misleading posts as provided by Birdwatch contributors. This allows us to examine differences in the virality across different types of misinformation. The coefficient estimates in Fig. 2.4 show that misleading tweets are less viral than not misleading tweets if they belong to the misinformation sub-types "Factual Error" (coef: -0.251; $p < 0.01$), "Missing Important Context" (coef: -0.127; $p < 0.01$), "Unverified Claim as Fact" (coef: -0.300; $p < 0.01$) and, "Other" (coef: -0.221; $p < 0.01$). In contrast, tweets belonging to the misinformation sub-types "Manipulated Media" (coef: 0.461; $p < 0.01$), and "Satire" (coef: 0.411; $p < 0.01$) receive more retweets. These results suggest that there are significant differences in virality across different sub-types of misinformation. The coefficient estimates for the other variables do not differ qualitatively from the previously performed regressions.

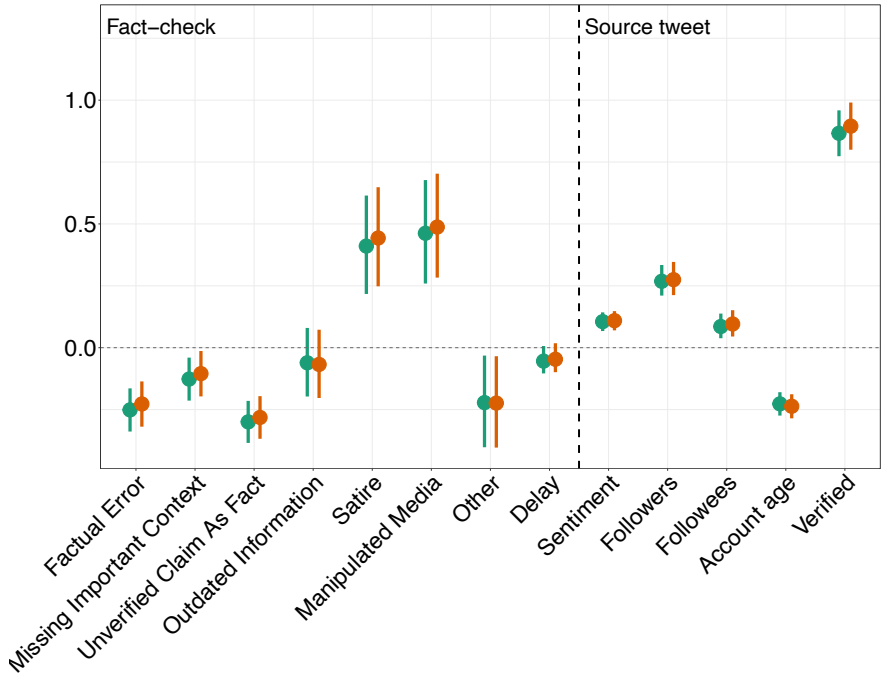


Figure 2.4: Coefficient estimates for negative binomial regression with the retweet count as dependent variable. Here, dummy variables referring to different sub-types of misinformation are included. Model (a) includes all posts (green), whereas Model (b) only includes the subset of posts classified as misleading (orange). The reference type in Model (a) are tweets classified as "not misleading," whereas the reference type in Model (b) are misleading tweets that have not been assigned to a subtype. The vertical bars represent 99 % confidence intervals. Month-year fixed effects are included.

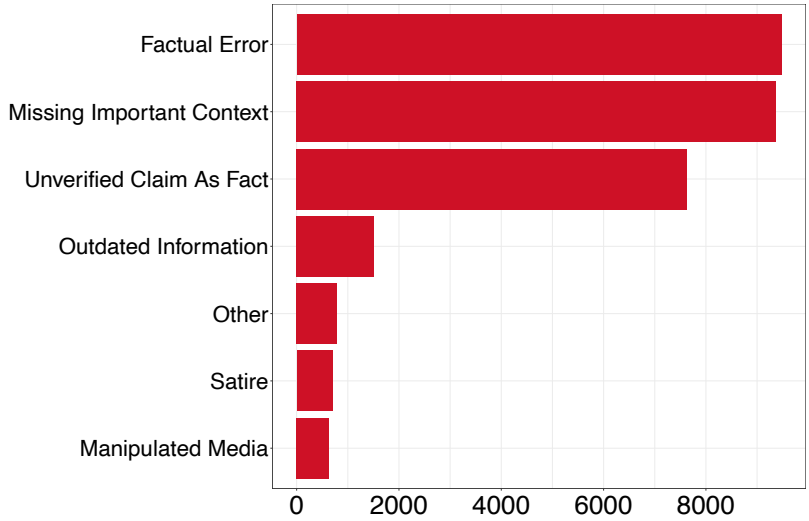


Figure 2.5: Barplot showing the number of tweets per checkbox answer option in response to the question "Why do you believe this tweet may be misleading?"

2.4.3 Comparison to Expert-Based Fact-Checking (RQ3)

In contrast to the work by Vosoughi et al. (2018), which found that expert fact-checked falsehood on Twitter is *more* viral than the truth, our analysis suggests that crowd fact-checked tweets perceived as misleading are *less* viral than those perceived as not misleading. A possible explanation for this finding lies in the sample selection, i. e., third-party fact-checking organizations vs. Birdwatch contributors might fact-check social media posts published by different account types.

To shed light on this question, Fig. 2.6 compares the mean values of different user characteristics of the authors of misleading and not misleading crowd fact-checked posts to those of authors true and false rumors in the dataset of expert fact-checked posts from Vosoughi et al. (2018). Compared to expert fact-checked tweets, we find that user accounts of authors of crowd fact-checked posts have, on average, ≈ 40 times more followers, 41.65 % more followees, and approximately twice the account age. Moreover, while 49.21 % percent of the accounts of authors of crowd fact-checked posts are verified by Twitter, this is only the case for 2.00 % of the authors of expert fact-checked posts. Two-sided t-tests confirm that each difference in means is statistically significant ($p < 0.01$). These findings suggest that social media users contributing to crowd-based fact-checking tend to fact-check posts from larger accounts with greater social influence, while expert fact-checks tend to target rumors that are shared by smaller accounts.

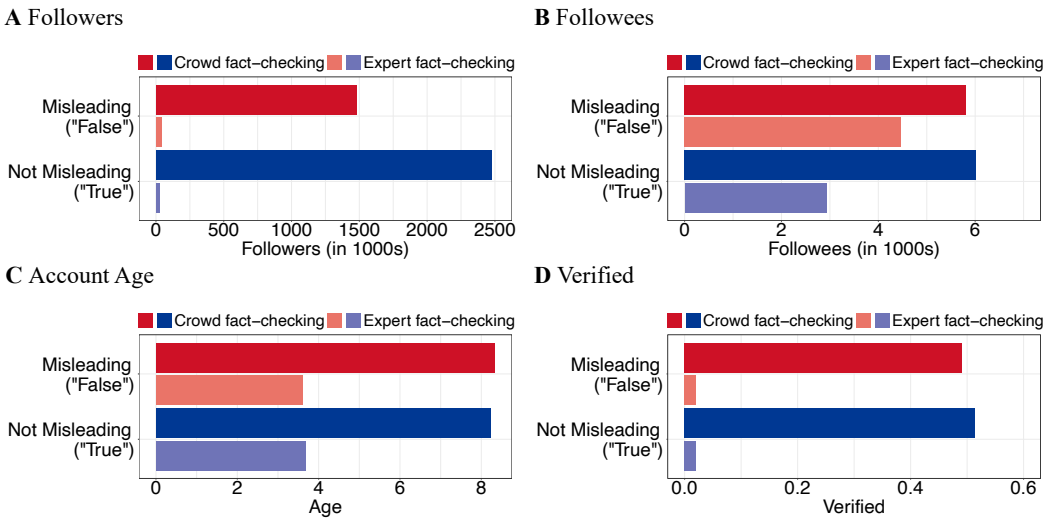


Figure 2.6: Comparison of characteristics (mean values) of authors of crowd fact-checked and expert fact-checked tweets for (a) the number of followers, (b) the number of followees, (c) the account age and, (d) the verified status. We compare the authors of misleading and not misleading crowd fact-checked posts on Birdwatch to those of true and false rumors in the dataset of expert fact-checked posts from Vosoughi et al. (2018).

We further observe that, for expert fact-checked posts, falsehood tends to originate from accounts with relatively more followers, while we observe the opposite pattern for crowd fact-checked posts (see Fig. 2.6). Specifically, we find that authors of crowd fact-checked posts perceived as misleading have 67.71 % more followers than accounts of posts perceived as not misleading ($p < 0.01$). In contrast, authors of falsehood in expert fact-checked posts have 34.04 % less followers than authors of the true tweets ($p < 0.01$). This suggests that fact-checks from Birdwatch contributors are more likely to endorse/emphasize the accuracy of not misleading tweets authored by influential users with a wide reach. Opposite to this, expert fact-checked tweets authored by influential accounts are more likely to convey false information. Since author characteristics are inherently linked to the virality of posts (e. g., users with a wider reach can generate more retweets), the observed differences in fact-checking targets provide a (partial) explanation for the overall higher virality of not misleading posts in the case of Birdwatch.

2.4.4 Perceived Reliability of Community-Created Fact-Checks (RQ4)

In order to assess the perceived reliability of the community-based fact-checks from Birdwatch, we conducted a user study on the online survey platform Prolific (www.prolific.co). We recruited $n = 7$ participants, four women and three men, who were on average 35 years old. All participants were based in the US, and English native speakers. All but one participant indicated that they are familiar with Twitter and regularly share content on social media. Participants were presented with a randomized sample of 300 tweets (150 not misleading and 150 misleading) and the corresponding fact-checks from Birdwatch (fact-checking label and text explanation). Note that we purposely presented the participants with both the source tweet and the fact-check (instead of only the source tweet). In the absence of a ground truth (which might require expert assessment), we were interested in the *perceived* reliability of the fact-checks rather than testing how much one crowd agrees with another. As such, for each tweet, participants were asked for their assessment on (i) the extent to which they agree with

the fact-checking label, and (ii) whether they perceive the fact-check as purposely deceptive (e. g., because of motivated reasoning, manipulation attempts, etc.). The participants answered both questions on a 5-point Likert scale, ranging from 1 ("strongly disagree") to 5 ("strongly agree").

Fig. 2.7 visualizes the distribution of the median votes for the individual tweets across all response options. We first evaluate the extent to which the participants agree with the fact-checks from Birdwatch. We find that the participants at least somewhat agree with 73.33 % of the community-created fact-checks performed by Birdwatch users. Interestingly, the agreement is lower for tweets categorized as misleading (72.00 %) than for tweets categorized as not misleading (74.67 %).

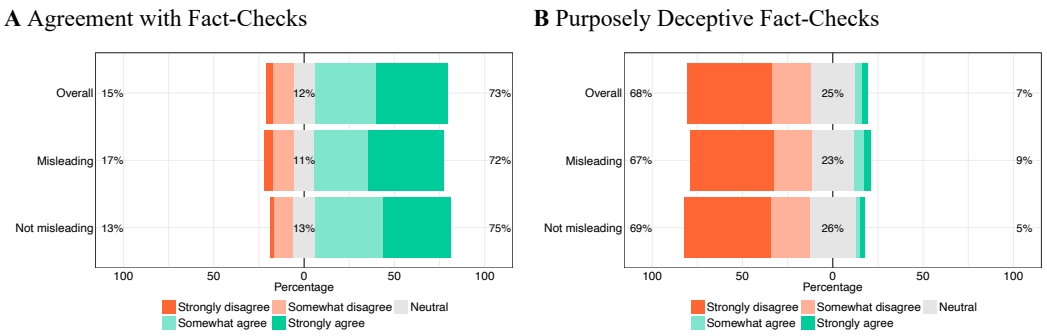


Figure 2.7: User study evaluating the perceived reliability of community-based fact-checks from Birdwatch. $n = 7$ participants were recruited via Prolific. Here we report the median responses to the questions (a) "Do you agree with the fact-checking label?" and (b) "Do you feel that the fact-check is purposely deceptive?"

We find a consistent pattern for the second question item: the median ratings of the seven participants suggest that only a relatively small share of 7.00 % of fact-checks are perceived as purposely deceptive. Notably, we again observe considerable differences across fact-checks reporting misleading vs. not misleading tweets. Specifically, fact-checks reporting misleading tweets are more likely to be perceived as being purposely deceptive (9.33 %) than fact-checks reporting not misleading tweets (4.67 %).

The participants showed statistically significant inter-rater agreements. Kendall's W was 0.43 ($p < 0.01$) for the first question item (agreement with the fact-checking label); and 0.32 ($p < 0.01$) for the second question item (purposely deceptive fact-checks).

In sum, the results of our user study suggest that the vast majority of community-created fact-checks are perceived as being reliable. This supports the results of previous experimental works, which suggest that the risk of users purposely trying to "game the system" is tolerable e. g., Allen et al., 2021. Even though inaccurate fact-checks and misuse of the platform cannot be prevented completely, community-based fact-checking should be seen as one tool (as part of a larger toolset) that may help to combat the spread of misinformation on social media (Epstein et al., 2020; Godel et al., 2021).

2.4.5 Robustness Checks

We conducted an extensive set of checks that yielded consistent findings: (1) we controlled for outliers in the dependent variables; (2) we ran separate regressions for misleading vs. not

misleading posts; (3) we calculated variance inflation factors for all independent variables and found that all remain below the critical threshold of four; (4) we repeated our analysis with user-specific random effects; (5) we incorporated quadratic effects; (6) we included interaction terms between user-specific variables and the fact-checking label; (7) we evaluated alternative approaches to handle multiple fact-checks for the same tweet (e. g., majority vote). In all of these checks, our findings are supported. Detailed results are reported in the Appendix.

2.5 Discussion

Summary of findings: This study is the first to examine the diffusion of misleading vs. not misleading posts on social media that have been fact-checked by the crowd. Our key findings are as follows: (i) community fact-checked misleading tweets receive 36.85 % fewer retweets than not misleading tweets (RQ1). (ii) There are significant differences in virality across different sub-types of misinformation (RQ2). Specifically, we find that misleading tweets are less viral than not misleading tweets across almost all sub-types of misinformation, except for (the relatively rare categories) satire, manipulated media, and outdated information. (iii) The fact-checking targets significantly differ between community fact-checkers and expert fact-checkers (RQ3). In particular, the crowd tends to fact-check posts from accounts with greater social influence (e. g., high-follower accounts).

As an additional contribution, we conducted a user study to assess the perceived reliability of (real-world) community-created fact-checks (RQ4). We find that users agree with a relatively high share (73.33 %) of community-created fact-checks, whereas only a relatively small share (7.00 %) is perceived as being purposely deceptive (e. g., due to manipulation attempts). These results corroborate previous findings of experimental studies, which suggested that crowds can achieve a high level of accuracy when fact-checking social media content e. g., Allen et al., 2021.

Research implications: In contrast to previous research examining the spread of misinformation that has been fact-checked by third-party organizations (Pröllochs et al., 2021a; Solovev & Pröllochs, 2022b; Vosoughi et al., 2018), we find that community fact-checked misleading posts receive fewer retweets than not misleading posts. The diverging results may be a consequence of differences in the sample selection. While third-party organizations tend to fact-check posts on topics experts believe are of broad public interest and/or particularly concerning to society, community fact-checked posts comprise posts that have been deemed to be worth fact-checking by actual social media users.

Our analysis suggests that crowd vs. experts focus on different targets when fact-checking social media content. We find that community fact-checkers tend to fact-check posts from larger accounts with high social influence, while expert fact-checks tend to target rumors shared by smaller accounts. Furthermore, community fact-checkers are relatively more likely to endorse/emphasize the accuracy of not misleading posts authored by influential users (i. e., users with a wide reach). This pattern is opposite to expert fact-checking where posts authored by influential accounts are relatively more likely to convey misinformation. Since author characteristics are inherently linked to the virality of posts (e. g., users with a wider reach can generate more retweets), the observed differences in fact-checking targets provide a (partial) explanation for the higher virality of not misleading community fact-checked posts. Note, however, that author characteristics are unlikely to be the only reason. In our explanatory regression analysis, we find the pattern that community fact-checked posts are more viral to

persist – even after controlling for the social influence of the author. This suggests that there might be additional differences between experts and the crowd in how fact-checking targets are selected (see *Limitations and future research*). Importantly, while our study complements earlier work studying the diffusion of expert fact-checked posts, we do not claim that the selection by the crowd is more representative for the population of misinformation on social media as a whole. Rather, our results imply that the crowd focuses on different targets when fact-checking social media content and that sample selection plays a key role when studying misinformation diffusion. Compiling a representative sample of *all* misinformation circulating on social media presents an important – yet difficult – challenge for future research.

Practical implications: From a practical perspective, policy initiatives around the world oblige social media platforms to develop countermeasures against misinformation. Community-based fact-checking opens new avenues to increase the scalability and speed of fact-checking of social media content. Furthermore, the community-based approach has the potential to to overcome trust issues associated with expert-created fact-checks (Allen et al., 2020). The observed differences in the selection of fact-checking targets between community and expert fact-checkers suggest that both approaches might complement each other well. Here, community-created fact-checking may help to identify misinformation that is actually of interest to actual social media users – and which may go unnoticed on third-party fact-checking organizations. The results of our user study further suggest that the vast majority of community-created fact-checks are perceived as being reliable. Although misuse of the platform cannot be prevented completely, previous research suggests that many issues with bad actors can effectively be addressed using sophisticated ranking mechanisms (e. g., helpfulness ratings), incentivizing high-quality fact-checks (e. g., blocking malicious contributors) or performing additional community-based content moderation efforts (Epstein et al., 2020; Godel et al., 2021). In sum, community-based fact-checking systems (as part of a larger toolset) allow social media platforms for improved coverage and may help to combat misinformation on social media more effectively.

Limitations and future research: Our work has a number of limitations, which provide promising opportunities for future research. First, similar to related studies e. g., Solovev and Pröllochs, 2022b; Vosoughi et al., 2018, we do not make causal claims. Future work should thus seek to validate our results in controlled experiments. Second, our user study evaluates the *perceived* reliability of community-based fact-checks. While earlier experimental studies have already shown that crowds can achieve a high level of accuracy when fact-checking social media content e. g., Allen et al., 2021, it is necessary to further investigate the performance of the crowd in the field (e. g., via expert assessments of Birdwatch notes). Also, more research is necessary to better understand the role of manipulation attempts, and the conditions under which the wisdom of crowds can be unlocked for fact-checking. Third, our study shows that the fact-checking targets in community vs. expert fact-checks differ in terms of their author characteristics (e. g., number of followers). Future research should complement this analysis with a fine-grained study of additional characteristics of the fact-checked posts. For instance, it is a promising extension to employ topic modeling to study how the virality varies across topics (e. g., politics, health, entertainment, etc.) and other misinformation characteristics (e. g., novelty, believability). Fourth, our results are limited to Twitter's Birdwatch pilot. As such, the restricted set of Birdwatch contributors might not be representative for the overall user base on Twitter. Fifth, in recent years, Twitter and other platforms have increased their content moderation efforts. Compared to earlier periods, this may have implications

regarding what users post on social media (e. g., due to fear of getting banned) and some particularly egregious misinformation may have been removed in line with platform guidelines. However, community-created fact-checks did not have special effects on the way people see tweets or other system recommendations during Birdwatch's pilot phase (X, 2021). Sixth, the community-created fact-checks in our study were not visible to the vast majority of Twitter users (i. e., only to pilot participants), whereas Twitter's goal is that Birdwatch will be available to everyone on Twitter. Future research may expand the current investigation by studying how (community-based) fact-checking *labels* influence users' sharing behavior on social media.

2.6 Conclusion

The spread of misinformation on social media is a pressing societal problem that platforms, policymakers, and researchers continue to grapple with. As a countermeasure, recent research proposed to build on crowd wisdom to fact-check social media content. In this study, we empirically analyzed the spread of posts that have been fact-checked by the crowd on Twitter's Birdwatch platform. Different from earlier studies that have analyzed the spread of misinformation fact-checked by third-party organizations, we find that crowd fact-checked misleading posts are less viral than not misleading posts. Our results also suggest that there are significant differences in virality across different sub-types of misinformation (e. g., factual errors, missing context, satire). Altogether, our findings offer insights into how misleading vs. not misleading posts spread and highlight the crucial role of sample selection when studying misinformation on social media.

Bibliography

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393.

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539.

Allen, J., Martel, C., & Rand, D. G. (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *CHI*.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Bär, D., Pröllochs, N., & Feuerriegel, S. (2023a). Finding qs: Profiling qanon supporters on parler. *ICWSM*.

Bär, D., Pröllochs, N., & Feuerriegel, S. (2023b). New threats to society from free-speech social media platforms. *Communications of the ACM*, *66*(10), 37–40.

Bhuiyan, M. M., Zhang, A. X., Sehat, C. M., & Mitra, T. (2020). Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, *4*, 1–26.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *PNAS*, *114*(28), 7313–7318.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *PNAS*, *113*(3), 554–559.

Dissanayake, I., Nerur, S., Singh, R., & Lee, Y. (2019). Medical crowdsourcing: Harnessing the "wisdom of the crowd" to solve medical mysteries. *Journal of the Association for Information Systems*, *20*(11), 4.

Drolsbach, C. P., & Pröllochs, N. (2023). Believability and harmfulness shape the virality of misleading social media posts. *WWW*.

Epstein, Z., Pennycook, G., & Rand, D. (2020). Will the crowd game the algorithm? using layperson judgments to combat misinformation on social media by downranking distrusted sources. *Chi*.

Frey, V., & van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, *67*(7), 4273–4286.

Friggeri, A., Adamic, L. A., Eckles, D., & Cheng, J. (2014). Rumor cascades. *ICWSM*.

Geissler, D., Bär, D., Pröllochs, N., & Feuerriegel, S. (2022). Russian propaganda on social media during the 2022 invasion of ukraine. *arXiv:2211.04154*.

Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. (2021). Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, *1*(1), 1–36.

Han, Y., Ozturk, P., & Nickerson, J. V. (2021). Leveraging the wisdom of crowd to address societal challenges: A revisit to the knowledge reuse process for innovation through analytics. *Journal of the Association for Information Systems*, *forthcoming*.

Jakubik, J., Vössing, M., Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). Online emotions during the storming of the us Capitol: Evidence from the social media network Parler. *ICWSM*.

Lazer, D., 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.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. *Icjai*.

Micallef, N., Armacost, V., Memon, N., & Patil, S. (2022). True or false: Studying the work practices of professional fact-checkers. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–44.

Micallef, N., He, B., Kumar, S., Ahamad, M., & Memon, N. (2020). The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. *International Conference on Big Data*.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly*, *37*(2), 407–426.

Oh, O., Kwon, K. H., & Rao, H. R. (2010). An exploration of social media in extreme events: Rumor theory and twitter during the Haiti earthquake 2010. *International Conference on Information Systems (icis)*.

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on wikipedia readers and readership. *Journal of the Association for Information Science and Technology*, *65*(12), 2381–2403.

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

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *PNAS*, *116*(7), 2521–2526.

Pew Research Center. (2016). News use across social media platforms 2016 [Accessed: 2022-04-02].

Poynter. (2019). Most republicans don't trust fact-checkers, and most Americans don't trust the media [Accessed: 2022-04-02].

Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. *ICWSM*.

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021a). Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports*, *11*, 22721.

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021b). Emotions in online rumor diffusion. *EPJ Data Science*, *10*(1), 51.

Pröllochs, N., & Feuerriegel, S. (2023). Mechanisms of true and false rumor sharing in social media: Collective intelligence or herd behavior? *arXiv*, (2207.03020).

Qazvinian, V., Rosengren, E., Radev, D. R., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. *Emnlp*.

Rho, E. H. R., & Mazmanian, M. (2020). Political hashtags & the lost art of democratic discourse. *CHI*.

Rinker, T. W. (2019). *sentimentr: Calculate text polarity sentiment* [version 2.7.1]. Buffalo, New York. https://help.twitter.com/en/rules-and-policies/media-policy

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature human behaviour*, *7*(5), 812–822.

Shore, J., Baek, J., & Dellarocas, C. (2018). Network structure and patterns of information diversity on twitter. *MIS Quarterly*, *42*(3), 849–972.

Solovev, K., & Pröllochs, N. (2022a). Hate speech in the political discourse on social media: Disparities across parties, gender, and ethnicity. *WWW*.

Solovev, K., & Pröllochs, N. (2022b). Moral emotions shape the virality of covid-19 misinformation on social media. *WWW*.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 Boston marathon bombing. *Iconference*.

Statista. (2022). Number of monetizable daily active twitter users (mdau) in the united states from 1st quarter 2017 to 2nd quarter 2022. https://www.statista.com/statistics/970911/monetizable-daily-active-twitter-users-in-the-united-states/

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, *29*(4), 217–248.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explorations Newsletter*, *21*(2), 80–90.

X. (2021). Introducing Birdwatch, a community-based approach to misinformation.

Appendix 2.A Regression Results Without Outliers

To assess the robustness of our analysis regarding outliers, we remove tweets with the top 1% highest values for the retweet count. The results are presented Table 2.1. All results are robust and confirm our previous findings.

Table 2.1: Regression Results Without Outliers

Dependent Variable: Number of Retweets (*RetweetCount*)			
	Source Tweet	*Fact-Checking Label*	*Misinformation Types*
	Model 1	Model 2	Model 3
Misleading		−0.281***	
		(0.067)	
Factual Error			−0.233***
			(0.033)
Missing Important Context			−0.093***
			(0.032)
Unverified Claim As Fact			−0.273***
			(0.032)
Outdated Information			0.086*
			(0.052)
Satire			0.076
			(0.075)
Manipulated Media			0.479***
			(0.078)
Other			−0.203***
			(0.069)
Delay	−0.018	−0.017	−0.028
	(0.018)	(0.018)	(0.018)
Sentiment	0.076***	0.074***	0.070***
	(0.015)	(0.015)	(0.015)
Followers	0.214***	0.212***	0.213***
	(0.020)	(0.020)	(0.020)
Followees	0.133***	0.134***	0.144***
	(0.016)	(0.016)	(0.016)
Account age	−0.237***	−0.238***	−0.238***
	(0.016)	(0.016)	(0.016)
Verified	1.063***	1.070***	1.114***
	(0.033)	(0.033)	(0.034)
Intercept	7.125***	7.354***	7.335***
	(0.087)	(0.103)	(0.090)
Fixed effects (month-year)	Yes	Yes	Yes
AIC	217 196	218 106	216 960
Observations	15 103	15 103	15 103

Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; standard errors in parentheses
Note: Negative binomial regression explains the number of retweets of the fact-checked tweet. Month-year fixed effects are included.

Appendix 2.B Separate Regressions for Misleading and Not Misleading Tweets

We run separate for regressions for the subsets of misleading and not misleading tweets. The results remain robust (see Table 2.2).

Table 2.2: Regression Results for Subsets of Misleading and Not Misleading Tweets

Dependent Variable: Number of Retweets (*RetweetCount*)		
	Subset: *Misleading*	Subset: *Not Misleading*
	Model 1	Model 2
Delay	−0.041**	−0.286***
	(0.019)	(0.073)
Sentiment	0.115***	−0.016
	(0.016)	(0.070)
Followers	0.273***	0.270***
	(0.020)	(0.057)
Followees	0.084***	0.027
	(0.017)	(0.047)
Account age	−0.224***	−0.063
	(0.017)	(0.077)
Verified	0.803***	0.504***
	(0.035)	(0.158)
Intercept	7.751***	8.416***
	(0.098)	(0.234)
Fixed effects (month-year)	Yes	Yes
AIC	209 875	13 287
Observations	14 384	872

Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; standard errors in parentheses
Note: Negative binomial regression explains the number of retweets of the fact-checked tweet. Month-year fixed effects are included.

Appendix 2.C Variance Inflation Factors

We calculated variance inflation factors for all explanatory variables in our regression models for RQ1 and RQ2 (Table 2.3). The VIFs are substantially below the critical threshold of four. This indicates that multicollinearity is not an issue in our analysis.

Table 2.3: Variance Inflation Factors for Regression Models

	RQ1	**RQ2**
Misleading	1.019	
Delay	1.004	1.008
Sentiment	1.011	1.016
Followers	1.069	1.072
Followees	1.001	1.003
Account Age	1.155	1.177
Verified	1.193	1.251
Factual Error		1.093
Missing Important Context		1.082
Unverified Claim As Fact		1.129
Outdated Information		1.045
Satire		1.049
Manipulated Media		1.053
Other		1.021

Appendix 2.D Analysis With User-Specific Random Effects

Fact-checks on Birdwatch are performed by many different contributors. To account for this, we include random effects for the individual Birdwatch contributors into our regression model. The regression results are reported in Table 2.4. All results are robust and confirm our previous findings.

Table 2.4: Regression Results With User-Specific Random Effects

Dependent Variable: Number of Retweets (*RetweetCount*)			
	Source Tweet	*Fact-Checking Label*	*Misinformation Types*
	Model 1	Model 2	Model 3
Misleading		−0.456***	
		(0.068)	
Factual Error			−0.251***
			(0.034)
Missing Important Context			−0.127***
			(0.033)
Unverified Claim As Fact			−0.300***
			(0.033)
Outdated Information			−0.061
			(0.054)
Satire			0.411***
			(0.077)
Manipulated Media			0.462***
			(0.080)
Other			−0.222***
			(0.071)
Delay		−0.048***	−0.054***
		(0.018)	(0.018)
Sentiment	0.068***	0.068***	0.061***
	(0.014)	(0.014)	(0.014)
Followers	0.271***	0.267***	0.268***
	(0.019)	(0.019)	(0.019)
Followees	0.074***	0.076***	0.086***
	(0.016)	(0.016)	(0.016)
Account age	−0.213***	−0.216***	−0.227***
	(0.017)	(0.017)	(0.017)
Verified	0.774***	0.783***	0.866***
	(0.034)	(0.034)	(0.035)
Intercept	7.922***	8.266***	8.122***
	(0.089)	(0.105)	(0.092)
Fixed effects (month-year)	Yes	Yes	Yes
Random effects (user)	Yes	Yes	Yes
AIC	223 233	223 182	222 893
Observations	15 256	15 256	15 256

Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; standard errors in parentheses
Note: Negative binomial regression explains the number of retweets of the fact-checked tweet. Month-year fixed effects are included.

Appendix 2.E Quadratic Effects and Interaction Terms

As a robustness check, we include quadratic effects and interaction terms between the fact-checking label and the source tweet variables into our regression analysis. The results remain robust and support our findings (see Table 2.5).

Table 2.5: Regression Results With Quadratic Effects and Interaction Terms

Dependent Variable: Number of Retweets (*RetweetCount*)		
	Quadratic Effects	*Interaction Terms*
	Model 1	Model 2
Misleading	−0.479***	−0.659***
	(0.068)	(0.106)
Delay	0.086**	−0.242***
	(0.042)	(0.070)
Delay^2	−0.010***	
	(0.003)	
Sentiment	0.136***	0.020
	(0.016)	(0.066)
Sentiment^2	0.011	
	(0.007)	
Followers	0.548***	0.249***
	(0.045)	(0.054)
Followers^2	−0.029***	
	(0.006)	
Followees	0.115***	0.025
	(0.024)	(0.045)
Followees^2	−0.003*	
	(0.002)	
Account age	−0.416***	−0.058
	(0.022)	(0.073)
Account age^2	−0.341***	
	(0.021)	
Verified	0.739***	0.389***
	(0.035)	(0.150)
Misleading × Delay		0.201***
		(0.072)
Misleading × Sentiment		0.096
		(0.068)
Misleading × Followers		0.024
		(0.058)
Misleading × Followees		0.059
		(0.048)
Misleading × Account age		−0.166**
		(0.075)
Misleading × Verified		0.416***
		(0.155)
Intercept	8.641***	8.436***
	(0.107)	(0.130)
Fixed effects (month-year)	Yes	Yes
AIC	222 906	223 179
Observations	15 256	15 256

Significance levels: $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; standard errors in parentheses
Note: Negative binomial regression explains the number of retweets of the fact-checked tweet. Month-year fixed effects are included.

Appendix 2.F Alternative Handling of Multiple Fact-Checks

Our main analysis focuses on the temporally first fact-check after the tweet has been posted. As a robustness check, we evaluate whether our results are robust to alternative handling of multiple fact-checks. We repeated our analysis with the following variants: (i) we determined the fact-checking label via majority vote; (ii) we use Birdwatch's rating mechanism (see **Twitter.2021** for details) to identify the fact-check with which most users agree; (iii) we consider all fact-checks without any filtering.

The regression results are presented in Table 2.6. In all cases, we find qualitatively identical results that support our previous findings.

Table 2.6: Regression Results With Alternative Handling of Multiple Fact-Checks

Dependent Variable: Number of Retweets (*RetweetCount*)						
	(i) Majority Vote		*(ii) Highest Agreement*		*(iii) All Fact-Checks*	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Misleading		−0.630***		−0.728***		−0.849***
		(0.068)		(0.057)		(0.043)
Delay		0.018		0.011		−0.004
		(0.017)		(0.018)		(0.013)
Sentiment	0.115***	0.112***	0.112***	0.109***	0.046***	0.064***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.013)	(0.013)
Followees	0.087***	0.092***	0.074***	0.078***	0.021	0.033**
	(0.017)	(0.017)	(0.016)	(0.016)	(0.013)	(0.013)
Followers	0.284***	0.273***	0.271***	0.257***	0.319***	0.316***
	(0.020)	(0.020)	(0.019)	(0.019)	(0.014)	(0.014)
Account age	−0.223***	−0.223***	−0.213***	−0.214***	−0.190***	−0.192***
	(0.017)	(0.017)	(0.017)	(0.017)	(0.014)	(0.014)
Verified	0.789***	0.796***	0.774***	0.763***	0.717***	0.731***
	(0.035)	(0.035)	(0.034)	(0.034)	(0.029)	(0.029)
(Intercept)	7.879***	8.375***	7.922***	8.477***	8.821***	9.559***
	(0.091)	(0.107)	(0.089)	(0.100)	(0.070)	(0.079)
Fixed effects (month-year)	Yes	Yes	Yes	Yes	Yes	Yes
AIC	211 311	211 213	223 233	223 039	317 871	317 409
Observations	14 619	14 619	15 256	15 256	20 218	20 218

Significance levels: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$; standard errors in parentheses

Note: Negative binomial regression explains the number of retweets of the fact-checked tweet. Month-year fixed effects are included.

Chapter 3

Believability and Harmfulness Shape the Virality of Misleading Social Media Posts

Abstract

Misinformation on social media presents a major threat to modern societies. While previous research has analyzed the virality across true and false social media posts, not every misleading post is necessarily equally viral. Rather, misinformation has different characteristics and varies in terms of its believability and harmfulness – which might influence its spread. In this work, we study how the perceived believability and harmfulness of misleading posts are associated with their virality on social media. Specifically, we analyze (and validate) a large sample of crowd-annotated social media posts from Twitter's Birdwatch platform, on which users can rate the believability and harmfulness of misleading tweets. To address our research questions, we implement an explanatory regression model and link the crowd ratings for believability and harmfulness to the virality of misleading posts on Twitter. Our findings imply that misinformation that is (i) easily believable and (ii) not particularly harmful is associated with more viral resharing cascades. These results offer insights into how different kinds of crowd fact-checked misinformation spreads and suggest that the most viral misleading posts are often not the ones that are particularly concerning from the perspective of public safety. From a practical view, our findings may help platforms to develop more effective strategies to curb the proliferation of misleading posts on social media.

Keywords: Social media, misinformation, virality, community fact-checking, computational social science, explanatory modeling

3.1 Introduction

Social media disseminates vast amounts of misinformation (e. g., Ecker et al., 2022; Shao et al., 2016; Vosoughi et al., 2018). Several works have studied the diffusion of rumors of varying veracity, finding that misinformation spreads more virally than the truth (Bessi et al., 2015; Friggeri et al., 2014; Pröllochs et al., 2021a; Solovev & Pröllochs, 2022; Vosoughi et al., 2018). If misinformation becomes viral, it can have detrimental real-world consequences and affects how opinions are formed (Allcott & Gentzkow, 2017; Bakshy et al., 2015; Del Vicario et al., 2016; Oh et al., 2013). This has been observed, for example, during elections (e. g., Allcott & Gentzkow, 2017; Aral & Eckles, 2019; Bakshy et al., 2015; Grinberg et al., 2019) and crisis situations (e. g., Broniatowski et al., 2018; Geissler et al., 2022; Oh et al., 2010, 2013; Pennycook et al., 2020; Solovev & Pröllochs, 2022; Starbird et al., 2014; Zeng et al., 2016). As such, misinformation on social media threatens the well-being of society at large and demands effective countermeasures (Bär et al., 2023; Lazer et al., 2018; Pennycook et al., 2021).

While earlier research has analyzed differences in the spread of true and false social media posts (Pröllochs et al., 2021a; Solovev & Pröllochs, 2022; Vosoughi et al., 2018), not every misinforming post is necessarily equally viral. Rather, misinformation has different characteristics and varies in terms of its believability and harmfulness – which might influence its spread. For example, individuals using social media tend to be in a hedonic mindset and thus are looking for entertainment and fun (Kim & Dennis, 2019; Lutz et al., 2020). Thus, if a user does not believe the content of a post, there might be less incentive to share it and increase its reach. In a similar vein, research in psychology suggests that threats capture attention (Koster et al., 2004; Schmidt et al., 2015). Contextualized to misinformation on social media, this would imply that harmful misleading posts are detected more accurately – and, therefore, less likely to be shared. Overall, one may expect that the believability and harmfulness of misinformation play a crucial role in its spread. However, there is currently no study empirically analyzing the link between these attributes and virality on social media.

Research goal: We analyze the link between the believability and harmfulness of misleading posts and their virality on social media. In particular, we seek to answer two research questions:

- **(RQ1)** *Are misleading posts perceived as believable more viral than those perceived as not believable?*
- **(RQ2)** *Are misleading posts perceived as harmful more viral than those perceived as not harmful?*

Data & methods: We draw upon a large dataset of crowd-annotated tweets from Twitter's fact-checking system "Birdwatch" (Pröllochs, 2022). On Birdwatch, users can create "Birdwatch notes" that aim to identify misleading tweets directly on Twitter. A unique feature of fact-checking on Birdwatch is that users also categorize whether they perceive misleading tweets to be easily believable and/or harmful. For our analysis, we collect (and validate) Birdwatch notes for misleading tweets between the launch of Birdwatch in early 2021 and the end of February 2022. Subsequently, we perform an explanatory regression analysis and link the believability and harmfulness (as provided in Birdwatch notes) to the number of retweets (as a measure of virality) of the fact-checked post. In our analysis, we control for established predictors that may affect the retweet rate (e. g., social influence, sentiment). This approach

allows us to empirically test how the believability and harmfulness of misleading posts are associated with their virality on social media.

Contributions: Our study offers insights into how crowd fact-checked misinformation spreads on social media. Specifically, we demonstrate that misinformation that is (i) easily believable and (ii) not particularly harmful is associated with more viral resharing cascades. These findings imply that not all kinds of misinformation are equally viral; and that the most viral misleading posts are oftentimes not the ones that are particularly concerning from the perspective of public safety. In a next step, our findings may help platforms to implement more effective strategies for reducing the proliferation of misinformation.

3.2 Background

Community-based fact-checking: The concept of community-based fact-checking is a relatively novel approach that aims to tackle misinformation on social media by harnessing the "wisdom of crowds" (Frey & van de Rijt, 2021; Woolley et al., 2010). Specifically, the idea is to let regular social media users carry out fact-checking of social media posts (Allen et al., 2020, 2021; Bhuiyan et al., 2020; Epstein et al., 2020; Godel et al., 2021; Micallef et al., 2020; Pennycook & Rand, 2019). Compared to expert-based approaches to fact-checking (e.g., via third-party fact-checking organizations), community-based fact-checking is appealing as it allows for large numbers of fact-checks to be frequently and inexpensively acquired (Allen et al., 2021; Woolley et al., 2010). Moreover, it addresses the issue that many users do not trust the assessments of professional fact-checkers (e. g., due to alleged political biases) (Poynter, 2019). Experimental studies suggest that the crowd can be highly accurate in identifying misinformation and even relatively small crowds can yield performance similar to experts (Bhuiyan et al., 2020; Epstein et al., 2020; Pennycook & Rand, 2019).

Birdwatch: Informed by experimental studies, the social media platform Twitter has recently launched its community-based fact-checking system Birdwatch (Pröllochs, 2022; X, 2021). Different from earlier crowd-based fact-checking initiatives (Bakabar, 2018; Bhuiyan et al., 2020; Florin, 2010; O'Riordan et al., 2019), Birdwatch allows users to identify misinformation *directly* on the platform (see next section for details). Given the recency of the platform, research on Birdwatch is scant. Early works suggest that politically motivated reasoning might pose challenges in community-based fact-checking (Allen et al., 2022; Pröllochs, 2022). Notwithstanding, community-created fact-checks on Birdwatch have been found to be perceived as informative and helpful by the vast majority of social media users (Pröllochs, 2022). Furthermore, real-world community fact-checks have been shown to be effective in reducing users' propensity to reshare misinformation (Wojcik et al., 2022).

Virality of misinformation: Several works have analyzed the spread of social media posts for which veracity was determined based on the assessment of third-party fact-checking organizations (Friggeri et al., 2014; Pröllochs & Feuerriegel, 2023; Pröllochs et al., 2021a; Solovev & Pröllochs, 2022; Vosoughi et al., 2018). For instance, Friggeri et al. (2014) analyzed upload and deletion rates in $\approx 4\,000$ expert fact-checked rumors from Facebook. Another literature stream has analyzed the diffusion of true vs. false rumors on Twitter (Pröllochs & Feuerriegel, 2023; Pröllochs et al., 2021a; Solovev & Pröllochs, 2022; Vosoughi et al., 2018). The rumors (and their veracity) in these works were identified based on the presence of user comments referencing fact-checks carried out by third-party fact-checking organizations (see,

e. g., Vosoughi et al., 2018 for methodological details). These studies typically observed that false social media posts spread more viral than true posts.

Research gap: Existing research has primarily focused on studying the virality across true vs. false social media posts that have been fact-checked by expert fact-checkers. However, an understanding of how the virality of misinformation varies depending on its underlying characteristics is largely absent. Specifically, we are not aware of previous work empirically analyzing how the perceived believability and harmfulness of misleading posts are associated with their virality on social media. This presents our contribution.

3.3 Data and Methodology

3.3.1 Data Collection

To answer our research questions, we analyze a large dataset of crowd-annotated tweets that have been identified as being misleading during the pilot phase of Twitter's Birdwatch platform (Pröllochs, 2022; X, 2021). Birdwatch has been launched by Twitter on January 23, 2021, and aims to identify misleading social media posts by harnessing the wisdom of crowds. Different from earlier small-scale crowd-based initiatives to fact-checking (Bakabar, 2018; Bhuiyan et al., 2020; Florin, 2010; O'Riordan et al., 2019), Birdwatch allows users to identify misleading tweets *directly* on Twitter and write short (max 280 characters) fact-checks (so-called "Birdwatch notes") that add context to the tweet. Another unique feature of Birdwatch is that authors of Birdwatch notes additionally need to answer checkbox questions when identifying misleading posts. Here users can rate whether they perceive the misleading tweet to be easily believable and whether the tweet might cause considerable harm.

To participate in the pilot phase of the Birdwatch feature (only available in the US), Twitter users had to register and apply to become a contributor. In early 2022, Birdwatch had approximately 3 250 contributors, which is a relatively small fraction of all Twitter users (≈41.5 million daily active users Statista, 2022). Birdwatch notes were displayed directly on tweets to pilot participants (see example in Fig. 3.1); while all other Twitter users could view them on a separate Birdwatch website (birdwatch.twitter.com). Accordingly, the fact-checks were not directly visible to the vast majority of Twitter users. Birdwatch notes were thus unlikely to influence the diffusion of the fact-checked tweets during our study period.

For our analysis, we downloaded all Birdwatch notes between the launch of Birdwatch on January 23, 2021, and the end of February 2022 from the Birdwatch website[1], i. e., for an observation period of more than one year. The dataset contains a total number of 20 218 Birdwatch notes from 3 257 different contributors.

On Birdwatch, multiple users can write Birdwatch notes for the same tweet. Therefore, the data sometimes includes multiple Birdwatch notes for the same post (≈ 1.24 notes per tweet). As a result, different Birdwatch users might disagree on the characteristics of one tweet. To incorporate this, we used majority vote to determine the categorizations. We excluded tweets without a definite assessment (i. e., if two assessments stand in opposition) and tweets classified as not misleading.[2] This filtering step resulted in a dataset consisting of 13 732 tweets. Each

[1] Available via https://twitter.com/i/communitynotes/download-data.

[2] Birdwatch contributors can also endorse the accuracy of *not* misleading tweets (5.72 % of all Birdwatch notes). Since users cannot rate the believability and harmfulness of these tweets, Birdwatch notes for not misleading tweets are excluded from our analysis.

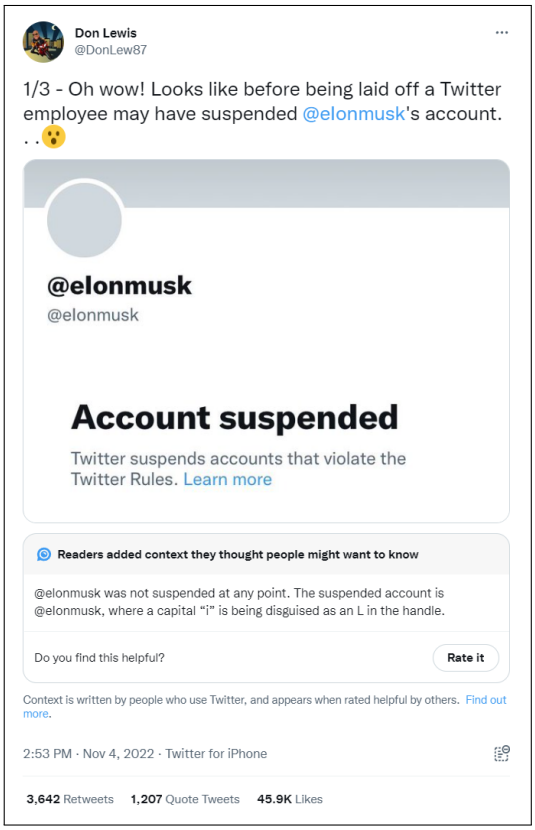


Figure 3.1: Example of a Birdwatch note identifying a misleading post on Twitter.

of the fact-checks addresses a single *misleading* tweet for which the Birdwatch contributor has assessed the believability and harmfulness.

We further mapped the *tweetID* referenced in each Birdwatch note to the underlying source tweets using the Twitter historical API. This allowed us to collect additional information concerning the fact-checked tweets and its author, namely, (a) the number of retweets, (b) the followers count, (c) the followees count, (d) the account age, and (e) whether the user has been verified by Twitter. Moreover, we calculated a sentiment score for each source tweet to control for its positivity/negativity in our later empirical analysis.[3]

3.3.2 Explanatory Regression Model

We specify an explanatory regression model that explains the virality of misleading tweets based on their believability and harmfulness. In our analysis, we use a common proxy for the virality of a resharing cascade, namely, the number of retweets (Han et al., 2020; Solovev & Pröllochs, 2022). Since the variance of the retweet count is larger than its mean, we have to adjust for overdispersion. Analogous to earlier research (e. g., Solovev & Pröllochs, 2022; Stieglitz & Dang-Xuan, 2013), we thus employ a negative binomial regression model.

[3] Analogous to prior work (e. g., Jakubik et al., 2023; Pröllochs et al., 2021a; Robertson et al., 2023), we use the NRC dictionary (Mohammad & Turney, 2013) to calculate a sentiment score measuring the share of positive vs. negative words. Here, we use the default implementation for sentiment analysis provided in the `sentimentr` R package.

Formally, the response variable in our negative binomial regression model is *RetweetCount*$_i$, which refers to the number of retweets received by tweet i. The key regressors are binary and indicate whether the tweet has been rated as believable (*Believable*$_i$, $= 1$ if true, otherwise 0) and harmful (*Harmful*$_i$, $= 1$ if true, otherwise 0) on Birdwatch. Concordant with earlier work (e. g., Solovev & Pröllochs, 2022; Stieglitz & Dang-Xuan, 2013; Vosoughi et al., 2018), we control for the social influence of the author of the source tweet (e. g., some authors have many followers and reach wider audiences). The control variables comprise the followers and followees count, the account age (in years), and the verification status. Furthermore, we control for the positivity/negativity ($Sentiment$) of the fact-checked tweet. This yields the model

with intercept β_0. Furthermore, we include month-year fixed effects u_i, which allow us to control for varying start dates and the age of the resharing cascades (e. g., Solovev & Pröllochs, 2022). In our regression analysis, all continuous variables are z-standardized to facilitate interpretability.

3.4 Empirical Analysis

3.4.1 Summary Statistics

We start our analysis by evaluating summary statistics. Out of all tweets, 94.20 % are rated as *Believable* and 74.60 % as *Harmful*. In total, the tweets have received 26.81 million retweets. On average, each tweet in our dataset has received 1 724 retweets. However, the number of retweets is higher for tweets perceived as believable. Specifically, the average number of retweets is 1 772 for believable tweets and 751 for not believable tweets. We further observe that tweets rated as harmful receive fewer retweets (1 607) than tweets rated as not harmful (1 832). Complementary cumulative distribution functions for the retweet count are shown in Fig. 3.2. The differences in the distributions are statistically significant according to two-tailed Kolmogorov-Smirnov (KS) tests ($p < 0.01$). Additionally, we calculated the correlation between the variables *Believable* and *Harmful*. Here we find a weak positive correlation of 0.181 ($p < 0.01$). This indicates that harmful posts can be but are not necessarily believable (and vice versa).

Note that the tweets in our dataset show substantial heterogeneity regarding the characteristics of the source accounts. On average, the authors of the tweets have 1.39 million followers (SD: 5.88 million), 5 795 followees (SD: 20 094), and an account age of 8.89 years (SD: 4.46). A total share of 47.90 % of all authors have been verified by Twitter (SD: 0.50). The mean sentiment of the tweets in our dataset is −0.005, i. e., slightly negative (SD: 0.26). To accommodate these potentially confounding factors, we estimate an explanatory regression model with control variables in the next section.

3.4.2 Regression Analysis

Coefficient estimates: We estimate a negative binomial regression to study the role of believability and harmfulness in the virality of misleading posts after controlling for confounding effects (e. g., varying social influence). Fig. 3.3 reports the coefficient estimates and 99% CIs. The dependent variable is the retweet count of the misleading tweet. We find that the coefficient for *Believable* is positive and statistically significant (coef: 1.154; $p < 0.01$). This estimate implies that misleading posts perceived as believable receive $e^{1.154} - 1 \approx 217.09$ % more

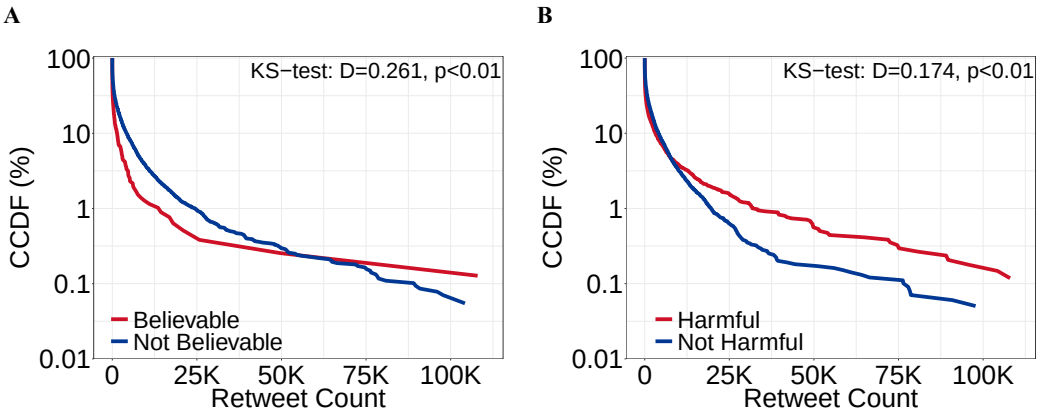


Figure 3.2: Complementary cumulative distribution functions showing the distribution of the retweet count separated by (a) believability and (b) harmfulness.

retweets. We further observe a negative and statistically significant coefficient for *Harmful* (coef: -0.533; $p < 0.01$). This implies that misleading posts perceived as harmful receive 41.32 % fewer retweets. In sum, we find that misinformation that is (i) easily believable and (ii) not particularly harmful is associated with more viral resharing cascades.

Interaction effect: Misleading posts can be categorized as (i) believable or harmful, (ii) believable and harmful, or (iii) neither believable nor harmful. To test whether the effects of different combinations of believability and harmfulness on virality differ, we reestimated our regression model with an interaction term between *Believable* $\times$ *Harmful* (see Fig 3.3). We observe that the coefficient of the interaction term is not statistically significant (coef: 0.038; $p = 0.796$). At the same time, the coefficients of *Believable* and *Harmful* remain stable. This suggests that the predictors' effects are additive and do not depend on each other.

Control variables: We also observe statistically significant coefficient estimates for the control variables in our regression analysis. Specifically, more retweets occur for source tweets authored by accounts with higher numbers of followers (coef: 0.303; $p < 0.01$) and followees (coef: 0.107; $p < 0.01$). Furthermore, more retweets are estimated for tweets from accounts that are younger in age (coef: -0.233; $p < 0.01$) and users with a verified status (coef: 0.953; $p < 0.01$). Analogous to prior work (e. g., Pröllochs et al., 2021a, 2021b), we also observe that resharing cascades are larger if they convey a more positive sentiment (coef: 0.098; $p < 0.01$).

Robustness checks: We carried out multiple checks that confirmed the robustness of our results. First, we checked our models for multicollinearity and ensured that the VIFs are below four. Second, we reestimated our models with a random-effects specification controlling for heteregoneity across the contributors on Birdwatch (i. e., user-specific effects). Third, we used alternative methods for handling multiple Birdwatch notes for the same source tweets (e. g., via Birdwatch's rating mechanism; see Pröllochs, 2022; X, 2021). In each of these checks, we found support for our findings.

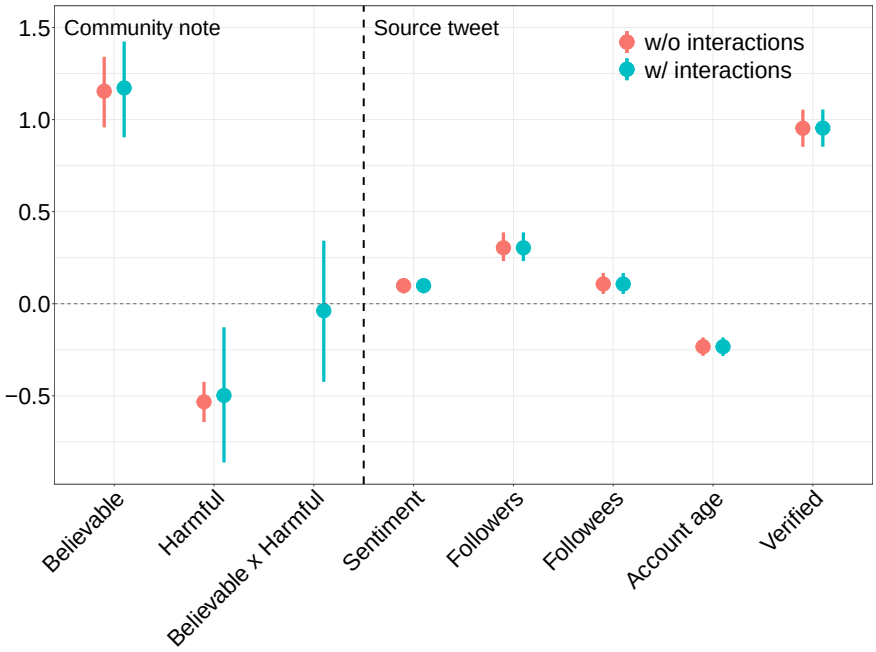


Figure 3.3: Negative binomial regression linking perceived believability and harmfulness to the number of retweets. Reported are models w/o (coral) and w/ (turquoise) an interaction term between believability and harmfulness. The circles show standardized coefficient estimates and the error bars indicate the 99 % CIs. Month-year fixed effects are included.

3.4.3 Validation Study

To validate the categorizations on Birdwatch, we carried out a user study with $n = 7$ participants via Prolific (www.prolific.com). All participants were English native speakers and based in the US Furthermore, six out of seven participants stated that they regularly use social media to share content. We asked the participants to rate the believability and harmfulness of 150 misleading tweets from Birdwatch on a 5-point Likert scale. The participants rated tweets categorized as believable by Birdwatch users as significantly more believable than tweets not categorized as believable ($M_{\text{Believable/Believable}} = 3.61$, $M_{\text{Believable/NotBelievable}} = 3.25$, $t = 3.03$, $p < 0.01$). Furthermore, tweets categorized as harmful by Birdwatch users were rated as significantly more harmful than misleading tweets not categorized as harmful ($M_{\text{Harmful/Harmful}} = 3.50$, $M_{\text{Harmful/NotHarmful}} = 3.02$, $t = 5.30$, $p < 0.01$). The inter-rater agreement was statistically significant for both believability ($W = 0.27$, $p < 0.01$) and harmfulness ($W = 0.43$, $p < 0.01$). These findings add to the validity of our results and confirm that the perceptions of independent annotators (that may have varying familiarity with the tweets' information) and the categorizations of (self-selected) Birdwatch users point in the same direction.

3.5 Discussion

Research implications: We contribute to research into misinformation by studying the link between specific attributes of misleading posts and their virality on social media. Specifically, we hypothesized that the virality of misleading posts differs depending on the perceived (i) believability and (ii) harmfulness. Our results suggest that misleading posts that are easily believable are more viral. From a theoretical perspective, a possible explanation lies in the hedonic mindset of social media users: if a user does not believe the content of a post,

increasing its reach might be less enjoyable (e. g., Johnson & Kaye, 2015; Kim & Dennis, 2019; Minas et al., 2014; Moravec et al., 2019). We further found that misleading posts perceived as harmful are less viral than those perceived as not harmful. This finding is concordant with research in psychology (e. g., Koster et al., 2004; Schmidt et al., 2015; Van Damme et al., 2008), suggesting that humans are more attentive if confronted with potentially harmful information. As a result, harmful misinformation might be detected more accurately and, therefore, less likely to be shared. Altogether, our work provides novel insights into how community fact-checked posts spread in a real-world environment and demonstrates that not all kinds of misinformation are equally viral. While previous research (e. g., Pröllochs et al., 2021a; Solovev & Pröllochs, 2022; Vosoughi et al., 2018) has analyzed differences in the spread of rumors of varying veracity, this study is the first to empirically study how the perceived believability and harmfulness of misleading posts are linked to their virality on social media.

Practical implications: Our findings are relevant for the design of more sophisticated strategies to counter misinformation. Community-based fact-checking has the potential to partially overcome the drawbacks of the experts' approach to fact-checking, e. g., in terms of speed, volume, and trust (Pennycook & Rand, 2019). Our observation that viral misleading posts tend to be easily believable and not particularly harmful implies that the most viral community fact-checked misinformation is often not particularly concerning from the perspective of public safety. In practice, this knowledge could be used by platforms to enhance the prioritization of posts for expert fact-checking. Our findings may also be relevant with regard to educational applications and for enhancing the accuracy of machine learning models for automatically detecting misleading posts.

Limitations and future work: As with others, our study is not free of limitations and offers potential for future work. First, analogous to earlier observational studies (e. g., Pröllochs et al., 2021a; Solovev & Pröllochs, 2022, 2023; Vosoughi et al., 2018), we demonstrate associations and not causal paths. Second, experimental studies in controlled settings may help to understand whether the perceptions regarding the believability and harmfulness of misinformation differ between community fact-checkers, experts, and regular social media users. Third, the restricted set of community fact-checked posts on Birdwatch may not reflect the overall population of misleading posts on social media. Thus, more research is necessary to better understand how the crowd selects posts for fact-checking (Drolsbach & Pröllochs, 2023). For instance, it would be interesting to understand whether Birdwatch users are more likely to fact-check tweets that are easier to judge in terms of their believability and harmfulness. Fourth, our analysis is limited to the social media platform Twitter and data from the Birdwatch pilot phase. In the future, community-based fact-checking on Twitter may evolve to a different steady-state due to a growing/more experienced user base and changes in functionality (e. g., Twitter recently rebranded Birdwatch to "Community Notes" X, 2021). Fifth, future work may analyze whether the observed spreading patterns are generalizable to posts from other fact-checking systems and social media platforms.

Bibliography

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393.

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539.

Allen, J., Martel, C., & Rand, D. G. (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *CHI*.

Aral, S., & Eckles, D. (2019). Protecting elections from social media manipulation. *Science*, *365*(6456), 858–861.

Bakabar, M. (2018). Crowdsourced factchecking [Accessed: 2022-04-02].

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). New threats to society from free-speech social media platforms. *Communications of the ACM*, *66*(10), 37–40.

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.

Bhuiyan, M. M., Zhang, A. X., Sehat, C. M., & Mitra, T. (2020). Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, *4*, 1–26.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *PNAS*, *113*(3), 554–559.

Drolsbach, C. P., & Pröllochs, N. (2023). Diffusion of community fact-checked misinformation on Twitter. *CSCW*.

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

Epstein, Z., Pennycook, G., & Rand, D. (2020). Will the crowd game the algorithm? using layperson judgments to combat misinformation on social media by downranking distrusted sources. *Chi*.

Florin, F. (2010). Crowdsourced fact-checking? what we learned from truthsquad [Accessed: 2022-04-02].

Frey, V., & van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, *67*(7), 4273–4286.

Friggeri, A., Adamic, L. A., Eckles, D., & Cheng, J. (2014). Rumor cascades. *ICWSM*.

Geissler, D., Bär, D., Pröllochs, N., & Feuerriegel, S. (2022). Russian propaganda on social media during the 2022 invasion of ukraine. *arXiv:2211.04154*.

Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. (2021). Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, *1*(1), 1–36.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378.

Han, Y., Lappas, T., & Sabnis, G. (2020). The importance of interactions between content characteristics and creator characteristics for studying virality in social media. *Information Systems Research*, *forthcoming*.

Jakubik, J., Vössing, M., Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). Online emotions during the storming of the us Capitol: Evidence from the social media network Parler. *ICWSM*.

Johnson, T. J., & Kaye, B. K. (2015). Reasons to believe: Influence of credibility on motivations for using social networks. *Computers in Human Behavior*, *50*, 544–555.

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.

Koster, E. H. W., Crombez, G., Van Damme, S., Verschuere, B., & De Houwer, J. (2004). Does imminent threat capture and hold attention? *Emotion*, *4*(3), 312–317.

Lazer, D., 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.

Lutz, B., Adam, M. T. P., Feuerriegel, S., Pröllochs, N., & Neumann, D. (2020). Affective information processing of fake news: Evidence from neurois. Springer.

Micallef, N., He, B., Kumar, S., Ahamad, M., & Memon, N. (2020). The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. *International Conference on Big Data*.

Minas, R. K., Potter, R. F., Dennis, A. R., Bartelt, V., & Bae, S. (2014). Putting on the thinking cap: Using neurois to understand information processing biases in virtual teams. *Journal of Management Information Systems*, *30*(4), 49–82.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Moravec, P. L., Minas, R. K., & Dennis, A. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly*, *43*(4), 1343–1360.

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly*, *37*(2), 407–426.

Oh, O., Kwon, K. H., & Rao, H. R. (2010). An exploration of social media in extreme events: Rumor theory and twitter during the Haiti earthquake 2010. *International Conference on Information Systems (icis)*.

O'Riordan, S., Kiely, G., Emerson, B., & Feller, J. (2019). Do you have a source for that? understanding the challenges of collaborative evidence-based journalism. *Opensym*.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *PNAS*, *116*(7), 2521–2526.

Poynter. (2019). Most republicans don't trust fact-checkers, and most Americans don't trust the media [Accessed: 2022-04-02].

Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. *ICWSM*.

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021a). Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific Reports*, *11*, 22721.

Pröllochs, N., Bär, D., & Feuerriegel, S. (2021b). Emotions in online rumor diffusion. *EPJ Data Science*, *10*(1), 51.

Pröllochs, N., & Feuerriegel, S. (2023). Mechanisms of true and false rumor sharing in social media: Collective intelligence or herd behavior? *arXiv*, (2207.03020).

Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature human behaviour*, *7*(5), 812–822.

Schmidt, L. J., Belopolsky, A. V., & Theeuwes, J. (2015). Attentional capture by signals of threat. *Cognition and Emotion*, *29*(4), 687–694.

Shao, C., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2016). Hoaxy: A platform for tracking online misinformation. *WWW Companion*.

Solovev, K., & Pröllochs, N. (2022). Moral emotions shape the virality of covid-19 misinformation on social media. *WWW*.

Solovev, K., & Pröllochs, N. (2023). Moralized language predicts hate speech on social media. *PNAS Nexus*, *2*(1), pgac281.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 Boston marathon bombing. *Iconference*.

Statista. (2022). Number of monetizable daily active twitter users (mdau) in the united states from 1st quarter 2017 to 2nd quarter 2022. https://www.statista.com/statistics/970911/monetizable-daily-active-twitter-users-in-the-united-states/

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, *29*(4), 217–248.

Van Damme, S., Crombez, G., & Notebaert, L. (2008). Attentional bias to threat: A perceptual accuracy approach. *Emotion*, *8*(6), 820–827.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M., Coleman, K., & Baxter, J. (2022). Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv*. https://arxiv.org/abs/2210.15723

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

X. (2021). Introducing Birdwatch, a community-based approach to misinformation.

Zeng, L., Starbird, K., & Spiro, E. S. (2016). Rumors at the speed of light? modeling the rate of rumor transmission during crisis. *Hawaii International Conference on System Sciences (hicss)*.

Chapter 4

Community notes increase trust in fact-checking on social media

Abstract

Community-based fact-checking is a promising approach to fact-check social media content at scale. However, an understanding of whether users trust community fact-checks is missing. Here, we presented n= 1810 Americans with 36 misleading and not misleading social media posts and assessed their trust in different types of fact-checking interventions. Participants were randomly assigned to treatments where misleading content was either accompanied by simple (i.e., context-free) misinformation flags in different formats (expert flags or community flags), or by textual "community notes" explaining why the fact-checked post was misleading Across both sides of the political spectrum, community notes were perceived as significantly more trustworthy than simple misinformation flags. Our results further suggest that the higher trustworthiness primarily stemmed from the context provided in community notes (i.e., fact-checking explanations) rather than generally higher trust towards community fact-checkers. Community notes also improved misinformation discernment, but did not consistently reduce sharing intentions for misleading posts beyond what was achieved with simple misinformation flags. Our work implies that context matters in fact-checking and that community notes might be an effective approach to mitigate trust issues with simple misinformation flags.

Keywords: misinformation, fact-checking, trust, social media

4.1 Main

Concerns about misinformation on social media have been rising in recent years, particularly given its potential impact on elections (Allcott & Gentzkow, 2017; Aral & Eckles, 2019; Bakshy et al., 2015; Grinberg et al., 2019; A. M. Guess et al., 2020; Moore et al., 2023), public health (Broniatowski et al., 2018; Gallotti et al., 2020; Rocha et al., 2021; Roozenbeek et al., 2020), and public safety (Bär et al., 2023; Oh et al., 2013; Starbird, 2017). Major social media providers such as X (formerly Twitter) and Facebook have thus been called upon to develop effective countermeasures to combat the spread of misinformation on their platforms (Calo et al., 2021; Donovan, 2020; Feuerriegel et al., 2023; Kozyreva et al., 2024; Lazer et al., 2018). To this end, a widely implemented approach is the use of professional fact-checkers to identify and label misleading posts (Instagram, 2019; Mosseri, 2016). The rationale is that if users are warned that a message is false, they should be less likely to believe it. Previous work has evaluated the effects of flags and warning labels on misinformation discernment and sharing intentions. Although some contradictory evidence exists (Moravec et al., 2019), there is a growing consensus that it can be effective to put misinformation flags on content contested by fact-checkers (Altay et al., 2023; Clayton et al., 2020; Kim & Dennis, 2019; Martel & Rand, 2023a; Mena, 2020; Moravec et al., 2020; Ng et al., 2021; Pennycook, Bear, et al., 2020; Pennycook, McPhetres, et al., 2020; Porter & Wood, 2021; Yaqub et al., 2020); for a review, see (Martel & Rand, 2023a).

However, current approaches to fact-checking social media content have several drawbacks that limit their full potential. First, due to the limited amount of fact-checks that experts can perform, they are unable to accommodate the amount and speed of content creation on social media (Martel et al., 2024; Pennycook & Rand, 2019). A large proportion of misinformation on social media thus goes unchecked. Second, a large proportion of Americans have concerns regarding the independence of the experts' assessment (Poynter, 2019; Straub & Spradling, 2022), and the stance on fact-checking is increasingly becoming a partisan issue. Adherents of the left tend to be less tolerable of the spread of misinformation and have greater trust in fact-checking (González-Bailón et al., 2022; Martel & Rand, 2023b; Nyhan & Reifler, 2015; Shin & Thorson, 2017). According to surveys, a majority of Republican partisans (70 %) and half of all U.S. adults believe that fact-checkers are biased and that their corrections cannot be trusted (Poynter, 2019). While research indicates that fact-checks are broadly effective (Martel & Rand, 2023b; Pan et al., 2022; Yaqub et al., 2020), their influence on users is smaller – although still significant – for those who distrust fact-checkers (Martel & Rand, 2023b). Hence, even when a social media post gets fact-checked and flagged, the impact on users may be limited due to a lack of trust (Brandtzaeg & Følstad, 2017).

Given these problems regarding scalability and trust, recent works have proposed outsourcing fact-checking of social media content to the actual social media users, i. e., non-expert fact-checkers in the crowd (Allen et al., 2020, 2021; Bhuiyan et al., 2020; Epstein et al., 2020; Godel et al., 2021; Micallef et al., 2020; Pennycook & Rand, 2019); for a review see (Martel et al., 2024). The rationale is that the "wisdom of crowds" (i. e., the aggregated assessments of non-expert fact-checkers) could result in an accuracy that is comparable to that of experts (Frey & van de Rijt, 2021; Martel et al., 2024; Woolley et al., 2010). A prominent implementation of community-based fact-checking is X's "Community Notes" feature (formerly known as "Birdwatch") (Allen et al., 2022; Chuai et al., 2024; Pröllochs, 2022; X, 2021). This feature allows regular social media users on X to identify posts (formerly "tweets") they believe are misleading and write (textual) annotations that provide *context* to the post, so-called

"community notes" (throughout this paper, we use the term "community notes" platform-agnostic to refer to textual fact-checking annotations carried out by community fact-checkers). Applying such a crowd-based approach to fact-checking social media posts has the potential to address the scalability issues with expert fact-checking. Compared to expert-based assessments, significantly larger quantities (Drolsbach & Pröllochs, 2023; Pröllochs, 2022) and a wider range of posts (Pilarski et al., 2024; A. Zhao & Naaman, 2023) could be fact-checked at a higher speed (A. Zhao & Naaman, 2023). However, evidence on whether exposing users to community notes mitigates trust issues with simple misinformation flags is missing.

Two distinctive factors may lead one to expect that community notes may be perceived as more trustworthy than simple misinformation flags. First, users may be more willing to trust their peers (i. e., community fact-checkers) than experts. Americans' trust in experts and established information sources has been in continued decline over the last years (Gottfried & Liedke, 2021; Kennedy et al., 2022). At the same time, people trust information sources more if they perceive the source as similar to themselves (Brinol & Petty, 2009; Ecker et al., 2022; Mackie et al., 1990; Siegrist, 2021). Hence, higher (perceived) source credibility may render community fact-checks to be perceived as more trustworthy. Second, community notes allow fact-checkers to add context by explaining why the fact-checked post was misleading and linking to relevant external sources (Pröllochs, 2022; X, 2021). People's trust in information increases with the amount of supporting evidence (Racherla et al., 2012; Schwarz & Jalbert, 2020; Siegrist, 2021) and presenting users with additional information that directly refutes a false statement may help to debunk misinformation (Kreps & Kriner, 2022). While context-free misinformation flags may create a gap in a user's mental state (e. g., "why is this post labeled as misleading?"), context refuting misleading claims may also fill in plausible details, which may strengthen the recall of correct information. Overall, both source effects (i. e., higher trust in peers) and the provision of additional context may make community notes to be perceived as more trustworthy than simple misinformation flags.

Previous work on community-based fact-checking has primarily focused on the accuracy of community fact-checkers. Experimental studies found that, while the assessment of individuals might be noisy and ineffective (Woolley et al., 2010), the crowd can be quite accurate in identifying misleading social media content. The assessment of even relatively small crowds is comparable to those of experts (Allen et al., 2021; Bhuiyan et al., 2020; Epstein et al., 2020; Godel et al., 2021; Martel et al., 2024; Pennycook & Rand, 2019; Resnick et al., 2021). Despite challenges with politically motivated flagging (Allen et al., 2022; Pröllochs, 2022), research further shows that users, to a large extent, perceive community-created fact-checks for social media posts as being informative and helpful (Pröllochs, 2022). Also, community fact-checkers surpass experts in the volume, variety, and speed of fact-checking (Drolsbach & Pröllochs, 2023; Pilarski et al., 2024; Pröllochs, 2022; Z. Zhao et al., 2020). Overall, prior studies suggest that community-based fact-checking may address the scalability problem of expert-based approaches. Yet, an understanding of whether users perceive community fact-checks as more trustworthy than misinformation flags is missing. Here, we add by assessing users' trust in community fact-checks for social media posts.

In this study, we conducted a pre-registered survey experiment with $n = 1810$ American residents (politically balanced) to assess their trust in different types of fact-checking interventions (see Methods). Participants in our experiment were randomly assigned to treatments where misleading content was accompanied by context-free misinformation flags in different formats (expert flags or community flags), or by textual community notes explaining why the

fact-checked post was misleading. Participants rated the perceived trustworthiness of each fact-check. We then implemented hierarchical regression models to analyze how trust in fact-checks varies across the different fact-checking interventions. Our objective is to understand whether people are more likely to trust community notes than simple misinformation flags and how the effects vary depending on the political concordance between participants and posts. We also examine source effects, i. e., whether fact-checks carried out by experts or the community are perceived as more trustworthy. This allows us to analyze whether trust in fact-checks is fostered by source effects (higher trust in peers) or the provision of additional context (i. e., fact-checking explanations). Furthermore, we study how users' ability to identify misinformation varies across the different fact-checking interventions.

4.2 Results

Here, we investigate users' trust in fact-checks for social media posts with a preregistered (https://aspredicted.org/rb45k.pdf) four-condition survey experiment (see Methods). We presented $n = 1810$ American residents (M_{age} = 42 years, 51 % female; politically balanced; see SI, Supplement A for further descriptive statistics) with 36 (18 misleading and 18 non-misleading) social media posts and assessed their trust in different types of fact-checks. Participants were recruited via Prolific across multiple sessions and randomly assigned to a control, treatments where all misleading content was accompanied by simple (i. e., context-free) misinformation flags (expert flags and community flags), or a treatment where all misleading content was supplemented with textual community notes explaining why a post was misleading. Each community note in our experiment represented a real-world fact-check from X's Community Notes platform that has been rated as helpful by multiple users with diverse viewpoints (see Methods). All posts and fact-checks were presented in a standard "X format" (see example in Figure 4.1), and were prevalidated (see SI, Supplement E for details) to appeal to subjects with different political views (pro-Republican, pro-Democrat, and politically neutral). Following our preregistration, we further asked the participants for an assessment of the misleadingness of the posts. This allowed us to analyze how users' ability to identify misinformation varied across the four conditions.

4.2.1 Trust in fact-checks

The data suggest significant differences in the perceived trustworthiness of the fact-checks presented to participants across the experimental conditions and political leanings. The participants' mean trust ratings (7-point Likert scale normalized to the interval [0, 1]) are shown in Figure 4.2. On average, simple expert flags were rated as slightly more trustworthy ($M = 0.59$) than simple community flags ($M = 0.57$). According to two-tailed t-tests, this difference was significant for the full sample ($P = 0.005$), but failed to reached statistical significance when looking at Biden supporters ($P = 0.117$) or Trump supporters ($P = 0.052$) separately. Compared to expert flags, exposing participants to textual community notes that elaborate on the reasons on why the fact-checked post was misleading significantly increased ($P < 0.001$) the perceived trustworthiness ($M = 0.63$). This improvement was significantly higher ($P < 0.001$) for Biden supporters ($M = 0.68$) than for Trump supporters ($M = 0.58$). We further observe that the average trust ratings differed according to the political orientation of both the participants and the fact-checked posts. Despite concerns with trust in fact-checkers (Poynter, 2019), Trump-supporting participants rated the majority of fact-checks as at least somewhat trustworthy ($M = 0.56$). However, participants preferring Biden exhibited a

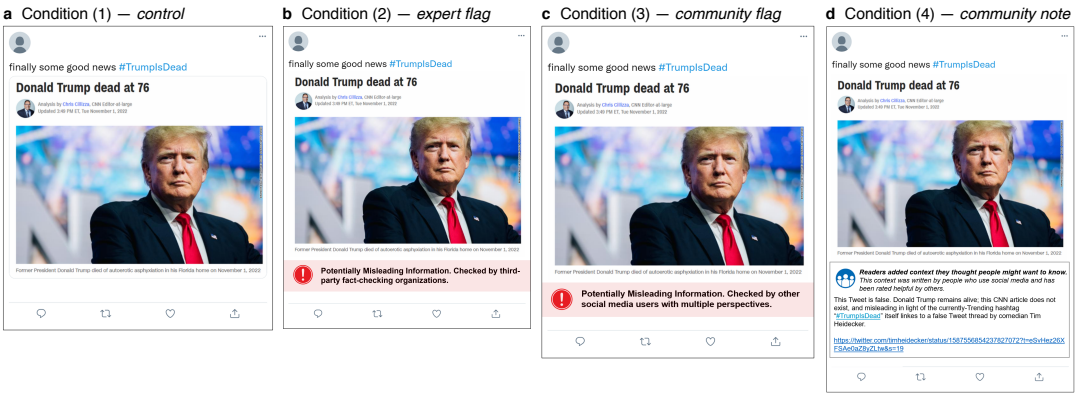


Figure 4.1: Example of a social media post and fact-checks shown to participants. Participants were randomly assigned to one of four conditions: (1) *Control*, in which the participants were presented only with the social media posts (i.e., without fact-checks), (2) *expert flag*, where misleading posts were supplemented by a flag indicating that expert fact-checkers categorized the content as being misleading, (3) *community flag*, where misleading posts were supplemented by a flag indicating that community fact-checkers categorized the content as being misleading, or (4) *community note*, where misleading posts were supplemented with textual community notes explaining why the information in the post is misleading. Posts were politically balanced to appeal to subjects with different political views (pro-Republican, pro-Democrat, and politically neutral).

significantly higher ($P < 0.001$) baseline trust in fact-checks ($M = 0.64$). Furthermore, fact-checks were, on average, perceived as significantly (each $P < 0.001$) less trustworthy if the misleading post was politically concordant ($M = 0.58$) as opposed to misleading posts that were politically neutral ($M = 0.60$) or politically discordant ($M = 0.62$).

Next, we fitted a hierarchical linear regression model with three-way interaction terms and random intercepts for subjects and posts to predict trustworthiness ratings (7-point Likert scale normalized to the interval [0, 1]). This allowed us to control for between-subject variations and study interaction effects (see Methods). Figure 4.3 visualizes the average marginal effects (AME) for our main explanatory variables (full estimation results are in SI, Supplement B). Consistent with the analysis of group-level means, the regression model predicts that community notes were perceived as significantly more trustworthy than simple expert flags. On average, replacing an expert flag with a community note would have increased trust in the fact-check by 4.8 percentage points (AME $= 0.048$; 95 % CI $= [0.042, 0.055]$; $P < 0.001$). In terms of percentages, this corresponds to an increase in trust of 8.2 %. Moreover, simple community flags were slightly less trustworthy than simple expert flags (AME $= -0.013$; 95 % CI $= [-0.020, -0.006]$; $P = 0.003$). This suggests that the higher trustworthiness of community notes primarily stemmed from the additional context (i. e., the fact-checking explanations) rather than the fact-checking source.

The increases in trust varied across the political leanings of participants and posts (see Figure 4.3b). For Biden supporters, replacing an expert flag with a community note would have increased trust in the fact-check by 7.4 percentage points (AME $= 0.074$; 95 % CI $= [0.061, 0.086]$; $P < 0.001$) for politically neutral posts, by 4.5 percentage points (AME $= 0.045$; 95 % CI $= [0.032, 0.058]$; $P < 0.001$) for politically discordant posts, and by 6.4 percentage points (AME $= 0.064$; 95 % CI $= [0.049, 0.080]$; $P < 0.001$) for politically concordant posts. For Trump supporters, replacing an expert flag with a community note would have increased trust in the fact-check by 4.2 percentage points (AME $= 0.042$;

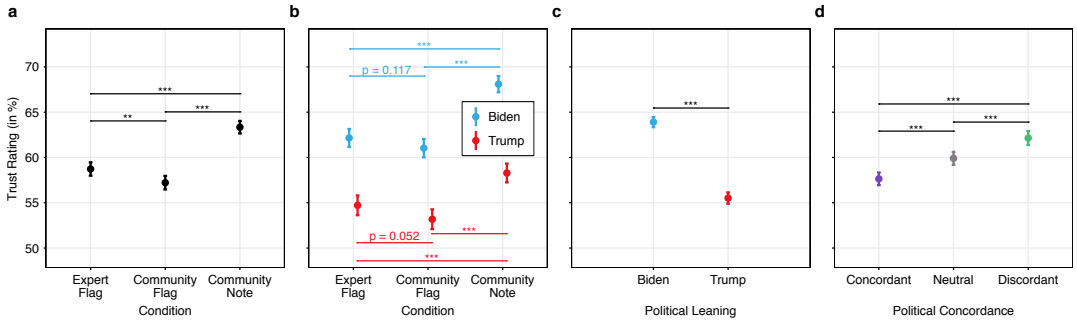


Figure 4.2: Trust in fact-checks for misleading posts. (a) Mean trust ratings across fact-checking interventions, i. e., experimental conditions. Note that trust in fact-checks was only assessed if a fact-check was actually presented to users, i. e., across experimental conditions 2–4. **(b)** Mean trust ratings across experimental conditions separated by participants' political leanings (leaning Trump vs. Biden). **(c)** Trust in fact-checks for participants leaning towards Trump vs. Biden. **(d)** Trust in fact-checks for politically concordant, neutral, and discordant misleading posts. The 7-point Likert scale responses were normalized to the interval [0, 1]. The y-axis begins at 50 %; the scale is the same across each panel. $N = 24\,003$ observations across 1 347 participants. Statistical significance ($^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$) is calculated using two-tailed t-tests.

95 % CI $= [0.025, 0.059]$; $P < 0.001$) for politically neutral posts, by 4.0 percentage points (AME $= 0.040$; 95 % CI $= [0.024, 0.056]$; $P < 0.001$) for politically discordant posts, and by 2.3 percentage points (AME $= 0.023$; 95 % CI $= [0.007, 0.039]$; $P = 0.010$) for politically concordant posts. Complementing earlier research on misinformation flagging (Jia et al., 2022), this implies that participants leaning towards Trump were more hesitant to trust the context provided in community notes when exposed to misinformation that was politically concordant (e. g., COVID-19 misinformation).

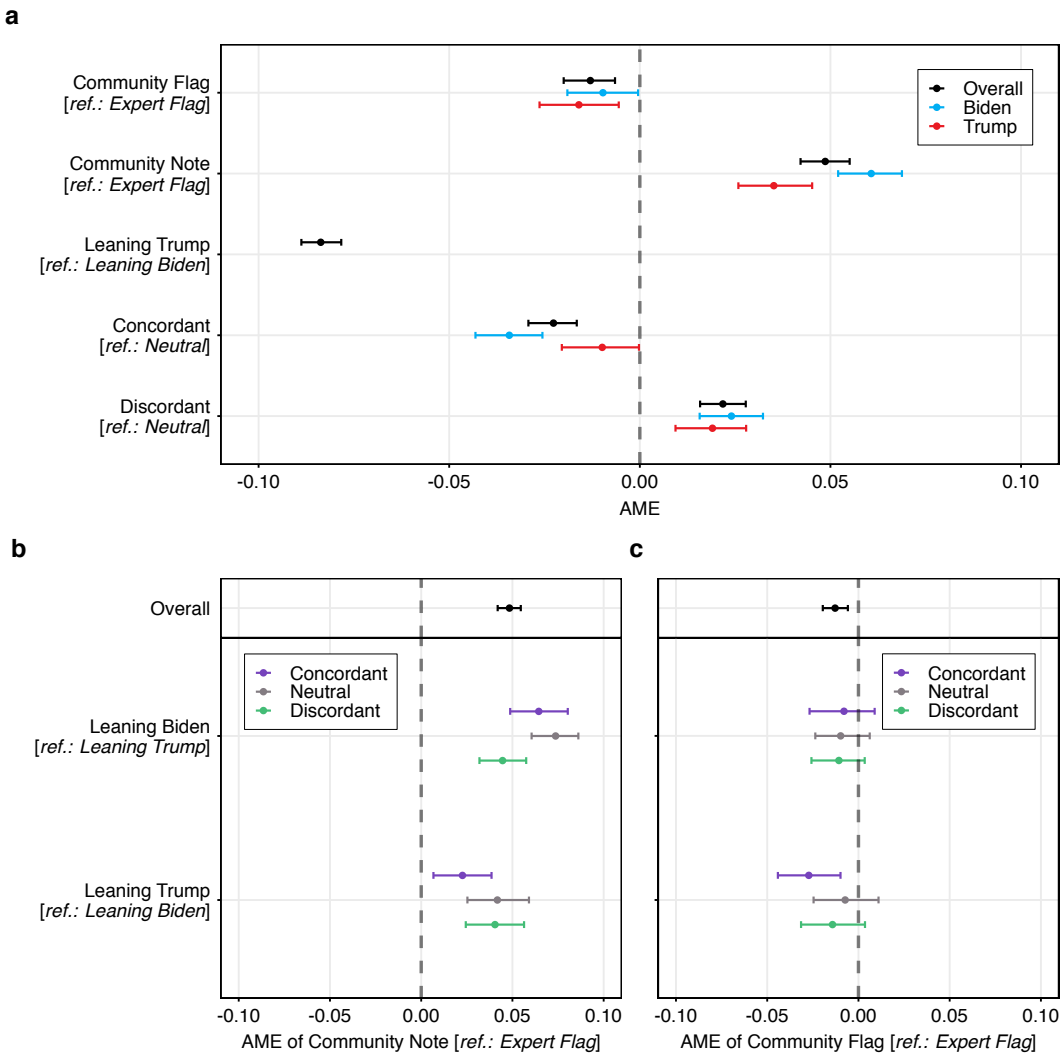


Figure 4.3: Community notes were rated as more trustworthy than simple misinformation flags. **(a)** Shown are the average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with interaction terms predicting trust in fact-checks (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in the average predicted trust ratings between the group of interest and the reference group (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in trust ratings). **(b)** AME of replacing an expert flag with a community note (i. e., the average difference between the predicted marginal effects for community notes vs. expert flags) across the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). **(c)** AME of replacing an expert flag with a community flag (i. e., the average difference between the predicted marginal effects for community flags vs. expert flags) across the political leanings of participants and the fact-checked posts. Control variables and random intercepts for posts and subjects were included. The 95 % confidence intervals (error bars) were derived using the bootstrap method for 1 000 resamples. $N = 24\,003$ observations across 1 347 participants. Full estimation results are in SI, Tables 4.5 and 4.7.

4.2.2 Identification of misleading and non-misleading posts

Next, we analyzed participants' identification of misleading and non-misleading posts. The majority of participants assigned to the control condition (in which no fact-check was shown)

successfully identified misleading posts. The participants' misleadingness ratings (7-point Likert scale normalized to the interval [0, 1]) for posts in the control condition were significantly higher (t-test: $t = 103.23$; $df = 15207$; $P < 0.001$, two-tailed) if they were misleading ($M = 0.73$) than if they were non-misleading ($M = 0.44$). Notably, for both misleading and non-misleading posts, the share of posts rated as at least somewhat misleading was relatively high. A plausible explanation is that we asked participants for an assessment of misleadingness (i. e., a broad concept) rather than just true and false veracity (see Søe, 2021). To quantify the effects of the different fact-checking interventions (i. e., conditions 2–4) on the identification of misleading and non-misleading posts, we fitted a hierarchical linear regression model with four-way interaction terms and random intercepts for subjects and posts to predict misleadingness ratings (see Methods). Figure 4.4 visualizes the average marginal effects (see SI, Section 4.B for full estimation results).

All fact-checking interventions significantly improved the identification of misleading social media posts (see Figure 4.4a,d,e). Compared to the control condition, displaying fact-checks increased the perceived misleadingness by, on average, 7.1 percentage points for expert flags (AME $= 0.071$; 95 % CI $= [0.063, 0.078]$; $P < 0.001$), 6.0 percentage points for community flags (AME $= 0.060$; 95 % CI $= [0.053, 0.067]$; $P < 0.001$), and 9.6 percentage points for community notes (AME $= 0.096$; 95 % CI $= [0.089, 0.102]$; $P < 0.001$). In terms of percentages, these numbers translate to an increase in perceived misleadingness of 9.7 % for expert flags, 8.2 % for community flags, and 13.1 % for community notes. Hence, participants were more likely to correctly identify misleading posts if exposed to a community note vs. an expert flag (AME $= 0.025$; 95 % CI $= [0.019, 0.032]$; $P < 0.001$), and less likely if exposed to a community flag vs. an expert flag (AME $= -0.011$; 95 % CI $= [-0.018, -0004]$; $P = 0.003$). The efficacy of the fact-checking interventions was slightly but significantly (each $P < 0.01$) larger for Biden supporters than for Trump supporters in the case of expert flags (Biden supporters: AME $= 0.085$; 95 % CI $= [0.076, 0.094]$; $P < 0.001$, Trump supporters: AME $= 0.055$; 95 % CI $= [0.044, 0.067]$; $P < 0.001$) and community notes (Biden supporters: AME $= 0.110$; 95 % CI $= [0.100, 0.120]$; $P < 0.001$, Trump supporters: AME $= 0.081$; 95 % CI $= [0.070, 0.091]$; $P < 0.001$), but did not significantly ($P = 0.443$) differ across political leanings for community flags (Biden supporters: AME $= 0.064$; 95 % CI $= [0.054, 0.074]$; $P < 0.001$, Trump supporters: AME $= 0.055$; 95 % CI $= [0.044, 0.066]$; $P < 0.001$). Replicating earlier work Garrett and Bond, 2021; Grinberg et al., 2019; A. Guess et al., 2019; Rathje et al., 2023, we further find that participants preferring Trump performed, on average, significantly worse at identifying misleading posts than participants preferring Biden (AME $= -0.016$; 95 % CI $= [-0.021, -0.011]$; $P = 0.001$). Also, participants were, on average, significantly more likely to correctly identify misleading posts that were politically discordant (AME $= 0.092$; 95 % CI $= [0.087, 0.097]$; $P < 0.001$) than those that were politically concordant (AME $= -0.057$; 95 % CI $= [-0.057, 0.064]$; $P = 0.001$).

The advantage of community notes over simple expert flags in improving the identification of misleading posts depended on the political congruence of the post (see Figure 4.4d). For Biden supporters, replacing an expert flag with a community note would have increased the perceived misleadingness by 2.4 percentage points for politically concordant posts (AME $= 0.024$; 95 % CI $= [0.007, 0.039]$; $P = 0.009$), and by 4.5 percentage points for politically neutral posts (AME $= 0.045$; 95 % CI $= [0.031, 0.059]$; $P < 0.001$). We observe no statistically significant effect for politically discordant posts (AME $= 0.005$; 95 % CI $= [-0.007, 0.016]$; $P = 0.439$). For Trump supporters, replacing an expert flag with a community note would

have increased the perceived misleadingness by 4.5 percentage points for politically concordant posts (AME $= 0.045$; 95 % CI $= [0.027, 0.065]$; $P < 0.001$), and by 4.4 percentage points for politically neutral posts (AME $= 0.044$; 95 % CI $= [0.026, 0.062]$; $P < 0.001$). For politically discordant posts, the effect was not statistically significant (AME $= -0.012$; 95 % CI $= [-0.027, 0.003]$; $P = 0.112$). Hence, both Biden supporters and Trump supporters were relatively more successful in identifying misinformation if exposed to community notes for politically concordant or neutral posts, but not for politically discordant posts.

Across all fact-checking interventions, there were significant treatment condition effects on untagged (i. e., non-labeled) non-misleading posts (see Figure 4.4b); that is, we observed an "implied truth" effect on untagged posts Pennycook, Bear, et al., 2020. Compared to the control condition (in which no fact-check was shown), displaying fact-checks on misleading posts decreased the perceived misleadingness of untagged non-misleading post by, on average, 5.2 percentage points for expert flags (AME $= -0.052$; 95 % CI $= [-0.059, -0.044]$; $P = 0.001$), 5.1 percentage points for community flags (AME $= -0.051$; 95 % CI $= [-0.058, -0.043]$; $P = 0.001$), and 3.2 percentage points for community notes (AME $= -0.032$; 95 % CI $= [-0.039, -0.025]$; $P = 0.001$). Hence, the treatment condition effect on untagged non-misleading posts was significantly less pronounced for community notes than for simple misinformation flags (each $P \leq 0.001$). There were no statistically significant variations in these effects for Biden vs. Trump supporters ($P = 0.306$ for expert flags, $P = 0.587$ for community flags, $P = 0.214$ for community notes).

Altogether, these results imply that community notes consistently improved the identification of misleading posts over simple misinformation flags. However, simple misinformation flags increased belief in untagged, non-misleading posts to a larger extent than community notes. As a result, when comparing the aggregated average marginal effects across both misleading and non-misleading posts (i. e., overall discernment), the advantage of community notes over simple misinformation flags was small (see Figure 4.4c). Compared to the control condition, displaying fact-checks on misleading posts increased the overall discernment of misleading vs. non-misleading posts by, on average, 6.1 percentage points for expert flags (AME $= 0.061$; 95 % CI $= [0.056, 0.066]$; $P < 0.001$), 5.5 percentage points for community flags (AME $= 0.055$; 95 % CI $= [0.050, 0.060]$; $P < 0.001$), and 6.4 percentage points for community notes (AME $= 0.064$; 95 % CI $= [0.059, 0.069]$; $P < 0.001$). The differences between these average marginal effects were statistically significant for expert flags vs. community flags ($P = 0.021$) and community notes vs. community flags ($P < 0.001$); however, not for community notes vs. expert flags ($P = 0.334$).

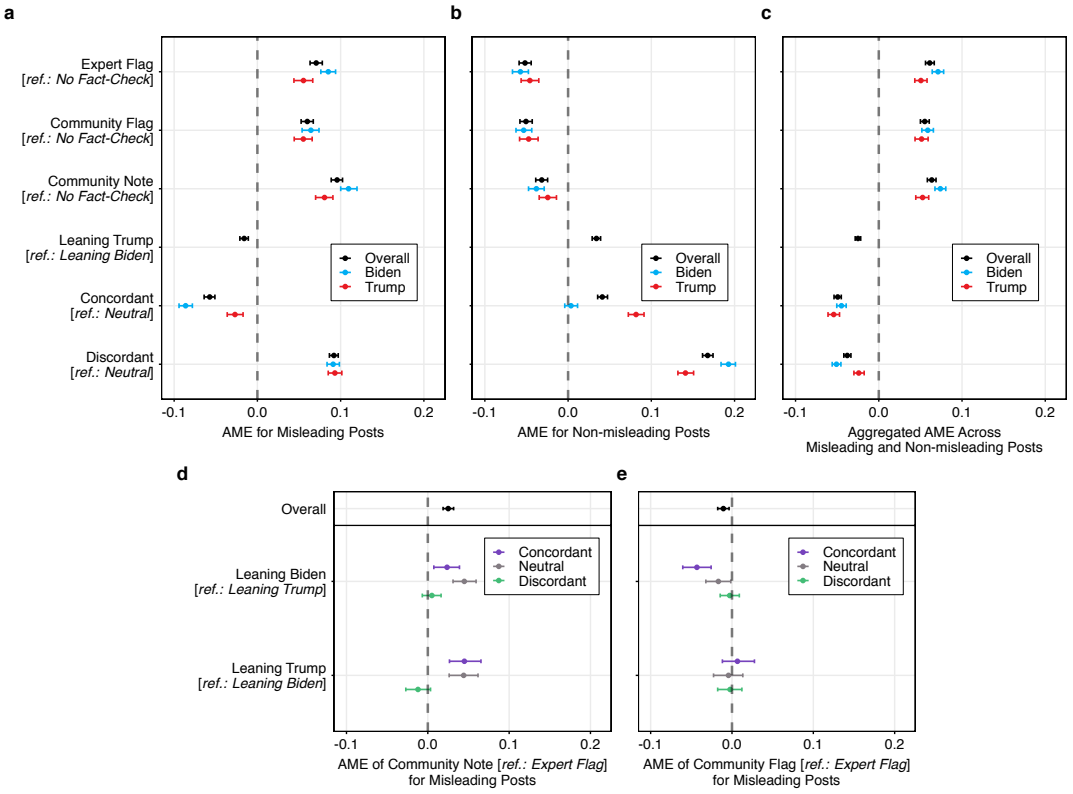


Figure 4.4: All fact-checking interventions improved the identification of misleading posts. Shown are the average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with interaction terms predicting the perceived misleadingness of a post (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in the average predicted misleadingness ratings between the group of interest and the reference group (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in perceived misleadingness). **(a)** AMEs for misleading posts. **(b)** AMEs for non-misleading posts. **(c)** Aggregated AMEs across misleading and non-misleading posts (i. e., overall discernment). **(d)** AMEs when replacing an expert flag with a community note (i. e., the average difference between the predicted marginal effects for community notes vs. expert flags) across the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). **(e)** AMEs when replacing an expert flag with a community flag (i. e., the average difference between the predicted marginal effects for community flags vs. expert flags) across the political leanings of participants and the fact-checked posts. Control variables and random intercepts for posts and subjects were included. The 95 % confidence intervals (error bars) were derived using the bootstrap method for 1 000 resamples. $N = 64\,454$ observations across 1 810 participants. Full estimation results are in SI, Tables 4.6 and 4.8.

4.2.3 Demographics, beliefs, and cognitive reflection

We examined differences across demographics and participants' beliefs towards various topics, which we also asked about as part of our survey (see Methods). Furthermore, participants had to complete a 4-item Cognitive Reflection Test to measure their level of reflective thinking (Mosleh et al., 2021; Thomson & Oppenheimer, 2016) and those in the treatment conditions were additionally asked how they would rate their general reliance on fact-checks. The average marginal effects of these variables on participants' trust in fact-checks and perceived misleadingness are visualized in Figure 4.5. Full results and descriptive statistics are in SI,

Supplement C and Table S1.

Figure 4.5a shows that under the control condition (in which no fact-check was shown), participants performed better in correctly identifying misleading posts if they were vaccinated against COVID-19 (AME $= 0.023$; 95 % CI $= [0.009, 0.037]$; $P = 0.004$) or have passed the CRT (AME $= 0.034$; 95 % CI $= [0.023, 0.045]$; $P < 0.001$). In contrast, participants were less accurate in identifying misleading posts if they belonged to an ethnical minority (AME $= -0.020$; 95 % CI $= [-0.032, -0.007]$; $P = 0.007$) or indicated a low willingness to take risks (AME $= -0.025$; 95 % CI $= [-0.037, -0.013]$; $P = 0.001$). We do not observe statistically significant links between the identification of misleading posts and other variables (see Figure 4.5a). Overall, these results align well with previous work examining predictors of susceptibility to misinformation (Arechar et al., 2023).

We further studied moderation effects, i. e., how the efficacy of fact-checking interventions varied depending on demographics, beliefs, and cognitive reflection (Figure 4.5b). Confirming previous work (Martel & Rand, 2023b), all considered fact-checking interventions had a broad efficacy. Across all conditions, belief in misinformation was consistently reduced across all demographics, i. e., misleading posts were more likely to be correctly identified (each $P < 0.001$).

When comparing the efficacy of different fact-checking interventions (Figure 4.5c), replacing an expert flag with a community note would have consistently increased trust in fact-checks (each $P < 0.001$); except for participants who indicated that their general reliance on fact-checks is low (AME $= 0.005$; 95 % CI $= [-0.008, 0.017]$; $P = 0.439$). The advantage of community notes over expert flags was particularly pronounced for male (AME $= 0.053$; 95 % CI $= [0.044, 0.062]$; $P < 0.001$) and young participants (AME $= 0.054$; 95 % CI $= [0.041, 0.067]$; $P < 0.001$), ethnic minorities (AME $= 0.067$; 95 % CI $= [0.053, 0.078]$; $P < 0.001$), participants who stated that they did not believe in God(s) (AME $= 0.077$; 95 % CI $= [0.068, 0.086]$; $P < 0.001$), and those who passed the CRT (AME $= 0.074$; 95 % CI $= [0.065, 0.083]$; $P < 0.001$). Regarding the identification of misleading posts, the highest improvements were observed for female participants (AME $= 0.035$; 95 % CI $= [0.026, 0.044]$; $P < 0.001$), participants who were not vaccinated against COVID-19 (AME $= 0.039$; 95 % CI $= [0.025, 0.051]$; $P < 0.001$), and for participants who indicated having a high willingness to take risks (AME $= 0.039$; 95 % CI $= [0.031, 0.047]$; $P < 0.001$).

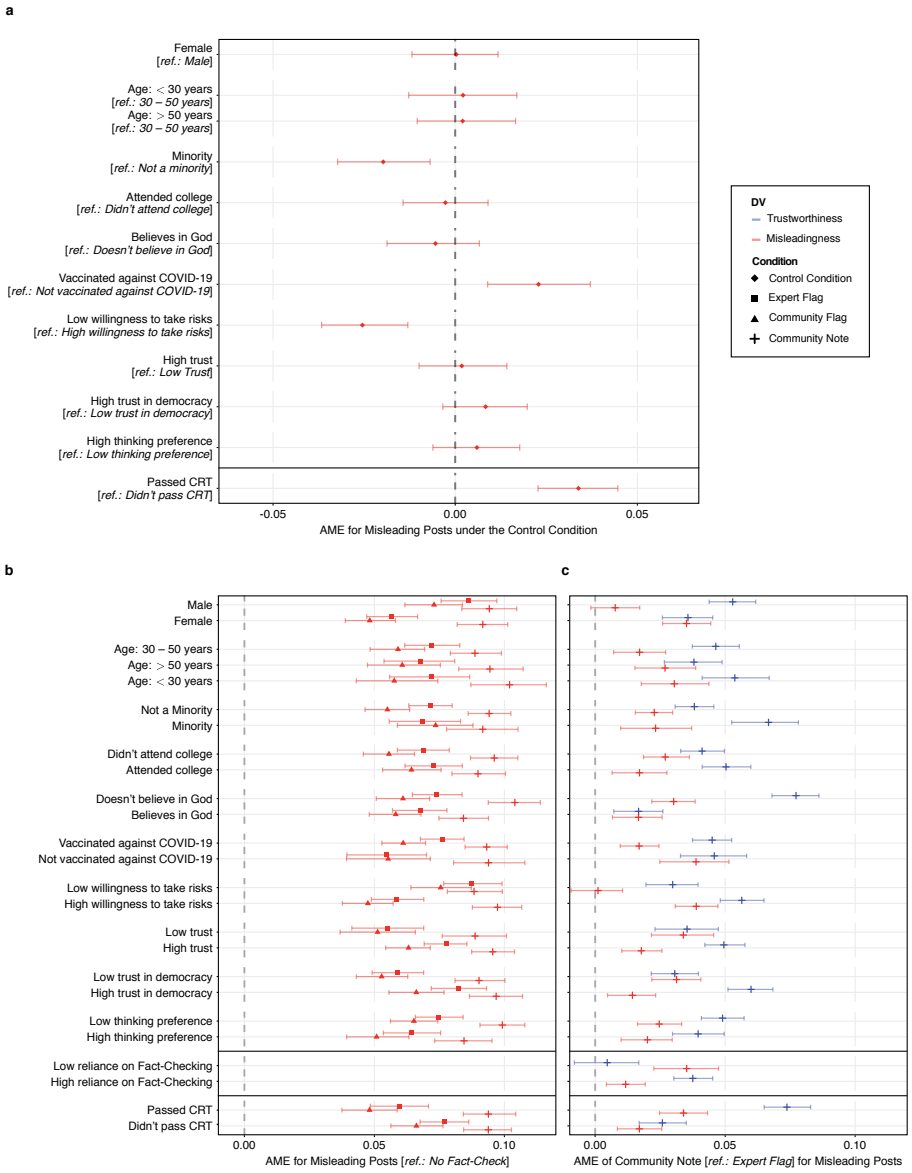


Figure 4.5: Effects of demographics, fact-checking reliance, and cognitive reflection. (a) Average marginal effects (AME) for misleading posts under the control condition. **(b)** AMEs for misleading posts across different fact-checking conditions (relative to the control condition). **(c)** AMEs when replacing an expert flag with a community note. AMEs (symbols) were calculated based on hierarchical linear regression models with interaction terms across two dependent variables (DV): trustworthiness (in blue), and misleadingness (in red). The 7-point Likert-scale responses were rescaled to the interval [0, 1]. Control variables and random intercepts for posts and subjects were included. The 95 % confidence intervals (error bars) were derived using the bootstrap method for 1 000 resamples. Full results are in SI, Tables 4.10 to 4.16.

4.2.4 Effect of presentation format and context

We carried out an additional experiment to study the extent to which different presentation formats influence the perceived trustworthiness of community notes. In the additional experiment, $n = 675$ participants ($M_{age} = 43$, 51% female, politically balanced) were randomly assigned

to one of two conditions, representing two different presentation formats of community notes (see Figure 4.6a). The two formats differed in such a way that the textual community note was presented either without an explicit warning label (i. e., former condition 4) or with an explicit warning label indicating that community fact-checkers categorized the content as being misleading (i. e., a combination of former conditions 3 and 4). Apart from the treatments, the experiment had the same elements and was conducted in exactly the same way as the original experiment (see Methods).

We fitted hierarchical linear regression models to quantify the effects of the alternative presentation format of community notes on trust in fact-checks and the identification of misleading and non-misleading posts (see SI, Supplement C). Figure 4.6b/c visualizes the average marginal effects (AME). Regarding trust in fact-checks, there were, on average, no statistically significant differences between the two presentation formats (AME $= -0.001$; 95 % CI $= [-0.008, 0.007]$; $P = 0.846$). However, there were significant differences when looking at Biden vs. Trump supporters individually. For Biden supporters, the alternative presentation format with an explicit warning label decreased the trust in community notes by 2.6 percentage points (AME $= -0.026$; 95 % CI $= [-0.036, -0.016]$; $P = 0.001$). In contrast, the alternative presentation format increased trust in community notes by 2.5 percentage points for participants leaning towards Trump (AME $= 0.025$; 95 % CI $= [0.015, 0.036]$; $P < 0.001$).

We linearly aggregated the effect sizes from both experiments to provide an estimate for the isolated effect of adding context to community fact-checks (i. e., warning label vs. warning label + context). Compared to context-free community warning labels (i. e., condition 3 in the main experiment), community notes combining the same warning label with context (i. e., the alternative presentation format in the additional experiment) were, on average, linked to 5.35 percentage points higher trust in fact-checks (95 % CI $= [0.047, 0.060]$; $P < 0.001$). In terms of percentages, this translates to an increase in trust of 9.35 %. Notably, the isolated effect of fact-checking context on trust was larger for Trump supporters with 6.65 percentage points (95 % CI $= [0.057, 0.076]$; $P < 0.001$) than for Biden supporters with 3.47 percentage points (95 % CI $= [0.027, 0.044]$; $P < 0.001$).

The alternative presentation format improved the identification of misleading and non-misleading posts. On average, presenting community notes with an explicit warning label increased the perceived misleadingness of misleading posts by 2.0 percentage points (AME $= 0.020$; 95 % CI $= [0.012, 0.027]$; $P < 0.001$), and decreased the perceived misleadingness of non-misleading posts by 1.4 percentage points (AME $= -0.014$; 95 % CI $= [-0.023, -0.005]$; $P = 0.001$). The variation across the political leanings of the participants was small. For Biden supporters, we observe a smaller positive effect for misleading posts (AME $= 0.014$; 95 % CI $= [0.004, 0.024]$; $P = 0.011$) and no statistically significant effect for non-misleading posts (AME $= -0.012$; 95 % CI $= [-0.023, 0.000]$; $P = 0.050$). For Trump supporters, we observe a slightly larger positive effect for misleading posts (AME $= 0.026$; 95 % CI $= [0.015, 0.037]$; $P < 0.001$) and a slightly larger negative effect for non-misleading posts (AME $= -0.017$; 95 % CI $= [-0.028, -0.004]$; $P = 0.011$). In terms of overall discernment, community notes with an explicit warning label significantly increased participants' accuracy by, on average, 1.66 percentage points (AME $= 0.017$; 95 % CI $= [0.011, 0.022]$; $P < 0.001$). This effect was again larger for Trump supporters (AME $= 0.021$; 95 % CI $= [0.012, 0.029]$; $P < 0.001$) than for Biden supporters (AME $= 0.013$; 95 % CI $= [0.005, 0.020]$; $P = 0.003$).

We again related these numbers back to the results from our main experiment to assess the isolated effect of context on the identification of misleading vs. non-misleading posts. Compared to context-free community warning labels, adding context increased the perceived misleadingness of misleading by posts by 5.54 percentage points on average (95 % CI $= [0.049, 0.062]$; $P < 0.001$). In terms of percentages, this number translates to an increase in perceived misleadingness of 6.99 %. The increase was slightly higher for participants leaning towards Biden with 5.95 percentage points (95 % CI $= [0.051, 0.068]$; $P < 0.001$) than for those leaning towards Trump with 5.07 percentage points (95 % CI $= [0.041, 0.061]$; $P < 0.001$). For non-misleading posts, the effect of additional context was small (0.53 percentage points overall; 0.37 percentage points for Biden supporters; and 0.70 percentage points for Trump supporters) and not statistically significant ($P = 0.155$ overall; $P = 0.494$ for Biden supporters; and $P = 0.234$ for Trump supporters). The improvement in overall discernment of misleading vs. non-misleading posts was 2.51 percentage points on average (95 % CI $= [0.020, 0.030]$; $P < 0.001$), 2.78 percentage points for Biden supporters (95 % CI $= [0.021, 0.034]$; $P < 0.001$), and 2.19 percentage points for Trump supporters (95 % CI $= [0.015, 0.030]$; $P < 0.001$).

Overall, our additional experiment indicates that presentation effects for community notes were relatively small. We observed, on average, no statically significant differences in trustworthiness between community notes presenting solely context vs. community notes combining context with explicit warning labels. In combination with the findings from our main experiment, this further supports that the higher trustworthiness of community notes (vs. simple misinformation flags) primarily stems from the additional context rather than different formats of labels/warning messages. Nonetheless, combining fact-checking explanations with explicit warning labels may have the potential to slightly improve users' ability to identify misinformation. A possible reason may be that the explicit warning labels offer users a more definite cue about whether or not a post is misleading, thereby encouraging them to adhere to the labels. In contrast, community notes that lack an explicit warning label and instead solely focus on offering context (i. e., a more "neutral" presentation) may imply to users that the judgment is, to a relatively larger extent, up to them. The consistent trust levels for community notes across both presentation formats indicate that, on average, these improvements may be achieved without compromising users' trust in fact-checks.

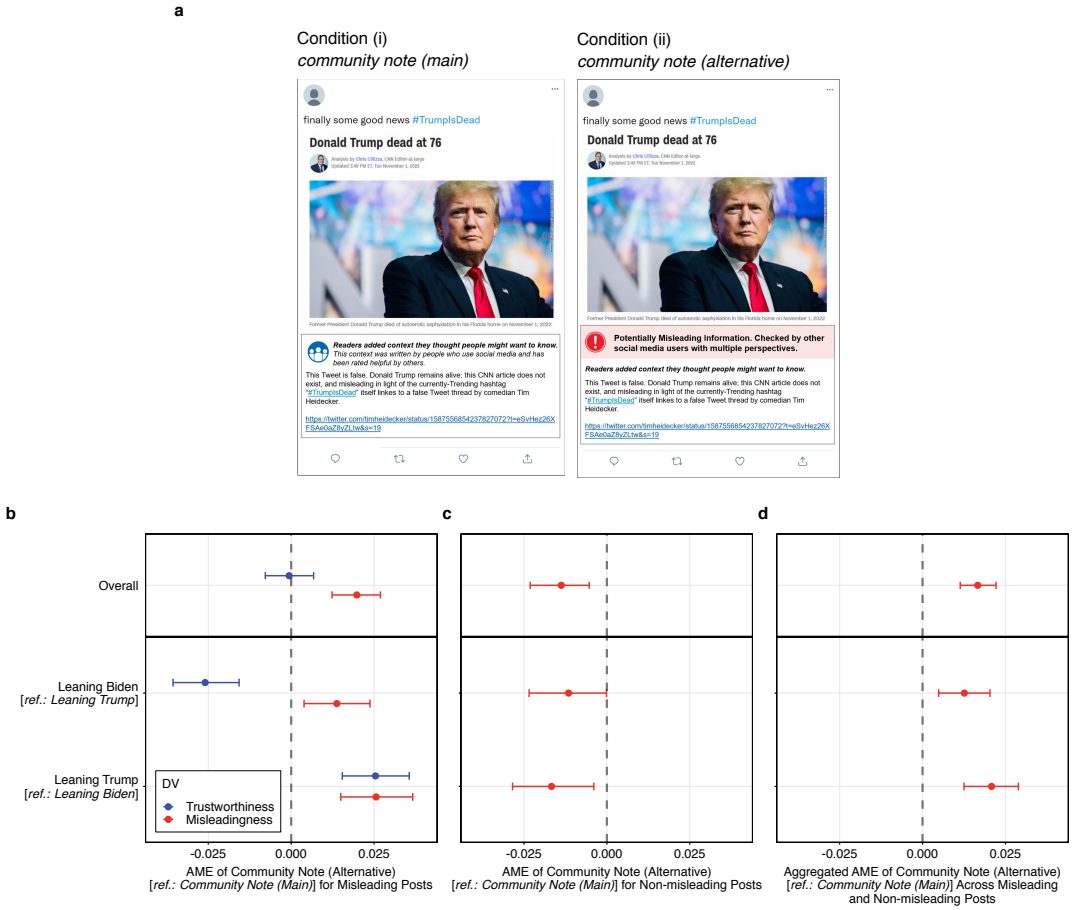


Figure 4.6: Additional experiment for effects of presentation format. (a) Example of a social media post and fact-checks shown to participants. Participants were randomly assigned to one of two conditions: (i) *community note (main)*, where misleading posts were supplemented with a textual community note without an explicit warning label (i. e., previous condition 4), or (ii) *community note (alternative)*, where misleading posts were supplemented by a combination of a textual community note and an explicit warning label indicating that community fact-checkers categorized the content as being misleading. Posts were politically balanced to appeal to subjects with different political views (pro-Republican, pro-Democrat, and politically neutral). **(b)** Average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with interaction terms predicting trustworthiness (blue), and misleadingness (red) (7-point Likert scale normalized to the interval [0, 1]) for misleading posts. **(c)** AMEs for non-misleading posts. **(d)** Aggregated AMEs for misleading and non-misleading posts (i. e., overall discernment). The 95 % confidence intervals (error bars) were derived using the bootstrap method for 1 000 resamples. $N = 12.168$ (24.300) observations across 675 participants. Full results are in SI, Tables 4.22 and 4.23.

4.3 Discussion

Here, we presented $n = 1810$ Americans with 36 misleading and non-misleading social media posts and assessed their trust in different types of fact-checking interventions. All posts represented real-world items from the social media platform X, and all fact-checks were surfaced by Community Notes' bridging algorithm. Across all demographics and political leanings, we found that text-based community notes explaining why a fact-checked post was misleading were perceived as significantly more trustworthy than simple (i. e., context-free)

misinformation flags. Only marginal differences were observed between simple expert flags and simple community flags. This implies that the increases in trustworthiness primarily stemmed from the context provided in community notes rather than the fact-checking source. Furthermore, in an additional experiment, we found no statically significant differences in trustworthiness across different presentation formats, that is, between community notes presenting solely text-based explanations vs. community notes combining warning labels and text-based explanations. Overall, our findings demonstrate that context matters in fact-checking on social media and that exposing users to text-based community notes might be an effective approach to mitigate trust issues with simple misinformation flags.

Previous work has evaluated the efficacy of simple misinformation flags regarding misinformation discernment and sharing intentions (Clayton et al., 2020; Kim & Dennis, 2019; Martel & Rand, 2023a; Mena, 2020; Moravec et al., 2020; Ng et al., 2021; Pennycook, Bear, et al., 2020; Pennycook, McPhetres, et al., 2020; Porter & Wood, 2021). Consistent with the vast majority of these works, we find that fact-checking interventions significantly improved the identification of misleading and non-misleading social media posts (we also analyzed sharing intentions; yet, we relegated these findings to the SI due to potential spillover effects when asking participants about misleadingness before assessing sharing intentions Epstein et al., 2023; see SI, Supplement C.1). However, the advantage of community notes (over simple misinformation flags) depended on the political congruence of the fact-checked post. Across both sides of the political spectrum, replacing a simple misinformation flag with a community note would have made users' assessments relatively more accurate for politically concordant misinformation, but not for politically discordant misinformation. Hence, the context in community notes appears to be particularly helpful in countering resistance to fact-checking for partisan-aligned misinformation (e. g., a Trump supporter exposed to COVID-19 misinformation). This is an encouraging finding, as politically concordant misleading posts are also the type of misinformation that people are more likely to believe and share on social media. In contrast, additional context might be (relatively) less helpful for politically discordant misinformation (e. g., a Biden supporter exposed to COVID-19 misinformation), which users might perceive as clearly false and which might result in them being more inclined to rely on their existing beliefs or knowledge. Interestingly, our analysis further showed that treatment condition effects on untagged non-misleading posts were significantly less pronounced for community notes than for simple misinformation flags. This indicates that the presence of community notes on misleading posts led to a less strong expectation that untagged, non-misleading posts were true.

Our findings reinforce the notion that partisanship plays a key role regarding the success of fact-checking interventions (Altay et al., 2023). While the majority of users across both sides of the political spectrum trusted the fact-checks in our study, participants preferring Trump over Biden were overall significantly less likely to do so. In particular, Trump supporters tended to distrust fact-checking context for politically concordant misinformation such as, for example, COVID-19 misinformation. These observations align with prior studies, suggesting that adherents to the left tend to be less tolerable to the spread of misinformation and have greater trust in fact-checking (González-Bailón et al., 2022; Shin & Thorson, 2017). Moreover, fact-checks were, on average, perceived as less trustworthy if the misleading post was politically concordant. This observation is supported by prior research that humans seek information that confirms their partisan preferences (Ditto et al., 2019) and that the perceived credibility of information is greater when it is consistent with the recipient's existing political beliefs

(Kelly, 2019; Traberg & van der Linden, 2022). Replicating earlier findings (Garrett & Bond, 2021; Grinberg et al., 2019; A. Guess et al., 2019; Rathje et al., 2023), we also observed that participants leaning to the political right were less accurate in identifying misleading posts, and more likely to re-share misinformation.

As with other research, ours is not free of limitations that offer opportunities for future research. First, analogous to earlier works (Epstein et al., 2020; Mena, 2020; Pennycook, Bear, et al., 2020), only posts that were verifiably misleading were presented with fact-checks and *all* misleading posts in the treatment conditions received flags/community notes. In practice, however, the validity of misinformation warnings can vary, and invalid misinformation warnings may lead individuals to discard authentic content (Freeze et al., 2021). Thus, future studies should examine how users' behavior changes if presented with erroneously labeled posts and the effects of labeling varying proportions of misleading posts. Second, our study focused on the cultural context of American misinformation and American participants. While previous research suggests that solutions to the misinformation challenge may be similarly effective around the globe (Arechar et al., 2023), future work should nonetheless explore how our findings translate to other countries and cultural contexts. Third, the survey participants recruited via Prolific were not nationally representative. Although our study was conducted among social media users, the survey population likely overrepresented higher-educated users. Fourth, additional limitations arise due to the selection of post and (high-quality) fact-checks surfaced by Community Notes' bridging algorithm. While we ensured that posts were politically balanced to appeal to subjects with different political views, more research is necessary to understand how the efficacy varies for community notes of varying quality, writing styles, and political stances. Fifth, future research may experiment with additional formats for the design of flags and notes. However, our study indicates that such presentation effects may be relatively small. Ultimately, our work was performed in an experimental context, which differs from the common environment users face when browsing through social media. However, the use of online surveys provided a sandbox to test otherwise challenging interventions, and previous works have shown that misinformation interventions replicate well in field experiments (Pennycook & Rand, 2021).

From a broader perspective, misinformation on social media poses serious threats to democracy and society as a whole. Major social media providers thus have been called upon to develop effective countermeasures to combat the spread of misinformation on their platforms. However, current approaches to fact-checking social media content do not live up to their full potential. A major challenge is that fact-checking on social media must deal with distrust towards fact-checkers. Our research demonstrates that community-based fact-checking systems (e. g., X's Community Notes) that focus on providing fact-checking context have the potential to at least mitigate trust issues that are common in traditional approaches to fact-checking. Fostering trust in fact-checking is vitally important, especially as we face upcoming events, such as elections, and emerging challenges due to the scalability of AI-generated misinformation.

4.4 Methods

4.4.1 Participants

We recruited a large sample of American residents (total n = 1810, M_{age} = 42 years, 51 % female, 52 % preferred Biden over Trump) via Prolific (https://www.prolific.co/) across seven experimental sessions conducted between April and June 2023 (see SI, Supplement D for

details). The study design was identical in all sessions. Although Prolific is not nationally representative, research shows that it provides high-quality data (Douglas et al., 2023).

To achieve a representative distribution of Trump and Biden supporters (Republicans are underrepresented on Prolific Douglas et al., 2023), we explicitly recruited Biden and Trump supporters, respectively, in sessions 2–7. To do so, we used the prescreening tool available through Prolific, contacting only participants who indicated they voted for Trump or Biden in the 2020 presidential election. Furthermore, we ensured that the participants were representative across genders in all sessions (women are overrepresented on Prolific). Detailed summary statistics are in SI, Supplement A.

4.4.2 Materials

All participants were presented with the same set of 36 social media posts (18 "misleading" and 18 "non-misleading" posts) across multiple topics (e.g., Politics, Business, Health, Celebrities). The full list of social media posts is in SI, Table S24. All posts were presented in a standard "X format" (see example in Figure 4.1). To minimize author-specific effects, the posts were anonymized, and all information about likes, shares, and comments was removed. The misleading social media posts and corresponding fact-checks have been manually selected from X's Community notes platform (available via https://twitter.com/i/communitynotes). At the time of selection, each fact-check has been classified as helpful by Community Notes' bridging algorithm, i. e., was rated helpful by multiple users with diverse viewpoints (Wojcik et al., 2022) (see Supplement E). The non-misleading social media posts have been manually selected from X to cover similar stories and topics. Hence, all fact-checks and posts in our study represent real-world items from a major social media platform. All posts were additionally fact-checked by three trained research assistants who accessed professional fact-checks (snopes.com, factcheck.org, etc.) and other reliable sources to ensure that the labels for misleading and non-misleading post items were accurate (see SI, Supplement E).

The post items in the misleading and non-misleading categories were politically balanced to appeal to subjects with different political views (see SI, Supplement E). Specifically, in each category, six posts were Republican-consistent (pro-Republican/anti-Democrat), six were Democrat-consistent (pro-Democrat/anti-Republican), and six were politically neutral. To examine differences in political orientation, pro-Democrat/anti-Republican posts were classified as concordant for participants favoring Biden over Trump and discordant for participants favoring Trump over Biden (and vice versa for pro-Republican/anti-Democrat posts). Politically neutral posts were classified as neutral for both groups of participants. The categorizations of the political orientations of the posts were validated with the help of three trained research assistants (see SI, Supplement E).

4.4.3 Procedure

The procedure was identical in all experimental sessions. Participants were randomly assigned to one of four conditions (see Figure 4.1) using a between-subject design: (1) *Control*, where the participants were presented only with the social media posts (i.e., without fact-checks), (2) *expert flag*, where participants were presented with the same posts, but misleading posts were supplemented by a flag indicating that expert fact-checkers categorized the content as being misleading, (3) *community flag*, where participants were presented with the same posts, but misleading posts were supplemented by a flag indicating that community fact-checkers

categorized the content as being misleading, or (4) *community note*, where participants were presented with the same posts, but misleading posts were supplemented with textual community notes explaining why the information in the post is misleading. In conditions (2–4), all misleading posts have been supplemented by the respective fact-check, and all non-misleading posts have been displayed without any fact-check. Posts were displayed to participants in a random order to minimize presentation-order effects.

Before starting the survey, participants were presented with the following instructions: "In this survey, you will be presented with a set of social media posts. Please read the posts carefully and answer the following questions for each post: (1) To the best of your knowledge, is the above social media post misleading? (2) How trustworthy do you think the [expert fact-check, community fact-check, textual community note] for the above social media post is? (3) If you were to see the above content on social media, how likely would you be to share it?" Question 2 was only asked for misleading posts in conditions (2–4). Participants assigned to conditions (2–4) have been provided with a general explanation of the origin of the fact-checks. For each social media post, participants were asked to answer the above questions on a 7-point Likert-scale. The Likert scales for the three questions (Q1–Q3) were specified as follows: (Q1) Extremely Accurate (1) to Extremely Misleading (7); (Q2) Extremely Untrustworthy (1) to Extremely Trustworthy (7); (Q3) Extremely Unlikely (1) to Extremely Likely (7).

At the end of the survey, we asked participants questions about demographics, their beliefs toward various topics, and they had to answer a 4-item cognitive reflection test (see SI, Supplement D).

4.4.4 Analysis

We implemented a hierarchical linear regression model with random intercepts for posts and subjects to predict participants' trust in fact-checks at the level of the response item. For the sake of interpretability, we normalized the 7-point Likert scale responses to the interval [0, 1]. This allows us to interpret the effects in our model as percentage point changes. The key explanatory variables were dummies referring to the experimental conditions. For our analysis of trust in fact-checks, we restricted the analysis to misleading posts and conditions 2–4 (baseline = Condition 2/Expert flag). The reason is that the question regarding the trustworthiness of fact-checks was only asked when a fact-check was presented and thus only for misleading posts in conditions 2–4. In addition, we controlled for the political leaning of the participants, i. e., whether they favored Biden or Trump (baseline = Biden), and whether a specific post was aligned with a participant's political leaning (concordant, neutral, or discordant; baseline = neutral). To examine interaction effects, we included three-way interaction terms between the condition dummies, the political leaning of the participants, and the political alignment of posts.

Additionally, we implemented hierarchical linear regression models to analyze how misinformation discernment varied across the four conditions. To this end, we used the misleadingness ratings for the posts as dependent variables. We again normalized the 7-point Likert scale responses to the interval [0, 1]. The key explanatory variables were dummies referring to the four experimental conditions (baseline = Condition 1/No fact-check), and a dummy indicating whether a post was non-misleading (= 1) or misleading (= 0). The control variables for the political leanings of the participants and the political alignment of the posts were the same as in the previous model. Furthermore, we included four-way interaction terms between the

condition dummies, the misleading dummy, the political leaning of the participants, and the political alignment of posts to study interaction effects. Random intercepts for posts and subjects were also included.

Since our models include multiple higher-order interactions, the coefficient estimates do not easily translate to our quantities of interest. Following best practices (Leeper, 2021), we thus calculated the average marginal effects (AME) to interpret the effect sizes. AME is the difference in the average predicted effect between the group of interest and the reference group. For instance, an AME of $+0.05$ for the Trump dummy variable indicates a 5 percentage point difference in the expected value of the dependent variable for participants favoring Trump (as opposed to those favoring Biden).

Multiple exploratory analyses and checks validated our results and confirmed their robustness (see SI, Supplement C): (i) we analyzed how sharing intentions varied across the four condition. Furthermore, we performed additional moderation analyses (ii) to examine differences across demographics and participants' beliefs towards various topics (e. g., belief in God, risk preferences); (iii) to examine how reliance on fact-checks is associated with participants' ratings; (iv) to examine the role of cognitive reflection based on the outcomes of a 4-item Cognitive Reflection Test. (v) We conducted an additional experiment to test the effects of alternative presentation formats on the efficacy of community notes. As part of our robustness checks, we also (vi) tested a wide range of alternative model specifications. For instance, we repeated our analyses using logistic mixed-effects models treating the Likert-scale responses as binary variables, and repeated our analysis using a linear regression model with robust standard errors clustered on both subjects and posts. Moreover, we tested model variants including by-post and by-subject random slopes. However, these "maximal" models failed to converge or resulted in singular fits, i. e., overfitting. Random slopes were therefore omitted and we instead opted for models with crossed random intercepts (Bates et al., 2015). Finally, (vi) we performed a variety of model checks (e. g., assessing variance inflation factors) to ensure that our estimates are robust. In all cases, our results were robust and consistently supported our findings.

Our models were implemented in R 4.3.2 using the `lme4` package and the `marginaleffects` package.

Bibliography

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393.

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539.

Allen, J., Martel, C., & Rand, D. G. (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *CHI*.

Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *HKS Misinformation Review*, *4*(4).

Aral, S., & Eckles, D. (2019). Protecting elections from social media manipulation. *Science*, *365*(6456), 858–861.

Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M. N., et al. (2023). Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, *7*(9), 1502–1513.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). New threats to society from free-speech social media platforms. *Communications of the ACM*, *66*(10), 37–40.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv*. https://arxiv.org/abs/1506.04967

Bhuiyan, M. M., Zhang, A. X., Sehat, C. M., & Mitra, T. (2020). Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction*, *4*, 1–26.

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

Brinol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, *20*(1), 49–96.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health*, *108*(10), 1378–1384.

Calo, R., Coward, C., Spiro, E. S., Starbird, K., & West, J. D. (2021). How do you solve a problem like misinformation? *Science Advances*, *7*(50), eabn0481.

Chuai, Y., Pilarski, M., Renault, T., Restrepo-Amariles, D., Troussel-Clément, A., Lenzini, G., & Pröllochs, N. (2024). Community-based fact-checking reduces the spread of misleading posts on social media. *arXiv preprint arXiv:2409.08781*.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., et al. (2020). Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*, 1073–1095.

Ditto, P. H., Liu, B. S., Clark, C. J., Wojcik, S. P., Chen, E. E., Grady, R. H., Celniker, J. B., & Zinger, J. F. (2019). At least bias is bipartisan: A meta-analytic comparison of

partisan bias in liberals and conservatives. *Perspectives on Psychological Science*, *14*(2), 273–291.

Donovan, J. (2020). Social-media companies must flatten the curve of misinformation. *Nature*.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparison between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE*, *18*(3), e0279720.

Drolsbach, C. P., & Pröllochs, N. (2023). Diffusion of community fact-checked misinformation on Twitter. *CSCW*.

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

Epstein, Z., Pennycook, G., & Rand, D. (2020). Will the crowd game the algorithm? using layperson judgments to combat misinformation on social media by downranking distrusted sources. *Chi*.

Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G., & Rand, D. (2023). The social media context interferes with truth discernment. *Science Advances*, *9*(9), eabo6169.

Feuerriegel, S., DiResta, R., Goldstein, J. A., Kumar, S., Lorenz-Spreen, P., Tomz, M., & Pröllochs, N. (2023). Research can help to tackle ai-generated disinformation. *Nature Human Behaviour*, *7*, 1818–1821.

Freeze, M., Baumgartner, M., Bruno, P., Gunderson, J. R., Olin, J., Ross, M. Q., & Szafran, J. (2021). Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Political Behavior*, *43*(4), 1433–1465.

Frey, V., & van de Rijt, A. (2021). Social influence undermines the wisdom of the crowd in sequential decision making. *Management Science*, *67*(7), 4273–4286.

Gallotti, R., Valle, F., Castaldo, N., Sacco, P., & De Domenico, M. (2020). Assessing the risks of 'infodemics' in response to Covid-19 epidemics. *Nature Human Behaviour*, *4*(12), 1285–1293.

Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances*, *7*(23), eabf1234.

Godel, W., Sanderson, Z., Aslett, K., Nagler, J., Bonneau, R., Persily, N., & Tucker, J. A. (2021). Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety*, *1*(1), 1–36.

González-Bailón, S., d'Andrea, V., Freelon, D., & De Domenico, M. (2022). The advantage of the right in social media news sharing. *PNAS Nexus*, *1*(3), pgac137.

Gottfried, J., & Liedke, J. (2021). Partisan divides in media trust widen, driven by a decline among republicans.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, *5*(1), eaau4586.

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour*, *4*(5), 472–480.

Instagram. (2019). Combatting misinformation on Instagram.

Jia, C., Boltz, A., Zhang, A., Chen, A., & Lee, M. K. (2022). Understanding effects of algorithmic vs. community label on perceived accuracy of hyper-partisan misinformation. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–27.

Kelly, D. (2019). Evaluating the news: (Mis)Perceptions of objectivity and credibility. *Political Behavior*, *41*(2), 445–471.

Kennedy, B., Tyson, A., & Funk, C. (2022). Americans' trust in scientists, other groups declines.

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.

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J. B., Barzilai, S., Basol, M., berinsky adam, a., Betsch, C., Cook, J., Fazio, L., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., … Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *Forthcoming*.

Kreps, S. E., & Kriner, D. L. (2022). The COVID-19 infodemic and the efficacy of interventions intended to reduce misinformation. *Public Opinion Quarterly*, *86*(1), 162–175.

Lazer, D., 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.

Leeper, T. J. (2021). Interpreting regression results using average marginal effects with R's margins.

Mackie, D. M., Worth, L. T., & Asuncion, A. G. (1990). Processing of persuasive in-group messages. *Journal of Personality and Social Psychology*, *58*(5), 812.

Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2024). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, *19*(2), 477–488.

Martel, C., & Rand, D. G. (2023a). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, *54*, 101710.

Martel, C., & Rand, D. G. (2023b). Fact-checker warning labels are effective even for those who distrust fact-checkers. *PsyArXiv*. https://doi.org/10.31234/osf.io/t2pmb

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy & Internet*, *12*(2), 165–183.

Micallef, N., He, B., Kumar, S., Ahamad, M., & Memon, N. (2020). The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. *International Conference on Big Data*.

Moore, R. C., Dahlke, R., & Hancock, J. T. (2023). Exposure to untrustworthy websites in the 2020 us election. *Nature Human Behaviour*, *7*, 1096–1105.

Moravec, P. L., Kim, A., & Dennis, A. R. (2020). Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research*, *31*(3), 987–1006.

Moravec, P. L., Minas, R. K., & Dennis, A. (2019). Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly*, *43*(4), 1343–1360.

Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (2021). Cognitive reflection correlates with behavior on Twitter. *Nature Communications*, *12*(1), 921.

Mosseri, A. (2016). Addressing hoaxes and fake news.

Ng, K. C., Tang, J., & Lee, D. (2021). The effect of platform intervention policies on fake news dissemination and survival: An empirical examination. *Journal of Management Information Systems*, *38*(4), 898–930.

Nyhan, B., & Reifler, J. (2015). Estimating fact-checking's effects: Evidence from a long-term experiment during campaign 2014.

Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly*, *37*(2), 407–426.

Pan, C. A., Yakhmi, S., Iyer, T. P., Strasnick, E., Zhang, A. X., & Bernstein, M. S. (2022). Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW1), 1–31.

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

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *PNAS*, *116*(7), 2521–2526.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402.

Pilarski, M., Solovev, K., & Pröllochs, N. (2024). Community notes vs. snoping: How the crowd selects fact-checking targets on social media. *ICWSM*.

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.

Poynter. (2019). Most republicans don't trust fact-checkers, and most Americans don't trust the media [Accessed: 2022-04-02].

Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. *ICWSM*.

Racherla, P., Mandviwalla, M., & Connolly, D. J. (2012). Factors affecting consumers' trust in online product reviews. *Journal of Consumer Behaviour*, *11*(2), 94–104.

Rathje, S., Roozenbeek, J., Van Bavel, J. J., & van der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis) information. *Nature Human Behaviour*, *7*(6), 892–903.

Resnick, P., Alfayez, A., Im, J., & Gilbert, E. (2021). Informed crowds can effectively identify misinformation. *arXiv*, (2108.07898). https://arxiv.org/abs/2108.07898

Rocha, Y. M., de Moura, G. A., Desidério, G. A., de Oliveira, C. H., Lourenço, F. D., & de Figueiredo Nicolete, L. D. (2021). The impact of fake news on social media and its influence on health during the covid-19 pandemic: A systematic review. *Journal of Public Health*, *31*, 1007–1016.

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., Van Der Bles, A. M., & Van Der Linden, S. (2020). Susceptibility to misinformation about Covid-19 around the world. *Royal Society Open Science*, *7*(10), 201199.

Schwarz, N., & Jalbert, M. (2020). When (fake) news feels true: Intuitions of truth and the acceptance and correction of misinformation. In *The psychology of fake news* (pp. 73–89). Routledge.

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

Siegrist, M. (2021). Trust and risk perception: A critical review of the literature. *Risk Analysis*, *41*(3), 480–490.

Søe, S. O. (2021). A unified account of information, misinformation, and disinformation. *Synthese*, *198*(6), 5929–5949.

Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. *ICWSM*.

Straub, J., & Spradling, M. (2022). Americans' perspectives on online media warning labels. *Behavioral Sciences*, *12*(3), 59.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*, 99–113.

Traberg, C. S., & van der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personality and Individual Differences*, *185*, 111269.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M., Coleman, K., & Baxter, J. (2022). Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv*. https://arxiv.org/abs/2210.15723

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*(6004), 686–688.

X. (2021). Introducing Birdwatch, a community-based approach to misinformation.

Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., & Patil, S. (2020). Effects of credibility indicators on social media news sharing intent. *CHI*.

Zhao, A., & Naaman, M. (2023). Variety, velocity, veracity, and viability: Evaluating the contributions of crowdsourced and professional fact-checking. *SocArXiv*. https://doi.org/10.31235/osf.io/yfxd3

Zhao, Z., Zhao, J., Sano, Y., Levy, O., Takayasu, H., Takayasu, M., Li, D., Wu, J., & Havlin, S. (2020). Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, *9*(1), 3035.

Appendix 4.A Descriptive statistics

4.A.1 Dependent variables

Figure 4.7 shows the distribution of the participants' responses to questions on trust in fact-checks (Figure 4.7a), misleadingness of posts (Figure 4.7b), and sharing intentions (Figure 4.7c) on 7-point Likert scales.

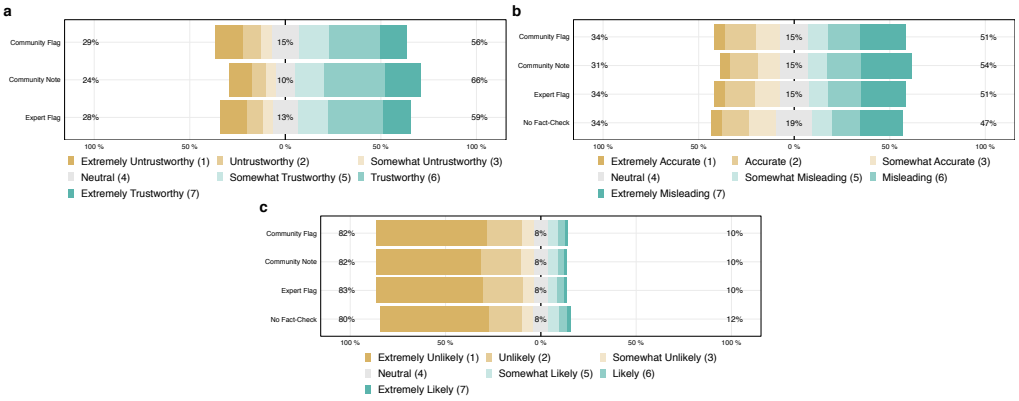


Figure 4.7: Distribution of participants' responses across experimental conditions on 7-point Likert scales. (a) Trustworthiness of the fact-check. (b) Misleadingness of the post. (c) Sharing intentions.

4.A.2 Demographics, beliefs, and CRT

Table 4.1 shows the frequencies of the responses to questions on demographics and beliefs. Evidently, the variables are similarly distributed across the experimental conditions. Overall, 51 % of all participants were female, 48 % male, and 1.4 % identify themselves as non-binary. On average, the participants in our study were 42 years old and the majority of all participants did at least attend college (83 %). Out of all participants, 97 % indicated to be native English speakers, 75 % belong to an ethnic majority, and 74 % are vaccinated against COVID-19. Furthermore, more than half of all participants indicated that they believe in God (54 %). Our sample was approximately balanced across political leanings. Overall, 43 % of participants identified as Democrats and 38 % as Republican (19.7 % as third party or other). During the 2020 presidency election, 48 % indicated that they voted for Biden, whereas 45 % voted for Trump. When being forced to choose between Biden an Trump, 52 % prefer Biden as president and 48 % prefer Trump.

Subjects in our study were asked a series of questions regarding their attitude towards risk, trust, and preference for analytical thinking on a 5-point Likert scale. In the median, survey participants were "undecided" on their attitude towards risk ($Median_{Risk} = 3$) and trust in democracy ($Median_{TrustinDem} = 3$), had "somewhat" trust in general ($Median_{Trust} = 4$), and did "not really" prefer doing something that requires little thought over something that is challenging ($Median_{ThinkingPreference} = 2$).

Subjects in our study were also asked to complete a 4-item Cognitive Reflection Test (CRT). Overall, 43 % passed the CRT, which means that they answered all 4 questions correctly. Notably, this share is comparatively high, pointing towards a high-quality sample of participants from Prolific.

Table 4.1: Descriptive statistics of participants' responses to questions on demographics, beliefs, and CRT. Values are reported as frequencies and percentages, unless otherwise stated.

Variable Participants (n)	Overall 1,810	No Fact-Check 463	Expert Flag 448	Community Flag 422	Community Note 477
Gender					
Female	923 (51%)	252 (54%)	228 (51%)	215 (51%)	228 (48%)
Male	861 (48%)	207 (45%)	215 (48%)	201 (48%)	238 (50%)
Non-binary	26 (1.4%)	4 (0.9%)	5 (1.1%)	6 (1.4%)	11 (2.3%)
Age					
Mean	42.00	42.41	41.62	42.13	41.84
Level of Education					
None	1 (<0.1%)	0 (0%)	1 (0.2%)	0 (0%)	0 (0%)
Less than high school degree	17 (0.9%)	2 (0.4%)	4 (0.9%)	6 (1.4%)	5 (1.0%)
High school diploma	272 (15%)	54 (12%)	70 (16%)	62 (15%)	86 (18%)
Attended college	478 (26%)	114 (25%)	127 (28%)	126 (30%)	111 (23%)
Bachelor's degree	731 (40%)	209 (45%)	177 (40%)	159 (38%)	186 (39%)
Graduate degree	311 (17%)	84 (18%)	69 (15%)	69 (16%)	89 (19%)
Proficiency in English					
Beginner	1 (<0.1%)	0 (0%)	1 (0.2%)	0 (0%)	0 (0%)
Intermediate	7 (0.4%)	3 (0.6%)	2 (0.4%)	0 (0%)	2 (0.4%)
Advanced	55 (3.0%)	12 (2.6%)	16 (3.6%)	16 (3.8%)	11 (2.3%)
Native Speaker	1,747 (97%)	448 (97%)	429 (96%)	406 (96%)	464 (97%)
Belief in God(s)					
I believe in God	972 (54%)	254 (55%)	230 (51%)	228 (54%)	260 (55%)
I don't believe in God	365 (20%)	85 (18%)	96 (21%)	83 (20%)	101 (21%)
I don't know whether or not God exists	330 (18%)	88 (19%)	87 (19%)	74 (18%)	81 (17%)
I don't really take a stance on God	143 (7.9%)	36 (7.8%)	35 (7.8%)	37 (8.8%)	35 (7.3%)
Vaccinated against COVID-19					
Unvaccinated	472 (26%)	119 (26%)	120 (27%)	105 (25%)	128 (27%)
Vaccinated	1,338 (74%)	344 (74%)	328 (73%)	317 (75%)	349 (73%)
Ethnicity					
Ethnic Majority	1,359 (75%)	341 (74%)	328 (73%)	323 (77%)	367 (77%)
Ethnic Minority	451 (25%)	122 (26%)	120 (27%)	99 (23%)	110 (23%)
Political Orientation					
Democrat	778 (43%)	187 (40%)	205 (46%)	176 (42%)	210 (44%)
Republican	683 (38%)	192 (41%)	160 (36%)	146 (35%)	185 (39%)
Third Party	103 (5.7%)	30 (6.5%)	27 (6.0%)	27 (6.4%)	19 (4.0%)
Other	246 (14%)	54 (12%)	56 (13%)	73 (17%)	63 (13%)
Vote in 2020					
Trump	811 (45%)	216 (47%)	198 (44%)	184 (44%)	213 (45%)
Biden	866 (48%)	209 (45%)	225 (50%)	199 (47%)	233 (49%)

Continued on next page

Variable	Overall	No Fact-Check	Expert Flag	Community Flag	Community Note
Other Candidate	17 (0.9%)	5 (1.1%)	1 (0.2%)	7 (1.7%)	4 (0.8%)
I did not vote for reasons outside my control	41 (2.3%)	11 (2.4%)	10 (2.2%)	12 (2.8%)	8 (1.7%)
I did not vote but I could have	64 (3.5%)	19 (4.1%)	12 (2.7%)	17 (4.0%)	16 (3.4%)
I did not vote out of protest	11 (0.6%)	3 (0.6%)	2 (0.4%)	3 (0.7%)	3 (0.6%)
Leaning Biden or Trump					
Biden	935 (52%)	230 (50%)	242 (54%)	217 (51%)	246 (52%)
Trump	875 (48%)	233 (50%)	206 (46%)	205 (49%)	231 (48%)
Willingness to take risks					
Median (IQR)	3.00 (2.00, 4.00)	3.00 (2.00, 4.00)	3.00 (2.00, 4.00)	3.00 (2.00, 4.00)	3.00 (2.00, 4.00)
Trust in general					
Median (IQR)	4.00 (3.00, 4.00)	4.00 (3.00, 4.00)	4.00 (3.00, 4.00)	4.00 (3.00, 4.00)	4.00 (3.00, 4.00)
Trust in Democracy					
Median (IQR)	3.00 (2.00, 4.00)	3.00 (2.00, 4.00)	4.00 (2.00, 4.00)	3.00 (2.00, 4.00)	4.00 (2.00, 4.00)
Thinking Preference					
Median (IQR)	2.00 (2.00, 3.00)	2.00 (2.00, 3.00)	2.00 (2.00, 3.00)	2.00 (2.00, 3.00)	2.00 (2.00, 3.00)
Cognitive Reflection Test					
Failed	1,029 (57%)	281 (61%)	256 (57%)	230 (55%)	262 (55%)
Passed	781 (43%)	182 (39%)	192 (43%)	192 (45%)	215 (45%)
General Reliance on Fact-Checks					
Low	886 (66%)	0 (0%)	297 (66%)	324 (77%)	265 (56%)
High	461 (34%)	0 (0%)	151 (34%)	98 (23%)	212 (44%)

4.A.3 Demographics, beliefs, and CRT of Trump vs. Biden supporters

To assess representativeness beyond gender and political orientation, we looked at how other variables (demographics, beliefs, etc.) differed between participants preferring Trump vs. Biden. Figure 4.8 visualizes the distributions of participants' responses for relevant variables.

We observe that the proportion of female participants is about the same in both groups, pro-Democrats and pro-Republicans (52% vs. 51.5%). At the same time, the share of participants older than 40 years is clearly smaller among pro-Democrats (43.7% vs. 60.5%). Further, we find that participants who favored Biden were slightly less likely to be native English speakers (95.5% vs. 97.5%), more often belonged to ethnic minorities (28.7 % vs. 20.9%), and were less likely to believe in God (35.4% vs. 74%) compared to those who supported Trump. A larger portion of participants leaning towards Biden (vs. those leaning towards Trump) possessed academic degrees (53% vs. 62%), received the COVID-19 vaccine (91.8% vs. 54.8%), and reported a high reliance on fact-checks (42% vs. 26.2%). Additionally, participants preferring Biden were more successful in passing the CRT than those preferring Trump (38.9% vs. 47.4%).

Note that these observations align well with expectations. For instance, analysis of voter demographics in various recent US elections revealed that younger people, ethnic minorities, and individuals holding academic degrees were more likely to support the Democratic Party than the Republican Party, while those identifying as religious were less prevalent among Democratic voters Doherty et al., 2018; Gramlich, 2020. Similarly, studies on attitudes towards COVID-19 vaccination showed that Democrats were more inclined to get vaccinated Pasquini and Saks, 2022; Ye, 2021. Moreover, distrust of fact-checkers was shown to be more widespread among pro-Republicans than among pro-Democrats, as pro-Republicans were more likely to perceive fact-checkers as biased Poynter, 2019.

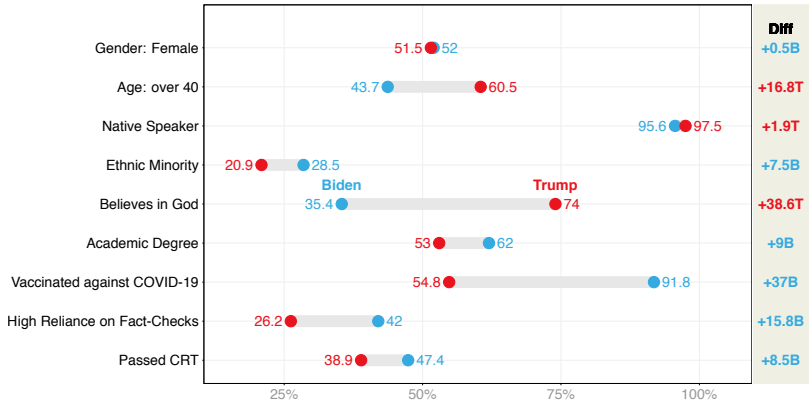


Figure 4.8: Difference in participants' responses to questions on demographics, beliefs, and CRT, depending on their political leaning. The values show the percentage share of pro-Republican/pro-Trump (red) and pro-Democrat/pro-Biden (blue) participants in the respective group. The column "Diff" reports the difference between both groups, "T" and "B" stand for Trump and Biden.

4.A.4 Attrition

Before conducting our analysis, we adjusted the dataset to exclude participants who either did not complete the survey in full, indicated they had responded randomly, engaged in online research during the survey, or indicated they do not use social media. Additionally, those who failed the attention check were removed (see SI, Section 4.D). Table 4.2 provides an overview of the number of participants excluded on the basis of these criteria.

We find that the variation in participant numbers across different conditions post-cleanup (see *Participants (after cleaning)* in Table 4.2) primarily reflects differences in the initial populations of these conditions (see *Participants (before cleaning)* in Table 4.2), i. e., randomness in the assignment via the survey tool. The share of participants excluded on the basis of the individual criteria was fairly consistent, and there were no statistically significant differences in the share of participants who completed the survey across the different conditions (Chi-square test: $\chi^2 = 9$; $P = 0.2133$, two-tailed). Overall, this suggests that attrition did not significantly affect the sizes of the treatment groups in our study.

Table 4.2: Distribution of different exclusion criteria per condition.

Variable	Overall	No Fact-Check	Expert Flag	Community Flag	Community Note
Participants (before cleaning)	2102	530	517	503	552
Participants (after cleaning)	1,810 (86.11%)	463 (87.35%)	448 (86.65%)	422 (83.90%)	477 (86.41%)
Completed Survey					
Yes	2014 (95.81%)	511 (96.42%)	492 (95.16%)	484 (96.22%)	527 (95.47%)
No	88 (4.19%)	19 (3.58%)	25 (4.84%)	19 (3.78%)	25 (4.53%)
Answered Randomly					
Yes	27 (1.19%)	6 (1.13%)	7 (1%)	5 (1%)	6 (0.5%)
No	1989 (94.62%)	505 (95.20%)	485 (94%)	479 (95%)	522 (94.5%)
N/A (didn't complete)	88 (4.19%)	19 (3.58%)	25 (4.84%)	19 (3.78%)	25 (4.53%)
Searched Online					
Yes	26 (1.14%)	5 (0.95%)	6 (1.16%)	5 (1.00%)	10 (1.81%)
No	1990 (94.67%)	506 (95.47%)	486 (94.00%)	479 (95.22%)	517 (93.66%)
N/A (didn't complete)	88 (4.19%)	19 (3.58%)	25 (4.84%)	19 (3.78%)	25 (4.53%)
Social Media Account					
Yes	1966 (93.53%)	500 (94.34%)	476 (92.07%)	473 (94.04%)	517 (93.66%)
No	48 (2.28%)	11 (2.08%)	16 (3.09%)	11 (2.18%)	10 (1.81%)
N/A (didn't complete)	88 (4.19%)	19 (3.58%)	25 (4.84%)	19 (3.78%)	25 (4.53%)
Passed attention check					
Yes	1900 (90.39%)	483 (91.13%)	472 (91.30%)	445 (88.47%)	500 (90.58%)
No	114 (5.42%)	28 (5.29%)	20 (3.86%)	39 (7.75%)	27 (4.89%)
N/A (didn't complete)	88 (4.19%)	19 (3.58%)	25 (4.84%)	19 (3.78%)	25 (4.53%)

4.A.5 Social media behavior

Table 4.3 reports the participants responses to questions regarding their social media behavior. 79 % of all participants use Facebook, 64 % use Twitter/X, and 68 % use Instagram. Prior to all analyses, all participants not using social media platforms were removed. Participants indicated that they are more likely to share content on political (41 %) and scientific (55 %) topics than on other topics. Out of all participants, 19 % stated that they do not share any content on social media. Participants where also asked to rate whether specific post characteristics are important for them when deciding about sharing a post. While accuracy and interestingness were extremely or very important to a majority of participants (91 % and 81 %), it was less important whether a post is funny (41 %) or surprising (22.5 %). Moreover, most participants indicated that it is at least moderately important for them that a post aligns with their beliefs (83 %). The participants' ratings on the importance of specific post characteristics did not differ drastically across different social media platforms and topics.

Table 4.3: Percentage of participants reporting having a social media account and type of content shared. Rows show participants' ratings of whether specific post characteristics are important for them when deciding about sharing a post.

	Social media outlets used							Type of content shared						
Variable	**Facebook**	**Twitter**	**Snapchat**	**Instagram**	**WhatsApp**	**TikTok**	**Other**	**Politics**	**Sports**	**Celebrities**	**Science**	**Business**	**Other**	**None**
Participants (n)	1,441	1,156	486	1,227	346	721	423	741	582	454	989	479	455	344
Post is Accurate														
Not at all	2.2%	1.5%	1.4%	2.0%	1.4%	1.1%	3.3%	0.5%	0.3%	0.7%	0.5%	0.2%	0.9%	8.4%
Slightly	0.9%	1.4%	1.2%	1.4%	0.3%	1.7%	0.7%	1.2%	1.4%	2.2%	1.4%	1.3%	1.3%	1.5%
Moderately	6.5%	7.8%	8.4%	7.7%	6.9%	7.6%	5.9%	8.1%	7.6%	8.1%	7.0%	7.5%	7.9%	5.2%
Very	26%	29%	29%	27%	27%	27%	25%	31%	32%	32%	30%	32%	26%	16%
Extremely	65%	61%	60%	62%	65%	63%	65%	59%	58%	57%	61%	59%	64%	69%
Post is Surprising														
Not at all	26%	24%	22%	25%	21%	22%	30%	18%	18%	12%	19%	18%	31%	46%
Slightly	22%	24%	26%	22%	16%	24%	24%	24%	23%	21%	24%	22%	27%	18%
Moderately	30%	32%	28%	31%	32%	31%	30%	32%	33%	33%	34%	32%	29%	22%
Very	16%	15%	17%	15%	22%	17%	13%	19%	21%	26%	17%	22%	9.5%	6.4%
Extremely	6.5%	5.5%	7.4%	6.2%	9.5%	6.4%	3.5%	6.6%	5.7%	7.5%	5.7%	6.1%	3.3%	8.4%
Post is Interesting														
Not at all	2.8%	1.6%	1.6%	2.3%	4.0%	1.2%	4.7%	0.3%	0.5%	0.2%	0.3%	0%	1.8%	12%
Slightly	2.2%	2.2%	1.9%	2.8%	2.0%	2.5%	3.5%	2.2%	1.7%	1.8%	1.9%	1.5%	2.2%	3.5%
Moderately	13%	14%	17%	14%	14%	13%	13%	13%	13%	10%	13%	14%	15%	14%
Very	43%	44%	42%	43%	43%	44%	43%	47%	47%	47%	46%	47%	45%	30%
Extremely	39%	38%	37%	38%	37%	40%	35%	38%	38%	41%	39%	38%	37%	40%

Continued on next page

Table 4.3: (continued)

	Social media outlets used							Type of content shared						
Variable	**Facebook**	**Twitter**	**Snapchat**	**Instagram**	**WhatsApp**	**TikTok**	**Other**	**Politics**	**Sports**	**Celebrities**	**Science**	**Business**	**Other**	**None**
Post is Aligned to the User's Beliefs														
Not at all	8.1%	8.2%	6.8%	8.1%	7.2%	6.4%	11%	4.7%	4.1%	4.8%	6.2%	4.6%	7.0%	19%
Slightly	8.1%	10%	9.3%	9.4%	6.6%	8.9%	11%	9.3%	12%	9.0%	12%	11%	8.6%	7.8%
Moderately	23%	24%	23%	22%	21%	23%	26%	26%	27%	25%	27%	31%	25%	15%
Very	30%	30%	32%	31%	33%	30%	30%	32%	32%	33%	30%	32%	31%	23%
Extremely	30%	27%	29%	30%	32%	31%	22%	28%	25%	29%	25%	23%	28%	34%
Post is Funny														
Not at all	11%	10%	8.4%	10%	6.9%	7.9%	13%	9.9%	6.7%	5.7%	8.3%	9.2%	9.9%	18%
Slightly	16%	18%	15%	18%	18%	15%	18%	20%	17%	15%	20%	19%	16%	13%
Moderately	31%	32%	30%	31%	29%	31%	34%	34%	33%	31%	33%	33%	35%	26%
Very	22%	22%	28%	22%	27%	25%	20%	23%	26%	25%	23%	24%	19%	18%
Extremely	19%	18%	19%	19%	19%	20%	15%	14%	17%	23%	16%	15%	20%	24%

4.A.6 Perception of fact-checks

Table 4.4 reports the participants awareness of fact-checks and the influence the fact-checks had on their responses during the survey. Among participants in conditions 2 – 4, a vast majority (75 %) was aware of fact-checking prior to participating. However, awareness was substantially higher for expert fact-checks (92 %) than for community flags (65 %) and community notes (66 %). In general, participants' reliance on fact-checks differed a lot, but was similarly distributed across all conditions. Overall, 34 % indicated to have a high or extreme reliance on fact-checks.

Furthermore, participants were asked to indicate what influence the presented fact-check had on their discernment of posts with and without fact-checks during the survey. Overall, the majority of participants stated the the presence of fact-checks made them rate potentially misleading post as less accurate (52 %) or the tag had not influence (20 %). Posts without fact-checks (i. e., that were not potentially misleading) have been rated as more accurate (38.5 %) or the tag had no influence (43 %). Interestingly, the share of participants in condition 4 (community note) indicating that they rated posts with a community note (slightly) more accurate was much higher than in the other conditions (44 % versus 14.5 % and 14.3 %).

Table 4.4: Participants' perception of fact-checks.

Variable **Participants (n)**	**Overall** 1,810	**No Fact-Check** 463	**Expert Flag** 448	**Community Flag** 422	**Community Note** 477
Aware of Fact-Checking (Prior to Survey)					
Not Aware of Fact-Checking	345 (26%)	–	37 (8.3%)	146 (35%)	162 (34%)
Aware of Fact-Checking	1,002 (74%)	–	411 (92%)	276 (65%)	315 (66%)
Reliance on Fact-Checks					
None (1)	193 (14%)	–	75 (17%)	76 (18%)	42 (8.8%)
Slight (2)	272 (20%)	–	94 (21%)	110 (26%)	68 (14%)
Moderate (3)	421 (31%)	–	128 (29%)	138 (33%)	155 (32%)
High (4)	328 (24%)	–	114 (25%)	67 (16%)	147 (31%)
Extreme (5)	133 (9.9%)	–	37 (8.3%)	31 (7.3%)	65 (14%)
Influence of Fact-Checks (Misleading Posts)					
Much less accurate (1)	223 (17%)	–	105 (23%)	61 (14%)	57 (12%)
Less accurate (2)	376 (28%)	–	148 (33%)	140 (33%)	88 (18%)
Slightly less accurate (3)	148 (11%)	–	52 (12%)	68 (16%)	28 (5.9%)
Tag had no influence (4)	274 (20%)	–	87 (19%)	97 (23%)	90 (19%)
Slightly more accurate (5)	179 (13%)	–	31 (6.9%)	37 (8.8%)	111 (23%)
More accurate (6)	118 (8.8%)	–	20 (4.5%)	16 (3.8%)	82 (17%)
Much more accurate (7)	29 (2.2%)	–	5 (1.1%)	3 (0.7%)	21 (4.4%)
Influence of Fact-Checks (Non-misleading Posts)					
Much less accurate (1)	36 (2.7%)	–	6 (1.3%)	14 (3.3%)	16 (3.4%)
Less accurate (2)	100 (7.4%)	–	16 (3.6%)	30 (7.1%)	54 (11%)
Slightly less accurate (3)	116 (8.6%)	–	23 (5.1%)	26 (6.2%)	67 (14%)
Tag had no influence (4)	575 (43%)	–	183 (41%)	183 (43%)	209 (44%)
Slightly more accurate (5)	366 (27%)	–	161 (36%)	118 (28%)	87 (18%)
More accurate (6)	122 (9.1%)	–	50 (11%)	38 (9.0%)	34 (7.1%)
Much more accurate (7)	32 (2.4%)	–	9 (2.0%)	13 (3.1%)	10 (2.1%)

Appendix 4.B Estimation results

To predict the participant's (i) trust in fact-checks and (ii) misleadingness ratings, we implemented a hierarchical linear regression model with random intercepts for posts and subjects. Average marginal effects (AMEs) and coefficient estimates are reported in Tables 4.5 and 4.7 (trust in fact-checks) and Tables 4.6 and 4.8 (misleadingness). The findings are described in detail in Section Results of the main paper.

Table 4.5: Average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with three-way interaction terms predicting the trustworthiness of a fact-check (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in the average predicted trustworthiness ratings between the group of interest and the reference group (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in trustworthiness ratings). The indented rows report the AMEs of an intervention depending on the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 24\,003$ observations across 1 347 participants. Coefficient estimates are in SI, Table 4.7.

	AME	Lower CI	Upper CI	*P*-value
Community Note [*ref.: Expert Flag*]	0.048	0.042	0.055	< 0.001
Leaning Biden	0.061	0.052	0.069	< 0.001
Leaning Biden, *Concordant*	0.064	0.049	0.080	< 0.001
Leaning Biden, *Neutral*	0.074	0.061	0.086	< 0.001
Leaning Biden, *Discordant*	0.045	0.032	0.058	< 0.001
Leaning Trump	0.035	0.026	0.045	< 0.001
Leaning Trump, *Concordant*	0.023	0.007	0.039	0.010
Leaning Trump, *Neutral*	0.042	0.025	0.059	< 0.001
Leaning Trump, *Discordant*	0.040	0.024	0.056	< 0.001
Community Flag [*ref.: Expert Flag*]	−0.013	−0.020	−0.006	0.003
Leaning Biden	−0.010	−0.019	0.000	0.045
Leaning Biden, *Concordant*	−0.008	−0.027	0.009	0.391
Leaning Biden, *Neutral*	−0.010	−0.024	0.006	0.188
Leaning Biden, *Discordant*	−0.011	−0.026	0.003	0.155
Leaning Trump	−0.016	−0.026	−0.006	0.006
Leaning Trump, *Concordant*	−0.027	−0.044	−0.010	0.001
Leaning Trump, *Neutral*	−0.007	−0.025	0.011	0.441
Leaning Trump, *Discordant*	−0.014	−0.031	0.004	0.118
Leaning Trump [*ref.: Leaning Biden*]	−0.084	−0.089	−0.078	0.001
Concordant [*ref.: Neutral*]	−0.023	−0.029	−0.017	0.001
Leaning Biden	−0.034	−0.043	−0.026	0.001
Leaning Trump	−0.010	−0.020	0.000	0.040
Discordant [*ref.: Neutral*]	0.022	0.016	0.028	< 0.001
Leaning Biden	0.024	0.016	0.032	< 0.001
Leaning Trump	0.019	0.009	0.028	< 0.001

Table 4.6: Average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with four-way interaction terms predicting the perceived misleadingness of a post (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in the average predicted misleadingness ratings between the group of interest and the reference group (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in misleadingness ratings). AMEs are reported separately for misleading (columns 2–5) and non-misleading posts (columns 6–9). The indented rows report the AMEs of an intervention depending on the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 64\,454$ observations across 1 810 participants. Coefficient estimates are in SI, Table 4.8.

	Misleading				**Non-misleading**			
	AME	**Lower CI**	**Upper CI**	***P*-value**	**AME**	**Lower CI**	**Upper CI**	***P*-value**
Expert Flag [*ref.: No Fact-Check*]	0.071	0.063	0.078	< 0.001	−0.052	−0.059	−0.044	0.001
Leaning Biden	0.085	0.076	0.094	< 0.001	−0.057	−0.067	−0.048	0.001
Leaning Trump	0.055	0.044	0.067	< 0.001	−0.046	−0.056	−0.035	0.001
Community Flag [*ref.: No Fact-Check*]	0.060	0.053	0.067	< 0.001	−0.051	−0.058	−0.043	0.001
Leaning Biden	0.064	0.054	0.074	< 0.001	−0.053	−0.063	−0.044	0.001
Leaning Trump	0.055	0.044	0.066	< 0.001	−0.047	−0.058	−0.036	0.001
Community Note [*ref.: No Fact-Check*]	0.096	0.089	0.102	< 0.001	−0.032	−0.039	−0.025	0.001
Leaning Biden	0.110	0.100	0.120	< 0.001	−0.038	−0.048	−0.029	0.001
Leaning Trump	0.081	0.070	0.091	< 0.001	−0.024	−0.035	−0.014	0.001
Community Note [*ref.: Expert Flag*]	0.025	0.019	0.032	< 0.001	0.020	0.013	0.027	< 0.001
Leaning Biden, *Concordant*	0.024	0.007	0.039	0.009	0.014	−0.001	0.029	0.060
Leaning Biden, *Neutral*	0.045	0.031	0.059	< 0.001	0.016	0.001	0.032	0.042
Leaning Biden, *Discordant*	0.005	−0.007	0.016	0.439	0.027	0.008	0.046	0.003
Leaning Trump, *Concordant*	0.045	0.027	0.065	< 0.001	0.035	0.016	0.053	< 0.001
Leaning Trump, *Neutral*	0.044	0.026	0.062	< 0.001	0.019	0.003	0.036	0.020
Leaning Trump, *Discordant*	−0.012	−0.027	0.003	0.112	0.009	−0.011	0.029	0.385
Community Flag [*ref.: Expert Flag*]	−0.011	−0.018	−0.004	0.003	0.001	−0.006	0.008	0.759
Leaning Biden, *Concordant*	−0.043	−0.061	−0.026	0.001	0.008	−0.010	0.024	0.334
Leaning Biden, *Neutral*	−0.017	−0.032	−0.001	0.033	−0.003	−0.017	0.011	0.721
Leaning Biden, *Discordant*	−0.003	−0.015	0.009	0.628	0.007	−0.012	0.025	0.459
Leaning Trump, *Concordant*	0.007	−0.012	0.028	0.532	−0.002	−0.021	0.018	0.876
Leaning Trump, *Neutral*	−0.004	−0.023	0.013	0.637	−0.009	−0.027	0.007	0.248
Leaning Trump, *Discordant*	−0.002	−0.018	0.012	0.780	0.007	−0.015	0.026	0.508
Leaning Trump [*ref.: Leaning Biden*]	−0.016	−0.021	−0.011	0.001	0.034	0.029	0.039	< 0.001
Concordant [*ref.: Neutral*]	−0.057	−0.064	−0.051	0.001	0.041	0.036	0.047	< 0.001

Continued on next page

	Misleading				Non-misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**	**AME**	**Lower CI**	**Upper CI**	***P*-value**
Leaning Biden	−0.086	−0.094	−0.078	0.001	0.003	−0.004	0.011	0.384
Leaning Trump	−0.027	−0.036	−0.017	0.001	0.082	0.072	0.091	< 0.001
Discordant [*ref.: Neutral*]	0.092	0.087	0.097	< 0.001	0.168	0.162	0.174	< 0.001
Leaning Biden	0.091	0.084	0.098	< 0.001	0.193	0.184	0.201	< 0.001
Leaning Trump	0.093	0.085	0.101	< 0.001	0.141	0.132	0.151	< 0.001

Table 4.7: Estimation results for a hierarchical linear regression model with three-way interaction terms predicting the trustworthiness of a fact-check (7-point Likert scale normalized to the interval [0, 1]). Random intercepts for posts and subjects are included. $N = 24\,003$ observations across 1 347 participants.

Dependent Variable: Trustworthiness					
Variable	**Coef.**	**Std. Error**	**Lower CI**	**Upper CI**	***P*-value**
Condition					
Misleading (Expert Flag) [*ref.*]	–	–	–	–	–
Misleading (Community Flag)	−0.010	0.025	−0.059	0.039	0.685
Misleading (Community Note)	0.074**	0.024	0.027	0.121	0.002
Political Leaning (Subject)					
Biden [*ref.*]	–	–	–	–	–
Trump	−0.080**	0.025	−0.129	−0.030	0.002
Political Concordance (Post)					
Neutral [*ref.*]	–	–	–	–	–
Concordant	−0.031	0.018	−0.067	0.004	0.083
Discordant	0.035	0.018	−0.001	0.070	0.054
Interactions					
Concordant × Trump	0.035	0.020	−0.004	0.074	0.081
Discordant × Trump	−0.013	0.020	−0.052	0.026	0.527
Misleading (Community Flag) × Concordant	0.001	0.011	−0.021	0.024	0.897
Misleading (Community Note) × Concordant	−0.010	0.011	−0.031	0.012	0.373
Misleading (Community Flag) × Discordant	−0.001	0.011	−0.023	0.021	0.955
Misleading (Community Note) × Discordant	−0.029**	0.011	−0.051	−0.008	0.007
Misleading (Community Flag) × Trump	0.003	0.036	−0.068	0.074	0.940
Misleading (Community Note) × Trump	−0.032	0.035	−0.100	0.037	0.366
Misleading (Community Flag) × Concordant × Trump	−0.021	0.016	−0.053	0.011	0.201
Misleading (Community Note) × Concordant × Trump	−0.010	0.016	−0.041	0.022	0.543
Misleading (Community Flag) × Discordant × Trump	−0.006	0.016	−0.038	0.026	0.725
Misleading (Community Note) × Discordant × Trump	0.027	0.016	−0.004	0.059	0.086
Intercept	0.619***	0.021	0.579	0.659	<0.001
Subject-level RE	YES				
Post-level RE	YES				
AIC	−2 710.56				
Participants (n)	1 347				
Observations (N)	24 003				

Significance: $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4.8: Estimation results for hierarchical linear regression model with four-way interaction terms predicting the perceived misleadingness of a post (7-point Likert scale normalized to the interval [0, 1]). Random intercepts for posts and subjects are included. $N = 64\,454$ observations across 1 810 participants.

Dependent Variable: Misleadingness					
Variable	**Coef.**	**Std. Error**	**Lower CI**	**Upper CI**	***P*-value**
Condition					
No Fact-Check [*ref.*]	–	–	–	–	–
Expert Flag	0.101***	0.010	0.080	0.121	<0.001
Community Flag	0.084***	0.011	0.063	0.105	<0.001
Community Note	0.146***	0.010	0.125	0.166	<0.001
Political Leaning (Subject)					
Biden [*ref.*]	–	–	–	–	–
Trump	−0.017	0.011	−0.038	0.003	0.101
Political Concordance (Post)					
Neutral [*ref.*]	–	–	–	–	–
Concordant	−0.077	0.050	−0.175	0.021	0.122
Discordant	0.138**	0.050	0.040	0.237	0.006
Misleading (Post)					
Misleading [*ref.*]	–	–	–	–	–
Non-misleading	−0.363***	0.050	−0.461	−0.265	<0.001
Interactions					
Concordant × Trump	0.066	0.051	−0.034	0.166	0.195
Discordant × Trump	−0.010	0.051	−0.110	0.089	0.836
Expert Flag × Concordant	0.004	0.012	−0.020	0.028	0.735
Community Flag × Concordant	−0.022	0.012	−0.047	0.002	0.072
Community Note × Concordant	−0.017	0.012	−0.041	0.006	0.152
Expert Flag × Discordant	−0.052***	0.013	−0.077	−0.028	<0.001
Community Flag × Discordant	−0.039**	0.013	−0.064	−0.014	0.003
Community Note × Discordant	−0.092***	0.012	−0.117	−0.068	<0.001
Expert Flag × Trump	−0.028	0.015	−0.058	0.001	0.059
Community Flag × Trump	−0.016	0.015	−0.046	0.014	0.293
Community Note × Trump	−0.029*	0.015	−0.058	0.000	0.046
Concordant × Non-misleading	0.089	0.071	−0.049	0.228	0.206
Discordant × Non-misleading	0.073	0.071	−0.066	0.212	0.304
Trump × Non-misleading	0.029*	0.012	0.005	0.053	0.018
Expert Flag × Non-misleading	−0.143***	0.012	−0.167	−0.119	<0.001
Community Flag × Non-misleading	−0.129***	0.012	−0.154	−0.105	<0.001
Community Note × Non-misleading	−0.172***	0.012	−0.195	−0.148	<0.001
Expert Flag × Concordant × Trump	−0.029	0.018	−0.064	0.006	0.110
Community Flag × Concordant × Trump	0.010	0.018	−0.026	0.045	0.598
Community Note × Concordant × Trump	−0.006	0.017	−0.041	0.028	0.717
Expert Flag × Discordant × Trump	0.025	0.018	−0.010	0.060	0.159
Community Flag × Discordant × Trump	0.013	0.018	−0.022	0.048	0.468
Community Note × Discordant × Trump	0.009	0.017	−0.025	0.043	0.605
Concordant × Trump × Non-misleading	0.023	0.072	−0.118	0.164	0.747
Discordant × Trump × Non-misleading	−0.035	0.072	−0.176	0.105	0.622
Expert Flag × Concordant × Non-misleading	−0.018	0.017	−0.052	0.016	0.302
Community Flag × Concordant × Non-misleading	0.019	0.018	−0.015	0.054	0.277
Community Note × Concordant × Non-misleading	0.001	0.017	−0.032	0.035	0.950
Expert Flag × Discordant × Non-misleading	0.021	0.017	−0.013	0.056	0.219
Community Flag × Discordant × Non-misleading	0.017	0.018	−0.018	0.052	0.348
Community Note × Discordant × Non-misleading	0.071***	0.017	0.037	0.106	<0.001
Expert Flag × Trump × Non-misleading	0.048**	0.018	0.014	0.082	0.006
Community Flag × Trump × Non-misleading	0.029	0.018	−0.006	0.064	0.100
Community Note × Trump × Non-misleading	0.052**	0.017	0.018	0.085	0.003
Expert Flag × Concordant × Trump × Non-misleading	0.007	0.025	−0.042	0.056	0.772
Community Flag × Concordant × Trump × Non-misleading	−0.033	0.025	−0.083	0.016	0.187
Community Note × Concordant × Trump × Non-misleading	0.004	0.025	−0.044	0.052	0.871
Expert Flag × Discordant × Trump × Non-misleading	−0.029	0.025	−0.078	0.020	0.247
Community Flag × Discordant × Trump × Non-misleading	−0.010	0.025	−0.060	0.039	0.687
Community Note × Discordant × Trump × Non-misleading	−0.032	0.025	−0.080	0.016	0.188

Continued on next page

Table 4.8: (continued)

Variable	Coef.	Std. Error	Lower CI	Upper CI	*P*-value
Intercept	0.712***	0.036	0.642	0.782	<0.001
Subject-level RE	YES				
Post-level RE	YES				
AIC	−4 352.41				
Participants (n)	1 810				
Observations (N)	64 454				

Standard errors are in parentheses; $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Appendix 4.C Additional analyses

4.C.1 Sharing intentions

Note: Participants in our survey also had to state whether they would consider sharing the presented posts on social media. However, research shows that asking about misleadingness before assessing sharing intentions can influence the outcome variable, that is, increase correlation between the identification of misleadingness and sharing intentions Epstein et al., 2023. The following results should therefore be interpreted with caution.

In our survey, participants assigned to the control condition had significantly lower (t-test: $t = -36.11$; $df = 13395$; $P < 0.001$, two-tailed) sharing intentions (7-point Likert scale normalized to the interval [0, 1]) for misleading ($M = 0.13$) than for non-misleading posts ($M = 0.23$). To quantify intervention effects, we fitted a hierarchical linear regression model with four-way interaction terms and random intercepts for subjects and posts to predict sharing intentions (see Section Methods of the main paper). Figure 4.9 shows the average marginal effects (see SI, Table 4.9 for AMEs).

All fact-checking interventions significantly reduced sharing intentions for misleading posts (see Figure 4.9a,c,d). Compared to the control condition, sharing intentions for misleading posts were, on average, 2.7 percentage points lower for participants exposed to expert flags (AME $= -0.027$; 95 % CI $= [-0.033, -0.021]$; $P = 0.001$), 3.5 percentage points lower for community flags (AME $= -0.035$; 95 % CI $= [-0.041, -0.029]$; $P = 0.001$), and 1.3 percentage points lower for community notes (AME $= -0.013$; 95 % CI $= [-0.018, -0.007]$; $P = 0.001$). In terms of percentages, these numbers translate to a reduction in sharing intentions of 21.3 % for expert flags, 27.6 % for community flags, and 10.3 % for community notes. Hence, compared to expert flags, community flags decreased sharing intentions for misleading posts (AME $= -0.008$; 95 % CI $= [-0.014, -0.003]$; $P = 0.001$), while community notes led to an increase (AME $= 0.014$; 95 % CI $= [0.014, 0.019]$; $P < 0.001$). The efficacy of the fact-checking interventions in reducing sharing intentions for misleading posts did not vary significantly for participants leaning towards Biden vs. Trump ($P = 0.817$ for expert flags, $P = 0.671$ for community flags, $P = 0.486$ for community notes). We further observe that Trump supporters had significantly higher baseline intentions to share misleading posts than Biden supporters (AME $= 0.018$; 95 % CI $= [0.014, 0.022]$; $P < 0.001$). Also, sharing intentions for misleading posts were, on average, higher for politically concordant posts (AME $= 0.022$; 95 % CI $= [0.017, 0.027]$; $P < 0.001$) than for politically discordant posts (AME $= -0.033$; 95 % CI $= [-0.038, -0.029]$; $P = 0.001$).

Similar to the previous analyses, the efficacy of community notes in reducing sharing intentions for misleading posts varied across political leanings (see Figure 4.9c). For Trump supporters, replacing an expert flag with a community note would have decreased sharing intentions for misleading posts by 2.2 percentage points if they were politically concordant (AME $= -0.022$; 95 % CI $= [-0.037, -0.008]$; $P = 0.001$). This effect was slightly positive for politically neutral (AME $= 0.015$; 95 % CI $= [0.002, 0.029]$; $P = 0.029$) and discordant posts (AME $= 0.035$; 95 % CI $= [0.022, 0.047]$; $P < 0.001$). We find a similar pattern for Biden supporters. Here, replacing an expert flag with a community note would have increased sharing intentions by 1.4 percentage points for politically discordant posts (AME $= 0.031$; 95 % CI $= [0.021, 0.042]$; $P < 0.001$), whereas the effects were smaller for politically neutral

(AME $= 0.014$; 95 % CI $= [0.002, 0.026]$; $P = 0.018$) and not statistically significant for politically concordant posts (AME $= 0.010$; 95 % CI $= [-0.002, 0.023]$; $P = 0.100$).

We also observe some treatment condition effects on untagged non-misleading posts (see Figure 4.9b). For Biden supporters, displaying fact-checks on misleading posts increased sharing intentions for untagged non-misleading posts by 1.4 percentage points for expert flags (AME $= 0.012$; 95 % CI $= [0.006, 0.023]$; $P = 0.003$), by 2.7 percentage points for community flags (AME $= 0.027$; 95 % CI $= [0.018, 0.036]$; $P < 0.001$), and by 1.0 percentage points for community notes (AME $= 0.010$; 95 % CI $= [0.002, 0.019]$; $P = 0.020$). For Trump supporters, we do not observe significant effects across any of the conditions ($P = 0.081$ for expert flags, $P = 0.447$ for community flags, $P = 0.218$ for community notes). When comparing the average marginal effects across all participants (i. e., both Biden and Trump supporters), the treatment condition effects on untagged non-misleading posts were statistically significant for community flags (AME $= 0.012$; 95 % CI $= [0.005, 0.009]$; $P < 0.001$), but not for expert flags (AME $= 0.003$; 95 % CI $= [-0.004, 0.009]$; $P = 0.340$) and community notes (AME $= 0.002$; 95 % CI $= [-0.004, 0.008]$; $P = 0.511$). Furthermore, we find that Trump supporters had slightly higher baseline intentions to share non-misleading posts than Biden supporters (AME $= 0.006$; 95 % CI $= [0.002, 0.011]$; $P = 0.005$), and sharing intentions for non-misleading posts were higher for politically concordant posts (AME $= 0.028$; 95 % CI $= [0.021, 0.033]$; $P < 0.001$) than for politically discordant posts (AME $= -0.072$; 95 % CI $= [-0.078, -0.067]$; $P = 0.001$).

In sum, these results suggest that community notes were less successful in reducing sharing intentions for misleading posts than simple misinformation flags. In particular, politically discordant misleading posts with community notes were more likely to be shared (by both Biden and Trump supporters) relative to discordant posts with simple misinformation flags. A potential explanation might be that the community notes themselves were actually perceived as politically concordant (because they were applied to politically discordant original posts). Thus, participants may have been inclined to share the post because they were sharing the community note appended to it, rather than endorsing the original post.

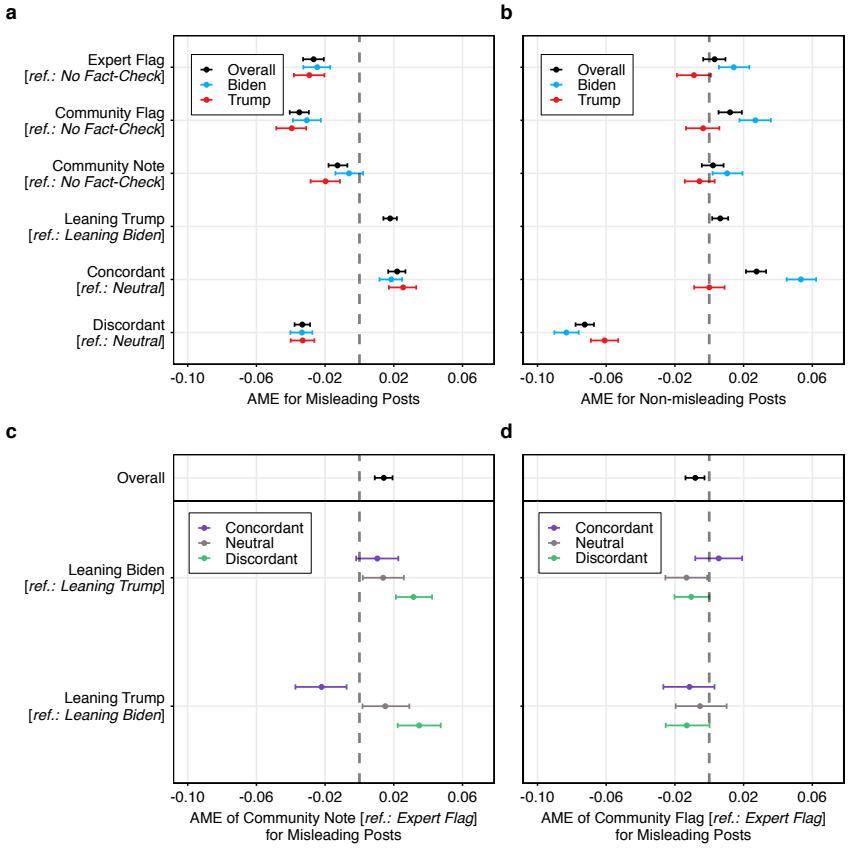


Figure 4.9: Community notes did not consistently reduce sharing intentions for misleading posts. Shown are the average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with interaction terms predicting sharing intentions (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in the average predicted sharing intentions between the group of interest and the reference group (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in sharing intentions). **(a)** AMEs for misleading posts. **(b)** AMEs for non-misleading posts. **(c)** AMEs when replacing an expert flag with a community note (i. e., the average difference between the predicted marginal effects for community notes vs. expert flags) across the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). **(d)** AMEs when replacing an expert flag with a community flag (i. e., the average difference between the predicted marginal effects for community flags vs. expert flags) across the political leanings of participants and the fact-checked posts. Control variables and random intercepts for posts and subjects were included. The 95 % confidence intervals (error bars) were derived using the bootstrap method for 1 000 resamples. $N = 64\,454$ observations across 1 810 participants. Full estimation results are in SI, Table 4.9.

Table 4.9: Average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with four-way interaction terms predicting sharing intentions (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in average predicted sharing intentions between the group of interest and the reference group. AMEs are reported separately for misleading and non-misleading posts. The indented rows report the AMEs depending on political leaning and congruence. Random intercepts and bootstrap-derived confidence intervals. $N = 64\,454$ observations across 1 810 participants.

Condition	Subgroup	Congruence	Misleading				Non-misleading			
			AME	Lower CI	Upper CI	P-value	AME	Lower CI	Upper CI	P-value
Expert Flag	All	–	−0.027	−0.033	−0.021	0.001	0.003	−0.004	0.009	0.340
Expert Flag	Leaning Biden	–	−0.025	−0.033	−0.017	0.001	0.014	0.006	0.023	0.003
Expert Flag	Leaning Trump	–	−0.029	−0.038	−0.020	0.001	−0.009	−0.019	0.001	0.081
Community Flag	All	–	−0.035	−0.041	−0.029	0.001	0.012	0.005	0.019	<0.001
Community Flag	Leaning Biden	–	−0.031	−0.039	−0.022	0.001	0.027	0.018	0.036	<0.001
Community Flag	Leaning Trump	–	−0.039	−0.048	−0.031	0.001	−0.004	−0.014	0.006	0.447
Community Note	All	–	−0.013	−0.018	−0.007	0.001	0.002	−0.004	0.008	0.511
Community Note	Leaning Biden	–	−0.006	−0.014	0.002	0.135	0.010	0.002	0.019	0.020
Community Note	Leaning Trump	–	−0.020	−0.028	−0.011	0.001	−0.006	−0.014	0.003	0.218
Community Note vs Expert Flag	All	–	0.014	0.009	0.019	<0.001	−0.001	−0.007	0.005	0.789
Community Note vs Expert Flag	Leaning Biden	Concordant	0.010	−0.002	0.023	0.100	−0.011	−0.027	0.006	0.188
Community Note vs Expert Flag	Leaning Biden	Neutral	0.014	0.002	0.026	0.018	0.006	−0.010	0.021	0.441
Community Note vs Expert Flag	Leaning Biden	Discordant	0.031	0.021	0.042	<0.001	−0.007	−0.019	0.006	0.276
Community Note vs Expert Flag	Leaning Trump	Concordant	−0.022	−0.037	−0.008	0.001	0.001	−0.017	0.018	0.858
Community Note vs Expert Flag	Leaning Trump	Neutral	0.015	0.002	0.029	0.029	0.002	−0.015	0.019	0.833
Community Note vs Expert Flag	Leaning Trump	Discordant	0.035	0.022	0.047	<0.001	0.006	−0.010	0.021	0.417
Community Flag vs Expert Flag	All	–	−0.008	−0.014	−0.003	0.001	0.009	0.002	0.016	0.010
Community Flag vs Expert Flag	Leaning Biden	Concordant	0.006	−0.008	0.019	0.442	0.023	0.006	0.041	0.010
Community Flag vs Expert Flag	Leaning Biden	Neutral	−0.013	−0.026	−0.001	0.037	0.023	0.008	0.039	0.004
Community Flag vs Expert Flag	Leaning Biden	Discordant	−0.010	−0.020	<0.001	0.053	−0.008	−0.021	0.005	0.220
Community Flag vs Expert Flag	Leaning Trump	Concordant	−0.012	−0.027	0.003	0.140	0.013	−0.005	0.033	0.150
Community Flag vs Expert Flag	Leaning Trump	Neutral	−0.005	−0.020	0.010	0.445	0.007	−0.009	0.026	0.391
Community Flag vs Expert Flag	Leaning Trump	Discordant	−0.013	−0.025	<0.001	0.053	−0.004	−0.019	0.011	0.598
Leaning Trump vs Biden	All	–	0.018	0.014	0.022	<0.001	0.006	0.002	0.011	0.005
Concordant vs Neutral	All	–	0.022	0.017	0.027	<0.001	0.028	0.021	0.033	<0.001
Concordant vs Neutral	Leaning Biden	–	0.018	0.012	0.025	<0.001	0.053	0.045	0.062	<0.001
Concordant vs Neutral	Leaning Trump	–	0.025	0.017	0.033	<0.001	<0.001	−0.009	0.009	0.986
Discordant vs Neutral	All	–	−0.033	−0.038	−0.029	0.001	−0.072	−0.078	−0.067	0.001
Discordant vs Neutral	Leaning Biden	–	−0.033	−0.040	−0.027	0.001	−0.083	−0.090	−0.076	0.001
Discordant vs Neutral	Leaning Trump	–	−0.033	−0.040	−0.026	0.001	−0.061	−0.069	−0.053	0.001

4.C.2 Demographics and beliefs

We repeated our regression analyses including information on the participants' demographics and beliefs (see Section Demographics, beliefs, and cognitive reflection in the main paper). This included the participants' gender (male, female, non-binary), level of education (attended college vs. didn't attend college), stance towards God (believes in God vs. doesn't believe in God), ethnicity (ethnic minority vs. not a minority), and COVID-19 vaccination status (vaccinated vs. not vaccinated). In addition, we included the following variables regarding the participants attitude: willingness to take risks (low vs. high), trust in people they interact with in their daily life (low vs. high), trust in democracy (low vs. high), and preference to perform tasks that require thinking (low thinking preference vs. high thinking preference). For all Likert-scale variables, a value of greater than "undecided" (3) was considered high (otherwise low).

Table 4.10 reports the AMEs for misleadingness ratings under the control condition. We further studied moderation effects, i. e., how the efficacy of fact-checking interventions varied depending on demographics and beliefs. The corresponding AMEs are reported in Tables 4.11 to 4.13.

Table 4.10: AMEs of demographics and beliefs on perceived misleadingness under the control condition (No Fact-Check) for misleading posts. The AMEs are calculated from hierarchical linear regression models with interaction terms. Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 64\,454$ observations across 1 810 participants for DV: $Misleadingness$

	Misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**
Dependent Variable: Misleadingness				
Female [*ref.: Male*]	0.000	−0.012	0.012	0.970
Attended college [*ref.: Didn't attend college*]	−0.003	−0.014	0.009	0.639
Age: $<$ 30 years [*ref.: 30 – 50 years*]	0.002	−0.013	0.017	0.760
Age: $>$ 50 years [*ref.: 30 – 50 years*]	0.002	−0.010	0.017	0.777
Believes in God [*ref.: Doesn't believe in God*]	−0.005	−0.019	0.007	0.398
Minority [*ref.: Not a minority*]	−0.020	−0.032	−0.007	0.007
Vaccinated against COVID-19 [*ref.: Not vaccinated against COVID-19*]	0.023	0.009	0.037	0.004
Low willingness to take risks [*ref.: High willingness to take risks*]	−0.025	−0.037	−0.013	0.001
High trust [*ref.: Low Trust*]	0.002	−0.010	0.014	0.796
High trust in democracy [*ref.: Low trust in democracy*]	0.008	−0.003	0.020	0.149
High thinking preference [*ref.: Low thinking preference*]	0.006	−0.006	0.018	0.323

Table 4.11: AME of replacing an expert flag with a community note on trust in fact-checks for misleading posts depending on participants' demographics and beliefs. The AMEs are calculated from a hierarchical linear regression model with interaction terms predicting predicting the trustworthiness of a fact-check (7-point Likert scale normalized to the interval [0, 1]). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 24\,003$ observations across 1 347 participants.

Dependent Variable: Trustworthiness				
	Misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**
AMEs of Community Note *[ref. Expert Flag]*				
Female	0.036	0.026	0.045	< 0.001
Male	0.053	0.044	0.062	< 0.001
Didn't attend college	0.041	0.033	0.050	< 0.001
Attended college	0.050	0.041	0.060	< 0.001
Age: < 30 years	0.054	0.041	0.067	< 0.001
Age: 30–49 years	0.046	0.037	0.055	< 0.001
Age: > 50 years	0.038	0.027	0.049	< 0.001
Doesn't believe in God	0.077	0.068	0.086	< 0.001
Believes in God	0.017	0.007	0.026	< 0.001
Not a Minority	0.038	0.031	0.046	< 0.001
Minority	0.067	0.053	0.078	< 0.001
Vaccinated against COVID-19	0.045	0.037	0.052	< 0.001
Not vaccinated against COVID-19	0.046	0.033	0.058	< 0.001
Low willingness to take risks	0.030	0.020	0.040	< 0.001
High willingness to take risks	0.056	0.048	0.065	< 0.001
Low trust	0.035	0.023	0.047	< 0.001
High trust	0.049	0.042	0.058	< 0.001
Low trust in democracy	0.031	0.022	0.040	< 0.001
High trust in democracy	0.060	0.051	0.068	< 0.001
Low thinking preference	0.049	0.041	0.057	< 0.001
High thinking preference	0.040	0.030	0.050	< 0.001

Table 4.12: AME of replacing an expert flag with a community note on misleadingness ratings for misleading posts depending on participants' demographics and beliefs. The AMEs are calculated from a hierarchical linear regression model with interaction terms predicting the perceived misleadingness of a post (7-point Likert scale normalized to the interval [0, 1]). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 64\,454$ observations across 1 810 participants.

Dependent Variable: Misleadingness				
	Misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**
AMEs of Community Note *[ref. Expert Flag]*				
Female	0.035	0.026	0.044	< 0.001
Male	0.008	−0.002	0.017	0.118
Didn't attend college	0.027	0.019	0.036	< 0.001
Attended college	0.017	0.007	0.028	< 0.001
Age: < 30 years	0.030	0.018	0.044	< 0.001
Age: 30–49 years	0.017	0.007	0.027	< 0.001
Age: > 50 years	0.027	0.015	0.039	< 0.001
Doesn't believe in God	0.030	0.022	0.038	< 0.001
Believes in God	0.017	0.007	0.026	0.003
Not a Minority	0.023	0.015	0.030	< 0.001
Minority	0.023	0.010	0.037	< 0.001
Vaccinated against COVID-19	0.017	0.010	0.025	< 0.001
Not vaccinated against COVID-19	0.039	0.025	0.051	< 0.001
Low willingness to take risks	0.001	−0.009	0.011	0.831
High willingness to take risks	0.039	0.031	0.047	< 0.001
Low trust	0.034	0.022	0.046	< 0.001
High trust	0.018	0.010	0.026	< 0.001
Low trust in democracy	0.031	0.022	0.041	< 0.001
High trust in democracy	0.014	0.005	0.023	0.003
Low thinking preference	0.025	0.016	0.033	< 0.001
High thinking preference	0.020	0.010	0.030	< 0.001

Table 4.13: AME of different fact-checking interventions (reference: control condition) on the identification of misleading posts depending on participants' demographics and beliefs. The AMEs are calculated from a hierarchical linear regression model with interaction terms predicting misleadingness (7-point Likert scale normalized to the interval [0, 1]). The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 64\,454$ observations across 1 810 participants.

Dependent Variable: Misleadingness				
	Misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**
AMEs of Expert Flag *[ref. No Fact-Check]*				
Female	0.057	0.047	0.067	< 0.001
Male	0.086	0.076	0.097	< 0.001
Didn't attend college	0.069	0.059	0.079	< 0.001
Attended college	0.073	0.062	0.084	< 0.001
Age: < 30 years	0.072	0.056	0.087	< 0.001
Age: 30–49 years	0.072	0.062	0.083	< 0.001
Age: > 50 years	0.068	0.054	0.081	< 0.001
Doesn't believe in God	0.074	0.065	0.084	< 0.001
Believes in God	0.068	0.057	0.078	< 0.001
Not a Minority	0.071	0.063	0.080	< 0.001
Minority	0.069	0.056	0.083	< 0.001
Vaccinated against COVID-19	0.076	0.068	0.085	< 0.001
Not vaccinated against COVID-19	0.055	0.039	0.070	< 0.001
Low willingness to take risks	0.087	0.077	0.099	< 0.001
High willingness to take risks	0.058	0.049	0.069	< 0.001
Low trust	0.055	0.041	0.069	< 0.001
High trust	0.078	0.069	0.086	< 0.001
Low trust in democracy	0.059	0.049	0.069	< 0.001
High trust in democracy	0.083	0.072	0.093	< 0.001
Low thinking preference	0.075	0.066	0.084	< 0.001
High thinking preference	0.064	0.053	0.075	< 0.001
AMEs of Community Flag *[ref. No Fact-Check]*				
Female	0.048	0.039	0.058	< 0.001
Male	0.073	0.062	0.084	< 0.001
Didn't attend college	0.056	0.046	0.065	< 0.001
Attended college	0.064	0.053	0.076	< 0.001
Age: < 30 years	0.058	0.043	0.074	< 0.001
Age: 30–49 years	0.059	0.048	0.069	< 0.001
Age: > 50 years	0.061	0.047	0.075	< 0.001
Doesn't believe in God	0.061	0.051	0.071	< 0.001
Believes in God	0.058	0.048	0.068	< 0.001
Not a Minority	0.055	0.046	0.064	< 0.001
Minority	0.074	0.059	0.088	< 0.001
Vaccinated against COVID-19	0.061	0.053	0.070	< 0.001
Not vaccinated against COVID-19	0.055	0.039	0.072	< 0.001
Low willingness to take risks	0.076	0.064	0.087	< 0.001
High willingness to take risks	0.047	0.038	0.057	< 0.001
Low trust	0.051	0.037	0.066	< 0.001
High trust	0.063	0.054	0.071	< 0.001
Low trust in democracy	0.053	0.043	0.063	< 0.001
High trust in democracy	0.066	0.056	0.077	< 0.001
Low thinking preference	0.065	0.056	0.074	< 0.001
High thinking preference	0.051	0.039	0.063	< 0.001
AMEs of Community Note *[ref. No Fact-Check]*				
Female	0.092	0.082	0.101	< 0.001
Male	0.094	0.084	0.105	< 0.001
Didn't attend college	0.096	0.087	0.105	< 0.001
Attended college	0.090	0.080	0.100	< 0.001
Age: < 30 years	0.102	0.087	0.116	< 0.001
Age: 30–49 years	0.089	0.079	0.099	< 0.001

Continued on next page

Continued from previous page

	AME	Lower CI	Upper CI	*P*-value
Age: $>$ 50 years	0.094	0.082	0.107	$<$ 0.001
Doesn't believe in God	0.104	0.094	0.114	$<$ 0.001
Believes in God	0.084	0.075	0.094	$<$ 0.001
Not a Minority	0.094	0.086	0.102	$<$ 0.001
Minority	0.092	0.078	0.105	$<$ 0.001
Vaccinated against COVID-19	0.093	0.085	0.101	$<$ 0.001
Not vaccinated against COVID-19	0.094	0.080	0.108	$<$ 0.001
Low willingness to take risks	0.088	0.078	0.099	$<$ 0.001
High willingness to take risks	0.097	0.088	0.107	$<$ 0.001
Low trust	0.089	0.076	0.101	$<$ 0.001
High trust	0.096	0.087	0.104	$<$ 0.001
Low trust in democracy	0.090	0.081	0.100	$<$ 0.001
High trust in democracy	0.097	0.087	0.107	$<$ 0.001
Low thinking preference	0.099	0.091	0.108	$<$ 0.001
High thinking preference	0.084	0.073	0.095	$<$ 0.001

4.C.3 Reliance on fact-checks

Participants in the treatment conditions were also asked to indicate whether they tend to rely on fact-checks in general (None (1) to Extreme (5)). We repeated our analysis with reliance on fact-checks as an additional explanatory variable (see Section Demographics, beliefs, and cognitive reflection in the main paper). Here, we code responses higher than "Moderate" (3) as a high reliance in fact-checks (otherwise = low). Table 4.14 reports the AMEs, i. e., how the efficacy of fact-checking interventions varied depending on participants' self-reported reliance on fact-checks across our three dependent variables.

Table 4.14: Average marginal effects (AME) of replacing an expert flag with a community note depending on participants' reliance on fact-checks. The AME are calculated from hierarchical linear regression models with interaction terms predicting trust in fact-checks and misleadingness for misleading posts. Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 24\,003$ observations across 1 347 participants for DV: $Trustworthiness$, and $N = 48\,249$ observations across 1 347 participants for DV: $Misleadingness$

	Misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**
Dependent Variable: Trustworthiness				
AMEs of Community Note *[ref. Expert Flag]*				
High reliance on fact-checks	0.038	0.030	0.045	< 0.001
Low reliance on fact-checks	0.005	−0.008	0.017	0.439
Dependent Variable: Misleadingness				
AMEs of Community Note *[ref. Expert Flag]*				
High reliance on fact-checks	0.012	0.004	0.019	0.005
Low reliance on fact-checks	0.035	0.023	0.047	< 0.001

4.C.4 Cognitive Reflection Test (CRT)

A common method to assess the level of a persons' reflective thinking is the so-called Cognitive Reflection Test (CRT). Participants in our study were asked to answer a 4-item CRT (for further details see Section 4.D). We classified four correct answers as "Passed CRT" and less than four correct answers as "Failed CRT". Table 4.15 reports the average marginal effects (AME) across our three dependent variables. Moderation effects, i. e., how the efficacy of fact-checking interventions varied depending on the outcomes of the CRT are reported in Table 4.16. The findings are described in detail in Section Demographics, beliefs, and cognitive reflection of the main paper.

Table 4.15: Average marginal effects (AME) of the outcome of a CRT ($= 1$ if passed; $= 0$ otherwise) on perceived misleadingness for misleading posts. The AME are calculated from linear mixed-effects regression models with interaction terms. Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 64\,454$ observations across 1 810 participants for DV: $Misleadingness$

	Misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**
Dependent Variable: Misleadingness				
Passed CRT [*ref.: Failed CRT*]	0.034	0.023	0.045	< 0.001

Table 4.16: Average marginal effects (AME) of different fact-checking interventions depending on the outcome of the CRT ($= 1$ if passed; $= 0$ otherwise) for misleading posts. The AME are calculated from linear mixed-effects regression models with interaction terms predicting trust in fact-checks and perceived misleadingness. Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 24\,003$ observations across 1 347 participants for DV: $Trustworthiness$, and $N = 64\,454$ observations across 1 810 participants for DV: $Misleadingness$.

	Misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**
Dependent Variable: Trustworthiness				
AMEs of Community Note *[ref. Expert Flag]*				
Passed CRT	0.074	0.065	0.083	< 0.001
Failed CRT	0.026	0.017	0.035	< 0.001
Dependent Variable: Misleadingness				
AMEs of Community Note *[ref. Expert Flag]*				
Passed CRT	0.034	0.025	0.043	< 0.001
Failed CRT	0.017	0.009	0.026	< 0.001
AMEs of Community Note *[ref. No Fact-Check]*				
Passed CRT	0.094	0.084	0.104	< 0.001
Failed CRT	0.094	0.084	0.103	< 0.001
AMEs of Community Flag *[ref. No Fact-Check]*				
Passed CRT	0.048	0.038	0.059	< 0.001
Failed CRT	0.066	0.056	0.076	< 0.001
AMEs of Expert Flag *[ref. No Fact-Check]*				
Passed CRT	0.060	0.048	0.071	< 0.001
Failed CRT	0.077	0.068	0.086	< 0.001

4.C.5 Analysis with hierarchical logistic regression models

We repeated our analysis with an alternative model specification using hierarchical logistic regression model and treating the Likert-scale responses as binary variables. Specifically, the dependent variable *Trustworthy* took the value $= 1$ if the fact-check was rated at least as somewhat trustworthy (i. e., participants gave a 5 or higher on the 7-point Likert scale) and $= 0$ otherwise. The dependent variable *Misleading* took the value $= 1$ if the fact-check was rated at least as somewhat misleading (i. e., participants gave a 5 or higher on the 7-point Likert scale) and $= 0$ otherwise. Table 4.17 reports the corresponding frequencies across the four experimental conditions. All explanatory variables and random effects specifications were the same as in our main analysis. The logistic mixed-effects models were implemented in R 4.3.2 using the `glmer` package and the `marginaleffects` package.

Consistent with our main analysis, we observe that users exposed to community notes perceived fact-checks as significantly more trustworthy (see Table 4.18) than those exposed to simple misinformation flags (all $P < 0.01$). Furthermore, all fact-checking interventions resulted in participants rating misleading posts as significantly more misleading (see Table 4.19). In sum, we find that all main findings are robust with similar effect sizes as in our main analysis.

Table 4.17: Frequency of ratings on the level of response items (per post and per participant) across the experimental conditions. Ratings given by the participants on a 7-point Likert scale were rescaled into binary variables that took the value $= 1$ if the rating was greater than Neutral (and $= 0$ otherwise).

Variable **Observations (N)**	**Overall** 64,454	**No Fact-Check** 16,205	**Expert Flag** 15,885	**Community Flag** 15,192	**Community Note** 17,172
Trustworthy					
Yes	14,514 (60%)	–	4,617 (59%)	4,260 (56%)	5,637 (66%)
No	9,489 (40%)	–	3,204 (41%)	3,336 (44%)	2,949 (34%)
Misleading					
Yes	43,099 (67%)	10,755 (66%)	10,473 (66%)	9,998 (66%)	11,873 (69%)
No	21,355 (33%)	5,450 (34%)	5,412 (34%)	5,194 (34%)	5,299 (31%)

Table 4.18: Average marginal effects (AME) and 95 % confidence intervals from a hierarchical logistic regression model with three-way interaction terms predicting whether a fact-check was rated as trustworthy ($0 = no$; $1 = yes$). AME is the difference in average predicted probability of whether a fact-check was rated as trustworthy between the group of interest and the reference group expressed as a proportion (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in predicted probabilities). The indented rows report the AMEs of an intervention depending on the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 500 resamples. $N = 24\,003$ observations across 1 347 participants.

	AME	**Lower CI**	**Upper CI**	***P*-value**
Community Note [*ref.: Expert Flag*]	0.065	0.055	0.074	< 0.001
Leaning Biden, *Concordant*	0.104	0.079	0.129	< 0.001
Leaning Biden, *Neutral*	0.120	0.098	0.144	< 0.001
Leaning Biden, *Discordant*	0.047	0.025	0.068	< 0.001
Leaning Trump, *Concordant*	0.013	−0.014	0.041	0.360
Leaning Trump, *Neutral*	0.052	0.026	0.080	< 0.001
Leaning Trump, *Discordant*	0.047	0.023	0.073	< 0.001
Community Flag [*ref.: Expert Flag*]	−0.024	−0.034	−0.014	0.002
Leaning Biden, *Concordant*	−0.024	−0.057	0.007	0.140
Leaning Biden, *Neutral*	−0.013	−0.039	0.012	0.286
Leaning Biden, *Discordant*	−0.019	−0.041	0.002	0.084
Leaning Trump, *Concordant*	−0.047	−0.075	−0.019	0.006
Leaning Trump, *Neutral*	−0.024	−0.054	0.004	0.088
Leaning Trump, *Discordant*	−0.017	−0.045	0.009	0.194
Leaning Trump [*ref.: Leaning Biden*]	−0.100	−0.109	−0.092	0.002
Concordant [*ref.: Neutral*]	−0.043	−0.055	−0.032	0.002
Discordant [*ref.: Neutral*]	0.048	0.038	0.058	< 0.001

Table 4.19: Average marginal effects (AME) and 95 % confidence intervals from a hierarchical logistic regression model with four-way interaction terms predicting whether a post was rated as misleading ($0 = no$; $1 = yes$). AME is the difference in average predicted probability of whether a fact-check was rated as misleading between the group of interest and the reference group expressed as a proportion (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in predicted probabilities). AMEs are reported separately for misleading (columns 2–5) and non-misleading posts (columns 6–9). The indented rows report the AMEs of an intervention depending on the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 500 resamples. $N = 64\,454$ observations across 1 810 participants.

	Misleading				Non-misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**	**AME**	**Lower CI**	**Upper CI**	***P*-value**
Expert Flag [*ref.: No Fact-Check*]	0.067	0.057	0.077	< 0.001	−0.078	−0.091	−0.064	0.002
Community Flag [*ref.: No Fact-Check*]	0.050	0.041	0.060	< 0.001	−0.072	−0.086	−0.060	0.002
Community Note [*ref.: No Fact-Check*]	0.080	0.071	0.090	< 0.001	−0.033	−0.046	−0.019	0.002
Community Note [*ref.: Expert Flag*]	0.013	0.006	0.021	0.006	0.046	0.033	0.060	< 0.001
Leaning Biden, *Concordant*	0.032	0.010	0.055	0.008	0.023	−0.005	0.052	0.088
Leaning Biden, *Neutral*	−0.002	−0.015	0.011	0.728	0.054	0.026	0.085	0.006
Leaning Biden, *Discordant*	−0.014	−0.024	−0.004	0.012	0.068	0.037	0.104	< 0.001
Leaning Trump, *Concordant*	0.068	0.045	0.094	< 0.001	0.080	0.046	0.112	< 0.001
Leaning Trump, *Neutral*	0.028	0.008	0.049	0.012	0.026	−0.006	0.060	0.122
Leaning Trump, *Discordant*	−0.032	−0.047	−0.016	0.002	0.023	−0.012	0.054	0.194
Community Flag [*ref.: Expert Flag*]	−0.017	−0.025	−0.008	0.002	0.006	−0.007	0.019	0.374
Leaning Biden, *Concordant*	−0.061	−0.086	−0.036	0.002	0.010	−0.020	0.039	0.496
Leaning Biden, *Neutral*	−0.009	−0.025	0.008	0.254	0.005	−0.026	0.037	0.742
Leaning Biden, *Discordant*	−0.009	−0.019	0.001	0.106	0.023	−0.011	0.059	0.196
Leaning Trump, *Concordant*	0.009	−0.018	0.039	0.566	0.022	−0.017	0.059	0.246
Leaning Trump, *Neutral*	−0.003	−0.027	0.019	0.798	−0.022	−0.055	0.011	0.204
Leaning Trump, *Discordant*	−0.023	−0.040	−0.007	0.010	−0.004	−0.037	0.033	0.838
Leaning Trump [*ref.: Leaning Biden*]	−0.029	−0.036	−0.022	0.002	0.049	0.039	0.058	< 0.001
Concordant [*ref.: Neutral*]	−0.070	−0.079	−0.061	0.002	0.042	0.029	0.053	< 0.001
Discordant [*ref.: Neutral*]	0.053	0.046	0.060	< 0.001	0.232	0.221	0.244	< 0.001

4.C.6 Analysis with cluster-robust standard errors

In our main analysis, we used hierarchical linear regression models with crossed random intercepts for posts and subjects. As an alternative model specification, we repeated our analysis using a linear regression model with robust standard errors clustered on both subjects and posts, analogous to earlier research Pennycook et al., 2020. The average marginal effects (AMEs) are visualized in Figures 4.10 and 4.11 and tabulated in Tables 4.20 and 4.21. Across all models, the results were qualitatively identical and consistently supported our findings.

Table 4.20: Average marginal effects (AME) and 95 % confidence intervals from a linear regression model with robust standard errors clustered on subjects and posts predicting the trustworthiness of a fact-check (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in average predicted probability of whether a fact-check was rated as trustworthy between the group of interest and the reference group expressed as a proportion (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in predicted probabilities). The indented rows report the AMEs of an intervention depending on the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 24\,003$ observations across 1 347 participants.

	Misleading			
	AME	**Lower CI**	**Upper CI**	***P*-value**
Community Note [*ref.: Expert Flag*]	0.048	0.038	0.059	< 0.001
Leaning Biden, *Concordant*	0.063	0.041	0.086	< 0.001
Leaning Biden, *Neutral*	0.074	0.051	0.098	< 0.001
Leaning Biden, *Discordant*	0.044	0.020	0.068	< 0.001
Leaning Trump, *Concordant*	0.023	−0.002	0.047	0.079
Leaning Trump, *Neutral*	0.042	0.016	0.068	0.004
Leaning Trump, *Discordant*	0.041	0.015	0.071	0.005
Community Flag [*ref.: Expert Flag*]	−0.013	−0.024	−0.002	0.017
Leaning Biden, *Concordant*	−0.010	−0.035	0.015	0.426
Leaning Biden, *Neutral*	−0.010	−0.032	0.015	0.404
Leaning Biden, *Discordant*	−0.010	−0.035	0.016	0.454
Leaning Trump, *Concordant*	−0.027	−0.052	−0.003	0.032
Leaning Trump, *Neutral*	−0.008	−0.034	0.022	0.504
Leaning Trump, *Discordant*	−0.013	−0.042	0.014	0.365
Leaning Trump [*ref.: Leaning Biden*]	−0.084	−0.092	−0.076	0.001
Concordant [*ref.: Neutral*]	−0.023	−0.032	−0.013	0.001
Discordant [*ref.: Neutral*]	0.022	0.011	0.032	< 0.001

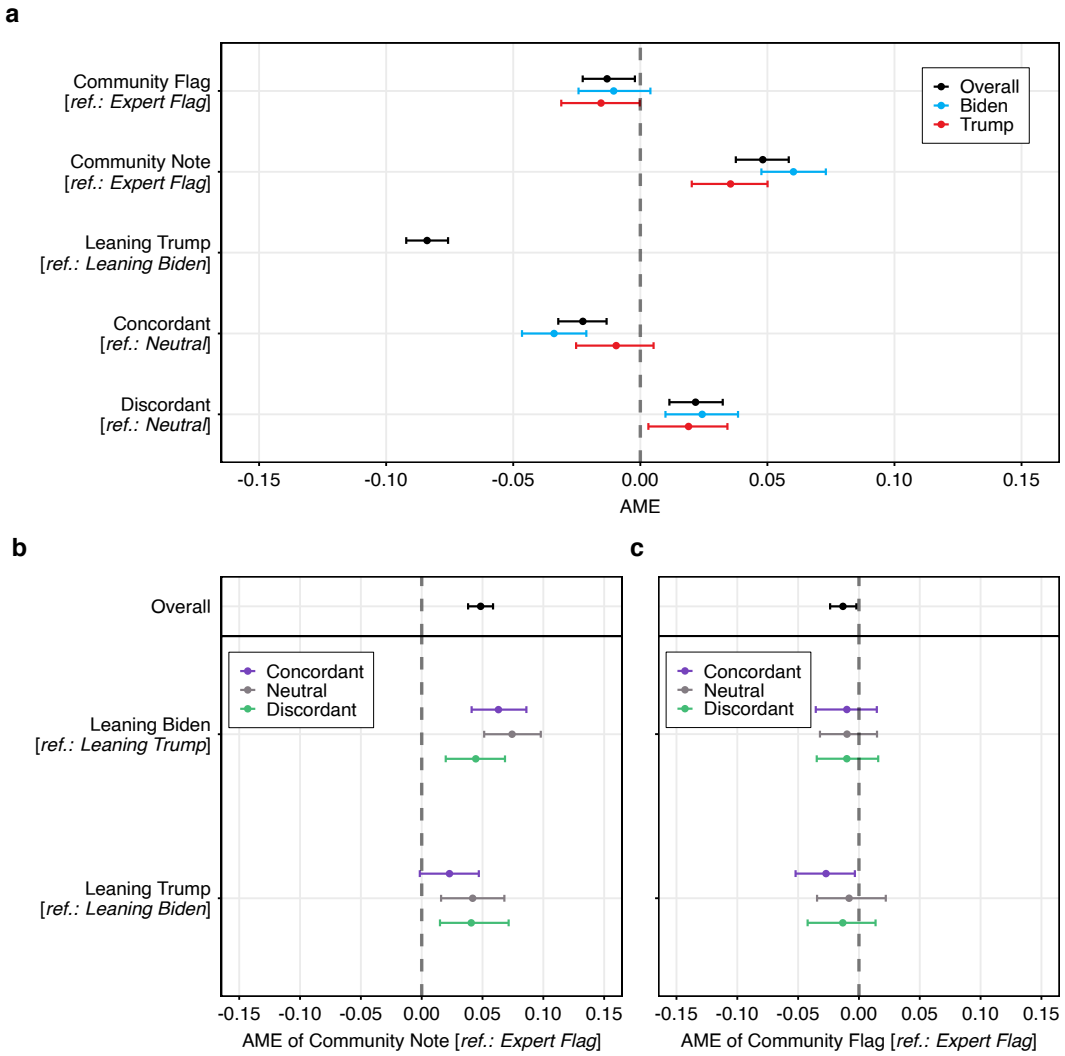


Figure 4.10: **(a)** Shown are the average marginal effects (AME) and 95 % confidence intervals from a linear regression model with robust standard errors clustered on subjects and posts predicting the trustworthiness of a fact-check (7-point Likert scale normalized to the interval [0, 1]). **(b)** AME of replacing an expert flag with a community note (i. e., the average difference between the predicted marginal effects for community notes vs. expert flags) across the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). **(c)** AME of replacing an expert flag with a community flag (i. e., the average difference between the predicted marginal effects for community flags vs. expert flags) across the political leanings of participants and the fact-checked posts. Control variables and random intercepts for posts and subjects were included. The 95 % confidence intervals (error bars) were derived using the bootstrap method for 1 000 resamples. $N = 24\,003$ observations across 1 347 participants. Full AMEs are in SI, Table 4.20.

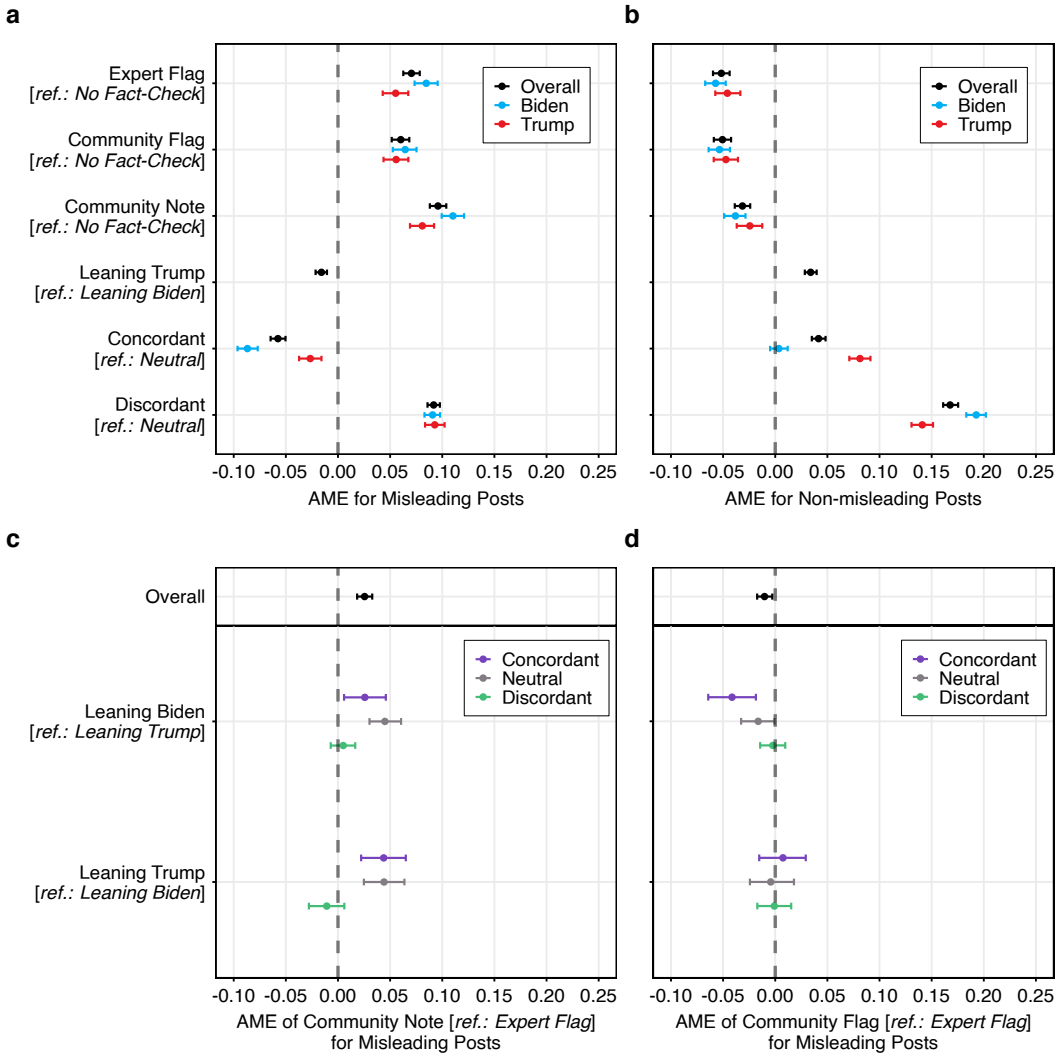


Figure 4.11: Shown are the average marginal effects (AME) and 95 % confidence intervals from a linear regression model with robust standard errors clustered on subjects and posts predicting the perceived misleadingness of a post (7-point Likert scale normalized to the interval [0, 1]). **(a)** AMEs for misleading posts. **(b)** AMEs for non-misleading posts. **(c)** AMEs when replacing an expert flag with a community note (i. e., the average difference between the predicted marginal effects for community notes vs. expert flags) across the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). **(d)** AMEs when replacing an expert flag with a community flag (i. e., the average difference between the predicted marginal effects for community flags vs. expert flags) across the political leanings of participants and the fact-checked posts. Control variables and random intercepts for posts and subjects were included. The 95 % confidence intervals (error bars) were derived using the bootstrap method for 1 000 resamples. $N = 64\,454$ observations across 1 810 participants. Full AMEs are in SI, Table 4.21.

Table 4.21: Average marginal effects (AME) and 95 % confidence intervals from a linear regression model with robust standard errors clustered on subjects and posts predicting the perceived misleadingness of a post (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in average predicted probability of whether a fact-Check was rated as misleading between the group of interest and the reference group expressed as a proportion (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in predicted probabilities). AMEs are reported separately for misleading (columns 2–5) and non-misleading posts (columns 6–9). The indented rows report the AMEs of an intervention depending on the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-Checked posts (concordant, neutral, discordant). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 64\,454$ observations across 1 810 participants.

	Misleading				**Non-misleading**			
	AME	**Lower CI**	**Upper CI**	***P*-value**	**AME**	**Lower CI**	**Upper CI**	***P*-value**
Expert Flag [*ref.: No Fact-Check*]	0.070	0.063	0.078	< 0.001	−0.052	−0.060	−0.044	0.001
Community Flag [*ref.: No Fact-Check*]	0.060	0.052	0.068	< 0.001	−0.051	−0.059	−0.042	0.001
Community Note [*ref.: No Fact-Check*]	0.096	0.088	0.104	< 0.001	−0.031	−0.039	−0.024	0.001
Community Note [*ref.: Expert Flag*]	0.026	0.018	0.033	< 0.001	0.020	0.013	0.028	< 0.001
Leaning Biden, *Concordant*	0.026	0.006	0.046	0.016	0.014	−0.004	0.031	0.123
Leaning Biden, *Neutral*	0.045	0.030	0.061	< 0.001	0.017	0.001	0.033	0.037
Leaning Biden, *Discordant*	0.005	−0.007	0.017	0.408	0.027	0.009	0.045	0.007
Leaning Trump, *Concordant*	0.044	0.022	0.065	< 0.001	0.035	0.014	0.055	< 0.001
Leaning Trump, *Neutral*	0.044	0.025	0.064	< 0.001	0.019	0.000	0.037	0.042
Leaning Trump, *Discordant*	−0.011	−0.028	0.006	0.188	0.009	−0.011	0.031	0.409
Community Flag [*ref.: Expert Flag*]	−0.010	−0.017	−0.003	0.010	0.001	−0.007	0.009	0.721
Leaning Biden, *Concordant*	−0.041	−0.064	−0.019	0.001	0.008	−0.011	0.026	0.386
Leaning Biden, *Neutral*	−0.016	−0.033	−0.001	0.036	−0.003	−0.018	0.014	0.743
Leaning Biden, *Discordant*	−0.002	−0.014	0.010	0.742	0.006	−0.014	0.026	0.534
Leaning Trump, *Concordant*	0.007	−0.015	0.029	0.501	−0.001	−0.022	0.020	0.945
Leaning Trump, *Neutral*	−0.004	−0.024	0.018	0.685	−0.009	−0.029	0.010	0.312
Leaning Trump, *Discordant*	−0.001	−0.017	0.015	0.911	0.007	−0.015	0.029	0.539
Leaning Trump [*ref.: Leaning Biden*]	−0.016	−0.022	−0.011	0.001	0.034	0.028	0.040	< 0.001
Concordant [*ref.: Neutral*]	−0.058	−0.065	−0.050	0.001	0.042	0.035	0.048	< 0.001
Discordant [*ref.: Neutral*]	0.092	0.086	0.098	< 0.001	0.168	0.161	0.175	< 0.001

4.C.7 Additional experiment for effects of presentation format

We conducted an additional experiment (see Section Effect of presentation format and context in the main paper) in which participants were randomly assigned to one of two conditions: (i) *community note (main)*, where misleading posts were supplemented with a textual community note without an explicit warning label (i. e., previous condition 4), or (ii) *community note (alternative)*, where misleading posts were supplemented by a combination of a textual community note and an explicit warning label indicating that community fact-checkers categorized the content as being misleading

We fitted hierarchical linear regression models to quantify the effects of the alternative presentation format of community notes on trust in fact-checks, and the identification of misleading and non-misleading posts. Note that due to the time that has passed since the first experiment was carried out and the resulting potential differences in the participants' level of knowledge on some of the topics covered in the social media posts, the answers cannot be directly compared across experiments. Therefore, we run a separate regression model for the new experiment. The key explanatory variable in the regression model was a binary dummy indicating the alternative presentation format. All other explanatory variables and random effects specifications were the same as in our main analysis. Tables 4.22 to 4.24 present the AMEs for our three dependent variables. The findings are described in Section Effect of presentation format and context of the main paper.

Table 4.22: Average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with three-way interaction terms predicting the trustworthiness of a fact-check (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in the average predicted trustworthiness ratings between the group of interest and the reference group (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in trustworthiness ratings). The indented rows report the AMEs of an intervention depending on the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 12\,150$ observations across 675 participants.

	AME	Lower CI	Upper CI	*P*-value
Community Note (Alternative) [*ref.: (Main)*]	−0.001	−0.008	0.007	0.846
Leaning Biden	−0.026	−0.036	−0.016	0.001
Leaning Biden, *Concordant*	−0.024	−0.041	−0.004	0.017
Leaning Biden, *Neutral*	−0.030	−0.049	−0.013	0.001
Leaning Biden, *Discordant*	−0.024	−0.041	−0.008	0.005
Leaning Trump	0.025	0.015	0.036	< 0.001
Leaning Trump, *Concordant*	0.031	0.013	0.049	< 0.001
Leaning Trump, *Neutral*	0.017	−0.003	0.034	0.093
Leaning Trump, *Discordant*	0.029	0.009	0.048	< 0.001
Leaning Trump [*ref.: Leaning Biden*]	−0.063	−0.070	−0.055	0.001
Concordant [*ref.: Neutral*]	−0.026	−0.035	−0.017	0.001
Leaning Biden	−0.039	−0.053	−0.027	0.001
Leaning Trump	−0.012	−0.026	0.001	0.067
Discordant [*ref.: Neutral*]	0.011	0.002	0.020	0.010
Leaning Biden	0.020	0.009	0.032	< 0.001
Leaning Trump	0.001	−0.012	0.014	0.858

Table 4.23: Average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with four-way interaction terms predicting the perceived misleadingness of a post (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in the average predicted misleadingness ratings between the group of interest and the reference group (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in misleadingness ratings). AMEs are reported separately for misleading (columns 2–5) and non-misleading posts (columns 6–9). The indented rows report the AMEs of an intervention depending on the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 24\,300$ observations across 675 participants.

	Misleading				**Non-misleading**			
	AME	**Lower CI**	**Upper CI**	***P*-value**	**AME**	**Lower CI**	**Upper CI**	***P*-value**
Community Note (Alternative) [*ref.: (Main)*]	0.020	0.012	0.027	< 0.001	−0.014	−0.023	−0.005	0.001
Leaning Biden	0.014	0.004	0.024	0.011	−0.012	−0.023	0.000	0.050
Leaning Biden, *Concordant*	0.010	−0.008	0.028	0.292	−0.002	−0.023	0.018	0.828
Leaning Biden, *Neutral*	0.008	−0.010	0.025	0.368	−0.008	−0.027	0.010	0.399
Leaning Biden, *Discordant*	0.023	0.006	0.039	0.005	−0.023	−0.045	−0.002	0.032
Leaning Trump	0.026	0.015	0.037	< 0.001	−0.017	−0.028	−0.004	0.011
Leaning Trump, *Concordant*	0.046	0.026	0.068	< 0.001	−0.018	−0.039	0.003	0.113
Leaning Trump, *Neutral*	0.014	−0.005	0.033	0.122	0.002	−0.020	0.023	0.865
Leaning Trump, *Discordant*	0.017	0.000	0.034	0.051	−0.030	−0.052	−0.008	0.011
Leaning Trump [*ref.: Leaning Biden*]	−0.010	−0.017	−0.002	0.016	0.019	0.010	0.027	< 0.001
Concordant [*ref.: Neutral*]	−0.048	−0.057	−0.038	0.001	0.051	0.041	0.061	< 0.001
Leaning Biden	−0.080	−0.092	−0.068	0.001	0.031	0.016	0.045	< 0.001
Leaning Trump	−0.015	−0.029	−0.001	0.030	0.072	0.056	0.087	< 0.001
Discordant [*ref.: Neutral*]	0.032	0.023	0.040	< 0.001	0.153	0.143	0.163	< 0.001
Leaning Biden	0.035	0.024	0.047	< 0.001	0.148	0.133	0.162	< 0.001
Leaning Trump	0.029	0.016	0.042	< 0.001	0.158	0.144	0.173	< 0.001

Table 4.24: Average marginal effects (AME) and 95 % confidence intervals from a hierarchical linear regression model with four-way interaction terms predicting sharing intentions (7-point Likert scale normalized to the interval [0, 1]). AME is the difference in average predicted sharing intentions between the group of interest and the reference group (e. g., an AME of $+0.05$ indicates a 5 percentage point difference in sharing intentions). AMEs are reported separately for misleading (columns 2–5) and non-misleading posts (columns 6–9). The indented rows report the AMEs of an intervention depending on the political leanings of participants (leaning Trump vs. Biden) and the political congruence of the fact-checked posts (concordant, neutral, discordant). Random intercepts for posts and subjects are included. The 95 % confidence intervals for the AMEs were derived using the bootstrap method for 1 000 resamples. $N = 24\,300$ observations across 675 participants.

	Misleading				**Non-misleading**			
	AME	**Lower CI**	**Upper CI**	***P*-value**	**AME**	**Lower CI**	**Upper CI**	***P*-value**
Community Note (Alternative) [*ref.: (Main)*]	−0.009	−0.016	−0.002	0.005	0.012	0.004	0.020	0.006
Leaning Biden	−0.021	−0.031	−0.012	0.001	0.015	0.005	0.026	0.006
Leaning Biden, *Concordant*	−0.015	−0.031	0.001	0.062	0.003	−0.017	0.024	0.764
Leaning Biden, *Neutral*	−0.024	−0.040	−0.008	0.006	0.022	0.003	0.042	0.024
Leaning Biden, *Discordant*	−0.023	−0.038	−0.008	0.003	0.019	0.001	0.037	0.024
Leaning Trump	0.003	−0.006	0.013	0.521	0.009	−0.002	0.020	0.150
Leaning Trump, *Concordant*	−0.001	−0.017	0.015	0.872	−0.001	−0.021	0.020	0.955
Leaning Trump, *Neutral*	0.014	−0.004	0.031	0.112	−0.002	−0.024	0.018	0.873
Leaning Trump, *Discordant*	−0.003	−0.020	0.013	0.706	0.029	0.010	0.048	< 0.001
Leaning Trump [*ref.: Leaning Biden*]	0.014	0.007	0.021	< 0.001	−0.002	−0.010	0.006	0.579
Concordant [*ref.: Neutral*]	0.010	0.002	0.019	0.022	−0.003	−0.014	0.008	0.583
Leaning Biden	0.009	−0.002	0.021	0.114	0.009	−0.005	0.024	0.241
Leaning Trump	0.010	−0.002	0.022	0.096	−0.015	−0.030	0.000	0.049
Discordant [*ref.: Neutral*]	−0.006	−0.014	0.002	0.150	−0.072	−0.081	−0.062	0.001
Leaning Biden	−0.009	−0.021	0.003	0.150	−0.076	−0.090	−0.063	0.001
Leaning Trump	−0.003	−0.015	0.008	0.599	−0.067	−0.081	−0.054	0.001

Appendix 4.D Further methodological details

4.D.1 Participants

Our preregistration (which can be found here: https://aspredicted.org/rb45k.pdf) describes that we "plan to recruit 1500 participants." In the end, we slightly exceeded this number: after handling all outliers 1810 valid participants remained (1,347 in treatment conditions). Participants were recruited between April 30th and June 15th in seven experimental sessions to achieve a balanced number of participants with different political leanings (Republicans are underrepresented on Prolific Douglas et al., 2023). Specifically, we explicitly recruited participants who indicated they voted for Biden (sessions 2, 5, and 6) or Trump (sessions 3, 4, and 7) in the 2020 presidential election using the prescreening tool available on Prolific. Participants who participated in one session were prevented from participating in subsequent session. The procedure was identical in all sessions. In our regression analysis, we control for potential pre-treatment effects by including random intercepts for subjects and posts. Participants were paid $4 for completing the survey (approx. 20min), which translated to an hourly wage of $12/h.

The participants were allocated to the individual sessions as follows:

- Session 1: April 30th, 2023. $n = 296$ finished the survey, participants who indicated responding randomly ($n = 8$), searching online for any of the headlines during the experiment ($n = 6$), do not have a social media account ($n = 9$), or failed the attention checks ($n = 29$) were removed from analysis. The final sample was $n = 260$ ($M_{Age} =$ 37, 52 % female).
- Session 2: May 15th, 2023. $n = 397$ finished the survey, participants who indicated responding randomly ($n = 5$), searching online for any of the headlines during the experiment ($n = 6$), do not have a social media account ($n = 5$), or failed the attention checks ($n = 19$) were removed from analysis. The final sample was $n = 374$ ($M_{Age} =$ 41, 51 % female).
- Session 3: May 16th, 2023. $n = 403$ finished the survey, participants who indicated responding randomly ($n = 5$), searching online for any of the headlines during the experiment ($n = 4$), do not have a social media account ($n = 11$), or failed the attention checks ($n = 30$) were removed from analysis. The final sample was $n = 362$ ($M_{Age} =$ 46, 52 % female).
- Session 4: May 25th, 2023. $n = 260$ finished the survey, participants who indicated responding randomly ($n = 4$), searching online for any of the headlines during the experiment ($n = 1$), do not have a social media account ($n = 7$), or failed the attention checks ($n = 25$) were removed from analysis. The final sample was $n = 227$ ($M_{Age} =$ 43, 53 % female).
- Session 5: May 26th, 2023. $n = 156$ finished the survey, participants who indicated responding randomly ($n = 2$), searching online for any of the headlines during the experiment ($n = 2$), do not have a social media account ($n = 6$), or failed the attention checks ($n = 13$) were removed from analysis. The final sample was $n = 137$ ($M_{Age} =$ 40, 48 % female).

Figure 4.12: CRT Questions

- Session 6: June 14th, 2023. $n = 252$ finished the survey, participants who indicated responding randomly ($n = 1$), searching online for any of the headlines during the experiment ($n = 3$), do not have a social media account ($n = 5$), or failed the attention checks ($n = 18$) were removed from analysis. The final sample was $n = 230$ ($M_{Age} =$ 40, 49 % female).
- Session 7: June 15th, 2023. $n = 250$ finished the survey, participants who indicated responding randomly ($n = 2$), searching online for any of the headlines during the experiment ($n = 4$), do not have a social media account ($n = 10$), or failed the attention checks ($n = 21$) were removed from analysis. The final sample was $n = 220$ ($M_{Age} =$ 45, 51 % female).

4.D.2 Participants in additional experiment

For the additional experiment studying the effects of different presentation formats of community notes, we recruited $n = 795$ participants between February 6th and February 13th, 2024 in one experimental session via Prolific.com. Survey design, prescreening, payment, and data cleansing were conducted identically to the main experiment. In addition, we excluded people who had already taken part in the survey last year.

Participants who indicated responding randomly ($n = 19$), searching online for any of the headlines during the experiment ($n = 26$), not having a social media account ($n = 15$), or failed the attention checks ($n = 64$) were removed from analysis. This led to a final sample of $n = 675$ participants ($M_{Age} = 42.88$, 50.81 % female).

4.D.3 Additional question items

Following the questions regarding trustworthiness, misleadingness, and sharing intentions, participants were asked to complete a 4-item Cognitive Reflection Test (CRT). A CRT is a common method to determine a person's level of reflective thinking. The purpose of its design is to assess an individual's inclination to replace an intuitive yet incorrect response with a more rational and accurate response Frederick, 2005. For our CRT, we used a combination of both numeric and non-numeric questions from different sources Frederick, 2005; Thomson and Oppenheimer, 2016 (see Figure 4.12).

Subsequently, participants were asked about their social media use:

- What type of social media accounts do you use (if any)? (Facebook, Twitter, Snapchat, Instagram, WhatsApp, TikTok, Other, None)
- Which of these types of content would you consider sharing on social media (if any)? (Political news, Sports news, Celebrity news, Science/technology news, Business news, Other, None)
- When deciding whether to share a piece of content on social media, how important is it to you that the content is... (Accurate, Surprising, Interesting, Aligned with my beliefs, Funny; each answered on a 5-point likert scale (Not al all, Slightly, Moderately, Very, Extremely))
- To what extend do you trust the information that comes from the following? (National news organizations, local news organization, friends and family, Social network sites (e.g., Facebook, Twitter), 3rd party fact-checkers (e.g., snopes.com, factcheck.org); each answered on a 5-point likert scale (not at all, a little, a moderate amount, a lot, a great deal))

Depending on the condition participants where assigned to (condition 2–4), we also asked questions about their awareness of the specific type of fact-checks and what influence the respective fact-checks had on their assessment regarding the accuracy of a post:

- Prior to you taking this study, were you aware of the existence of third-party fact-checking organizations (community-based fact-checking)? (Yes/No)
- To what extent did the "Checked by third-party fact-checking organizations" tag ("Checked by other social media users with multiple perspectives tag", "Community note") influence you opinion about the accuracy of the social media posts? (5-point likert scale; Not at all, Slightly, Moderately, Very, Extremely)
- We are interested in whether the "Checked by third-party fact-checking organizations" tag ("Checked by other social media users with multiple perspectives tag", "Community note") influenced your opinion about the accuracy of the social media posts that were tagged as potentially misleading. I rated "potentially misleading" posts as: (7-point likert scale; Much less accurate, Less accurate, Slightly less accurate, Tag had no influence, Slightly more accurate, More accurate, Much more accurate)
- We are interested in whether the "Checked by third-party fact-checking organizations" tag ("Checked by other social media users with multiple perspectives tag", "Community note") influenced your opinion about the accuracy of the social media posts that were NOT tagged as potentially misleading. I rated posts that were NOT "potentially misleading" as: (7-point likert scale; Much less accurate, Less accurate, Slightly less accurate, Tag had no influence, Slightly more accurate, More accurate, Much more accurate)

At the end of the survey, participants where asked to answer several demographic questions: age, gender, level of education, proficiency in English, US region where they live, stance toward god (or gods), whether they have been vaccinated against COVID-19, whether they see themselves as part of an ethnic minority, political orientation (Democrat, Republican, Third Party, Other), and questions on their voting behavior in the 2020 presidency election. First, they were asked who they voted for (Joe Biden, Donald Trump, Other Candidate, I did not vote for reasons outside my control, I did not vote but I could have, I did not vote out of protest)

and second, who they would prefer to be president, if they absolutely had to choose between Joe Biden and Donald Trump.

In addition, participants had to provide an assessment of their attitudes toward the following statements on risk aversion and trust (5-point likert scale; Not at all, Not really, Undecided, Somewhat, Very much):

- I am generally a person that is fully prepared to take risks.
- I usually have the feeling that I can trust the people I interact with in my daily life.
- I have a fundamental trust in democracy.
- I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.

At the end of the survey, participants were asked if they responded randomly at any point during the survey or searched for the content online. Participants who answered “yes” to any of these questions were excluded from the analysis. Furthermore, we excluded participants who indicated not to have any social media account or failed any attention check.

Appendix 4.E List of social media posts and fact-checks

All participants were presented with 36 social media posts (18 "Non-misleading" and 18 "misleading"). The posts for our study were selected as follows. First, misleading posts and corresponding fact-checks have been manually selected from X's Community Notes platform. Only posts that had been identified as helpful by the platform's bridging algorithm Wojcik et al., 2022 – meaning they were rated as helpful by users across diverse viewpoints and made visible on X (https://communitynotes.x.com/guide/en/contributing/notes-on-twitter) – were considered. Two research assistants (RAs) initially selected around 50 posts to cover a diverse set of posts across a wide variety of relevant topics (Politics, Business, Health, Climate Change, Celebrities, Other). From this initial selection, 18 posts were chosen for our study to appeal to subjects with different political views (6 "pro-Democrat," 6 "Neutral," and 6 "pro-Republican" posts). We aimed for a balanced distribution of the political leanings within the individual topic groups, where appropriate. For example, the topics "Politics" and "Health" comprised posts across all political leanings, whereas posts on "Celebrities" were all politically neutral. Second, Non-misleading posts have been manually selected via corresponding keyword searches from X to represent a similar distribution of topics. Here, we again started with a larger list of posts, which was then narrowed down to 18 posts that were evenly distributed across political lines (6 "pro-Democrat," 6 "Neutral," and 6 "pro-Republican" posts).

To ensure the correctness of the fact-checking labels, we have taken the following steps. First, all fact-checks for misleading posts were carried out by community fact-checkers on X's Community Notes platform and rated as helpful. Second, we ensured that we only include posts that we perceived as clearly misleading or non-misleading. Third, we had three trained RAs manually assess the veracity of the posts. The RAs were not aware of the fact-checking label and were tasked to manually assess professional fact-checks (snopes.com, factcheck.org, etc.) or other reliable sources to determine whether the posts were misleading (or Non-misleading). The assessments of the RAs were in perfect agreement with the fact-checking labels.

As a further check, we had three trained RA to validate our labels for the political orientation of the posts. The RAs were asked to assume that the content of all posts was entirely accurate and to evaluate whether the posts would have been more favorable to Democrats or Republicans (on a 5-point Likert-scale ranging from "more favorable to Democrats" (1) to "more favorable to Republicans" (5)). The distribution of participant's responses is shown in Figure 4.13. Overall, pro-Democrat posts ($M_{Misleading/Dem} = 1.40$, $M_{Non-misleading/Dem} = 1.53$) were rated as significantly less favorable for Republicans than pro-Republican posts ($M_{Misleading/Rep} = 4.47$, $M_{Non-misleading/Rep} = 4.10$). Pro-Democrat posts were rated as significantly more favorable to Democrats than politically neutral posts and pro-Republican posts were more favorable to Republicans than politically neutral posts. Statistically, each of these differences in means was significant according to two-sided t-tests (each $P < 0.001$).

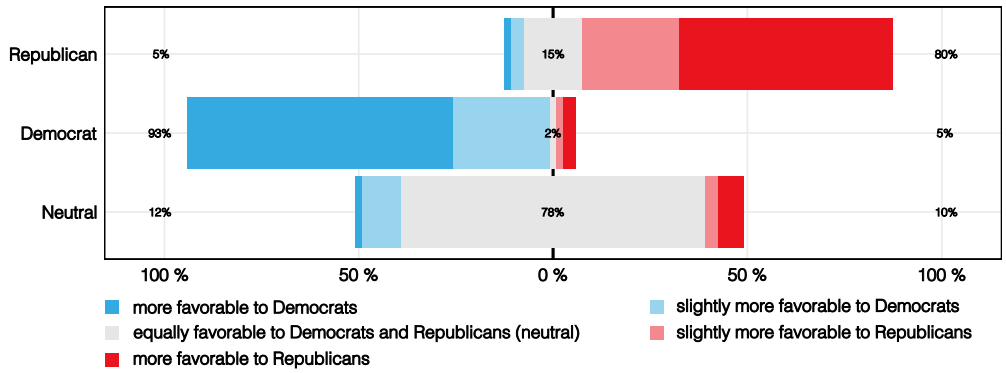


Figure 4.13: Distribution of participants' responses regarding the political orientation of posts on 5-point Likert scales.

Chapter 5

Characterizing AI-Generated Misinformation on Social Media

Abstract

AI-generated misinformation (e. g., deepfakes) poses a growing threat to information integrity on social media. However, prior research has largely focused on its potential societal consequences rather than its real-world prevalence. In this study, we conduct a large-scale empirical analysis of AI-generated misinformation on the social media platform X. Specifically, we analyze a dataset comprising 91 452 misleading posts, both AI-generated and non-AI-generated, that have been identified and flagged through X's Community Notes platform. Our analysis yields four main findings: (i) AI-generated misinformation is more often centered on entertaining content and tends to exhibit a more positive sentiment than conventional forms of misinformation, (ii) it is more likely to originate from smaller user accounts, (iii) despite this, it is significantly more likely to go viral, and (iv) it is slightly less believable and harmful compared to conventional misinformation. Altogether, our findings highlight the unique characteristics of AI-generated misinformation on social media. We discuss important implications for platforms and future research.

Keywords: Social Media, AI-generated misinformation, Community Notes, Virality

5.1 Introduction

Artificial intelligence (AI) technologies are rapidly transforming the social media landscape, enabling the creation of convincing synthetic content. A prominent example is deepfakes – AI-generated media that can realistically imitate real people's appearance, voice, or actions (Feuerriegel et al., 2023; Groh et al., 2024; Hancock & Bailenson, 2021; Sippy et al., 2024; Vaccari & Chadwick, 2020). These and other forms of AI-generated misinformation blur the line between authentic and fabricated content, making it increasingly difficult for users and platforms to discern what is real (Goldstein et al., 2023; Groh et al., 2024). As these tools become more sophisticated and accessible, they offer powerful new means for spreading misinformation (Feuerriegel et al., 2023; Goldstein et al., 2023). Their increased scalability, multilingualism, and multimodality further complicate detection, and pose significant challenges to the defense strategies previously employed by digital platforms and users (Feuerriegel et al., 2023).

Despite growing concerns, AI-generated misinformation on social media remains poorly understood. Prior research warns that such content can have serious societal consequences, including the erosion of trust in media, institutions, and democratic processes (Dobber et al., 2021; Goldstein et al., 2023; Hancock & Bailenson, 2021; Vaccari & Chadwick, 2020; Yan et al., 2025). Individuals with low media literacy may be particularly susceptible to these threats (Feuerriegel et al., 2023; Grinberg et al., 2019). While several studies have analyzed the ability of humans (Bashardoust et al., 2024; Bray et al., 2023; Cooke et al., 2024; Diel et al., 2024; Groh et al., 2024; Köbis et al., 2021; Kreps & Kriner, 2022; Somoray & Miller, 2023) and machine learning systems (Montserrat et al., 2020; Zi et al., 2020) to detect AI-generated content – typically finding that such detection is highly challenging – there is little empirical evidence on how AI-generated misinformation actually spreads on social media and how it differs from conventional forms of misinformation. This hinders efforts to design platform defenses, guide policy, and build public resilience (Feuerriegel et al., 2023). Our study addresses this gap with by characterizing AI-generated misinformation circulating on the social media platform X (formerly Twitter).

Research goal: In this paper, we conduct a large-scale empirical analysis to characterize AI-generated misinformation circulating on the social media platform X. Specifically, we address the following research questions:

- **RQ1:** *How does AI-generated misinformation differ from other forms of misinformation in terms of content attributes (e. g., modality, topics)?*
- **RQ2:** *What are the characteristics of accounts that disseminate AI-generated misinformation?*
- **RQ3:** *Is AI-generated misinformation more viral than other types of misinformation?*
- **RQ4:** *How does AI-generated misinformation differ in terms of its believability and harmfulness?*

Data & methods: We analyze a large dataset consisting of 91 452 misleading posts, both AI-generated and non-AI-generated, that have been identified and flagged on X's Community Notes platform (**Twitter.2021**; Pröllochs, 2022) between January 2023 and January 2025 (i. e., during an observation period of two years). Compared to alternative approaches (e. g.,

manual annotation; machine learning-based identification), Community Notes offers two key advantages for identifying AI-generated misinformation: (i) it enables large-scale detection (Pilarski et al., 2024; Pröllochs, 2022) and (ii) it achieves high identification accuracy enabled by the wisdom of crowds (Allen et al., 2021; Drolsbach & Pröllochs, 2023b; Martel et al., 2024). To characterize AI-generated misinformation, we employ a large language model (LLM) to annotate misleading posts across a wide range of dimensions (e. g., sentiment, topics, believability, harmfulness). Based on this data, we then empirically analyze how AI-generated misinformation on social media differs from traditional forms of misinformation. The data collection and the analysis follow common standards for ethical research (Rivers & Lewis, 2014).

Contributions: To the best of our knowledge, our study is the first large-scale study characterizing AI-generated misinformation circulating on social media. Our analysis contributes the following four main findings: (i) AI-generated misinformation is more often centered on entertaining content and tends to exhibit a more positive sentiment than conventional forms of misinformation, (ii) it is more likely to originate from smaller user accounts, (iii) despite this, it is significantly more likely to go viral, and (iv) it is slightly less believable and harmful as conventional misinformation. These findings highlight the distinct role of AI-generated within the broader misinformation ecosystem and offer important implications for platforms, policymakers, and researchers seeking to design countermeasures against AI-generated misinformation.

5.2 Background

The manipulation of media content has long been a concern; however, the rapid advancements of generative AI tools have significantly lowered the barriers to creating sophisticated AI-generated content (Feuerriegel et al., 2023; Westerlund, 2019). AI-generated content (e. g., deepfakes) – artificially generated videos, images, and audio designed to mimic real individuals – have become increasingly prevalent on social media, raising critical concerns about their role in spreading misinformation (Groh et al., 2024; Hancock & Bailenson, 2021; Vaccari & Chadwick, 2020). This emerging form of digital misinformation poses significant challenges to the defense strategies previously employed by digital platforms and users (Feuerriegel et al., 2023; Goldstein et al., 2023). Additionally, the increased scalability, multilingualism, and multimodality of AI-generated content further complicate its detection and mitigation (Feuerriegel et al., 2023; Timmerman et al., 2023).

Research on AI-generated misinformation is still in its early stages, with most existing studies focusing on their potential societal impacts, such as erosion of trust in media and democratic institutions (Dobber et al., 2021; Goldstein et al., 2023; Hancock & Bailenson, 2021; Vaccari & Chadwick, 2020). Despite growing concern, there is limited empirical evidence on the actual prevalence of AI-generated misinformation on social media and how it spreads relative to other forms of misinformation. Although AI-generated misinformation is often portrayed as a particularly dangerous type of digital misinformation (Hancock & Bailenson, 2021), it remains unclear what specifically distinguishes them from traditional misinformation in terms of its content characteristics, reach, and user engagement.

A separate body of research has focused on detecting AI-generated misinformation, typically through one of two approaches: (1) human-centered methods (Groh et al., 2024) and (2) automated, machine learning-based methods (Montserrat et al., 2020; Zhou et al., 2023).

Human-centered methods, which are dependent on individual judgements, often struggle to accurately identify AI-generated content (Bray et al., 2023; Cooke et al., 2024; Diel et al., 2024; Groh et al., 2024; Somoray & Miller, 2023). Further, they lack scalability, which makes them ineffective for large volumes of media (Groh et al., 2024). Yet, evidence from studies on traditional misinformation shows that crowd-based approaches, where multiple judgments are aggregated, can achieve accuracy levels comparable to those of experts and overcome scalability issues (Allen et al., 2021; Drolsbach & Pröllochs, 2023b; He et al., 2025; Martel et al., 2024). In contrast, automated detection approaches, e. g., using AI models to analyze inconsistencies in facial expressions, lighting, and audio-visual mismatches, offer a scalable solution. However, they often struggle with accuracy and robustness, leading to false positives or undetected deepfakes (Almars, 2021; Feuerriegel et al., 2023; Zhou et al., 2023). To enable comprehensive and reliable empirical analysis of AI-generated misinformation circulating on social media, there is thus a need for complementary detection strategies that are both scalable and accurate.

Our work: In this study, we characterize AI-generated misinformation that has been identified on X's Community Notes platform, i. e., via the wisdom of crowds. Compared to other identification strategies, this has two key advantages: (i) it overcomes the scalability limitations of human-centered methods (Groh et al., 2024); (ii) it addresses challenges with low accuracy of AI-based approaches for the detection of AI-generated content (Almars, 2021; Feuerriegel et al., 2023; Zhou et al., 2023). This unique data source enables us to empirically analyze what distinguishes AI-generated misinformation from traditional forms of misinformation in terms of reach, user engagement, and content characteristics. Addressing these questions, is a crucial first step for developing effective countermeasures against AI-generated misinformation.

5.3 Data and Methods

5.3.1 Data source

To address our research questions, we analyze a comprehensive dataset of crowd-annotated social media posts flagged as misleading on X's Community Notes between January 2023 and January 2025, i. e., for an observation period of two years.[1] X's Community Notes is a crowdsourced fact-checking system that enables users to add context to potentially misleading posts (Pröllochs, 2022; X, 2021). By incorporating contributions from diverse perspectives, the system strives to provide balanced and informative notes (Wojcik et al., 2022) to help users assess content accuracy (Drolsbach et al., 2024). Each Community Note consists of the assigned label (i. e., whether the post is considered misleading) and a textual description that explains why a post is misleading (e. g., because it is a deepfake; see examples in Fig. 5.1). Further, Community Notes features a rating mechanism, where other users can rate the helpfulness of each note (Solovev & Pröllochs, 2025). A note becomes visible on X for all users if (and only if) it reaches a certain helpfulness rating, reflecting agreement across diverse perspectives rather than simple majority approval. Community fact-checks identified as helpful on the Community Notes platform have been shown to be accurate (Drolsbach & Pröllochs, 2023b) and trustworthy (Drolsbach et al., 2024).

For our data collection, we filtered all Community Notes rated as helpful, resulting in a dataset of 91 452 community fact-checked posts. We downloaded both the source posts

[1] Available via https://communitynotes.x.com/guide/en/under-the-hood/download-data

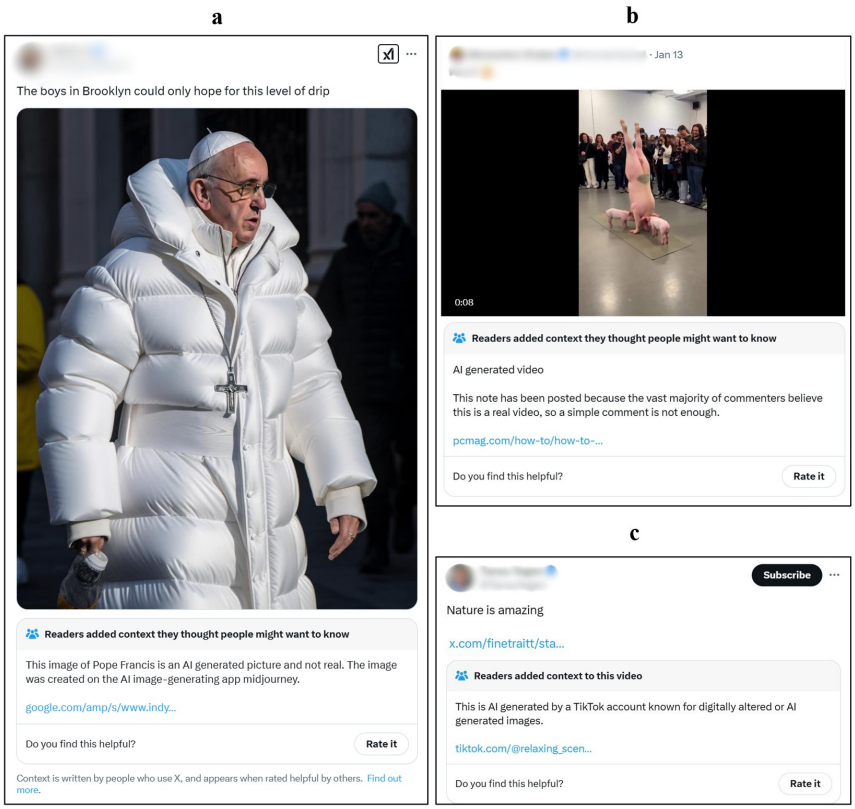


Figure 5.1: Examples of posts on X that shares a deepfake in form of an (**a**) Image, (**b**) Video, or (**c**) no media (here: URL to media content) and the corresponding Community Notes.

and all attached media (i. e., images and videos). To retrieve additional metadata about the fact-checked posts, we mapped the referenced *tweetID* to the original source tweet via the X Research API. This includes key attributes such as the number of retweets, likes and, impressions, the type of attached media as well as details about the author's profile such as the follower count, followee count, total tweet count, account age, and verification status. In addition, we estimated the authors' political partisanship and misinformation exposure on social media using the method proposed by Mosleh and Rand (2022). Partisanship scores range from -1 (Democrat) to $+1$ (Republican) and are based on the number of Democratic and Republican public figures followed by each user. The misinformation exposure score ($[0, 1]$) is derived from the proportion of followed public figures that have been rated false by PolitiFact.[2]

5.3.2 Identification of AI-generated misinformation

To distinguish AI-generated content from other misleading posts, we implemented an LLM-based identification approach. Specifically, we deployed an OpenAI Assistant (based on *gpt-4-turbo*), which was given the task of identifying whether a Community Note refers to AI-generated content. The assistant was instructed to act as a "professional annotator specializing on annotating social media content." The task was complemented by a clear

[2]We use the method available at https://github.com/mmosleh/minfo-exposure. Due to X API restrictions, partisanship and misinformation expose could only be calculated for authors of 32 070 posts in our dataset.

description of what is to be understood as AI-generated content. For each case, it returned a binary classification ("yes" or "no") of whether or not is AI-generated (see *SI, Sec.* Prompt: Identification of AI-Generated Posts for the full prompt).

Compared to alternative strategies (e. g., keyword-based heuristics, purely content-based methods), our approach to identify AI-generated misleading posts via Community Notes offers two main advantages: (i) it leverages crowd-sourced fact-checking explanations (i. e., Community Notes), which contain contextual insights not evident from the media content alone, and (ii) it achieves comparatively high accuracy (see validation below).

Validation: To validate the identification of AI-generated misinformation, we conducted a validation study involving three trained research assistants. Each assistant independently reviewed 400 randomly selected posts ($1/3$ AI and $2/3$ non-AI), including both the original post and the corresponding textual Community Note. For each post, they assessed whether it contained AI-generated content (Yes/No). The participants rated posts categorized as AI-generated significantly more often as AI-generated than posts categorized as non-AI-generated ($M_{AI/AI} = 0.75$, $M_{AI/nonAI} = 0.22$, $t = 19.318$, $p < 0.01$). An analysis of Fleiss's Kappa shows a fair interrater agreement, which is statistically significant ($\kappa = 0.322$, $z = 11.1$, $p < 0.01$). These findings strengthen the validity of our results, showing that individual annotators (that may have varying familiarity with the posts' information) and the LLM align well in their perceptions and categorizations of AI-generated misleading posts.

5.3.3 Annotation of post characteristics

We used an LLM-based classification approach (Chuai, Sergeeva, et al., 2025; Feuerriegel et al., 2025) to annotate the source posts with a wide variety of content characteristics. Since this task required processing not only textual content but also media elements (i. e., images/videos), it was computationally demanding. To balance API costs while ensuring representativeness, we randomly selected a balanced subset of 3 000 posts (1 500 AI/1 500 Non-AI). Subsequently, we deployed an OpenAI Assistant (*gpt-4-turbo*) configured to function as a professional content reviewer. The assistant was presented with the textual elements of the posts *and* the attached media (images or video snapshots). Further, it was provided with the information that all posts had been flagged as misleading by a community-based fact-checking initiative. Its evaluation focused on four core dimensions: *sentiment*, *topic*, *harmfulness*, and *believability*.

Sentiment (Feuerriegel et al., 2025) referred to the emotional tone or attitude expressed in the post, categorized as positive, neutral, or negative. The topic dimension (Feuerriegel et al., 2025) required the assistant to identify the main subject matter as one of the following topics: *Politics*, *Technology*, *Health*, *Crime*, *Business*, *Entertainment*, *Sports*, *Education*, *Satire*, and *Other*. Harmfulness (Drolsbach & Pröllochs, 2023a) assessed the potential of the content to cause real-world damage, including emotional distress, physical harm, or societal disruption, particularly if the misinformation were believed and acted upon (low, medium or high). Believability (Drolsbach & Pröllochs, 2023a) was defined as the degree to which a post could plausibly be accepted as true by a general audience (low, medium or high). These definitions were embedded directly in the system prompt to ensure consistent and interpretable outputs. The full prompt is available in the *SI, Sec.* Prompt: Annotation of post characteristics.

5.4 Empirical Analysis

5.4.1 Content characteristics (RQ1)

Our final dataset includes 91 452 posts across more than 60 languages, with 4 577 posts (i. e., 5.06%) identified as containing AI-generated content. English dominates the dataset, accounting for half of all posts (50.05%), followed by Japanese (12.15%), Spanish (10.94%), French (7.85%), Portuguese (5.83%), and German (2.61%). A long tail of low-frequency languages reflects the dataset's global scope, though many languages are represented by only a handful of posts. These posts contain either an image (55.95%), a video (37.16%), a GIF (0.02%), or no media (6.86%) (see Fig. 5.2). The latter mostly consists of URLs linking to AI-generated content on another website or are reposts of other posts containing media. Compared to other forms of misinformation, AI-generated misleading posts are 1.33 times more likely to contain media elements (i. e., images, video, or GIFs). Overall, the numbers reflect the growing importance of visual content on social media.

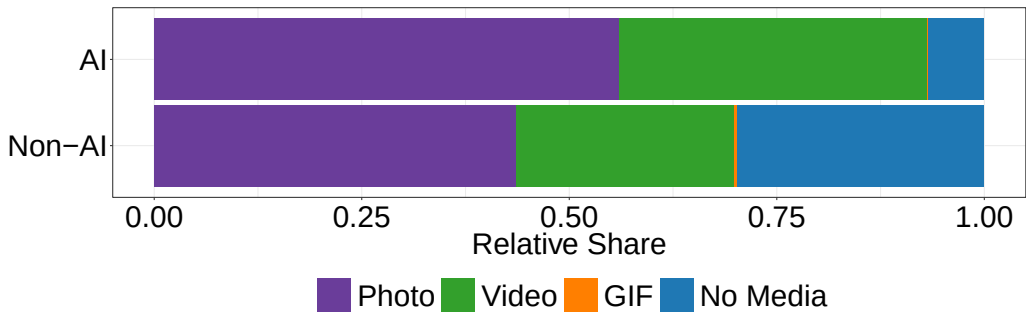


Figure 5.2: Distribution of media types/modalities in AI-generated vs. non-AI-generated misleading posts. $N = 91\,452$.

Sentiment: We applied LLM-based sentiment classification on a subset of the dataset ($N = 3000$) that incorporates both the textual content of each post *and* any attached media (see Sec. Annotation of post characteristics). Figure 5.3 presents the distribution of posts classified as negative, neutral or positive. The results indicate that a higher proportion of AI-generated posts exhibit positive sentiment compared to non-AI-generated posts (24% vs. 16%). In contrast, negative sentiment is more prevalent among non-AI-generated posts (51% vs. 43%). Further, Pearson's Chi-squared test revealed a statistically significant association between AI generation status and sentiment classification ($\chi = 36.85$, $p < 0.01$). This indicates that the distribution of sentiment categories differs systematically between AI-generated and non-AI-generated posts and suggests that AI-generated misinformation may be more likely to adopt a humorous or entertaining tone.

As a validation, we calculated sentiment scores on the full dataset ($N = 91,452$), this time using only the textual content of each post (i. e., without attached media).[3] Consistent with our main analysis, we find that the mean sentiment score is significantly higher for AI-generated posts than for non-AI-generated ones ($t = -3.63$, $p < 0.01$). However, the purely text-based approach classified a larger proportion of posts as neutral. This likely reflects the fact that

[3]We used the *Twitter-roBERTa-base model*, which is fine-tuned on tweets, to calculate sentiment scores for the textual content of each post (Barbieri et al., 2020; Loureiro et al., 2022).

much of the emotional tone in these posts is conveyed through attached media, rather than text alone.

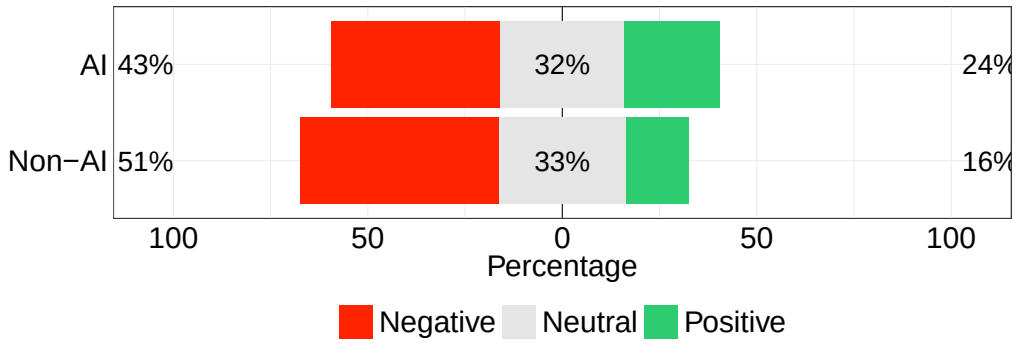


Figure 5.3: Distribution of sentiment in AI-generated vs. non-AI-generated misleading posts. $N = 3\,000$.

Topics: Based on the LLM-based topic classification (including both textual content & media), we observe notable differences in the distribution of specific topics between AI-generated and non-AI-generated content (see Fig. 5.4). Specifically, 30.40% of AI-generated posts are associated with the topic *Entertainment*, while only 16.60% of non-AI-generated posts focus on this topic. This suggests that AI-generated content may be more likely to focus on lighter, more engaging topics. In contrast, the share of posts related to the topic *Health* is higher among non-AI-generated content (9.40% vs. 2.27%), indicating a potential difference in the focus areas of human versus AI creators. For all other topics, the distribution remains largely similar between the two groups (i. e., the differences are $\leq 2.00\%$).

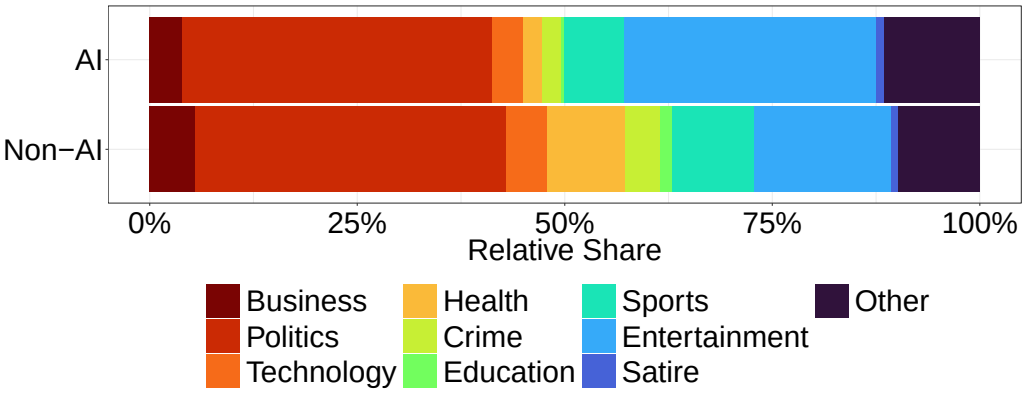


Figure 5.4: Distribution of topics in AI-generated vs. non-AI-generated misleading posts. $N = 3\,000$.

5.4.2 Author characteristics (RQ2)

Next, we analyze how the characteristics of authors of AI-generated misinformation differ from those of other forms misinformation (see Fig. 5.5). We observe that AI-generated misleading posts tend to originate from accounts with significantly more followers ($950,660$ vs. $585,671$), but fewer followees (6072 vs. 8153), and are, on average, older (8.32 vs. 7.31 years). These accounts have also posted and interacted with more content throughout their entire lifetime, with a higher mean tweet count ($73,688$ vs. $61,047$). The tweet count includes original posts, replies, retweets and quoted posts. Further, the number of posts per

author is lower for AI-generated posts (1.55 vs. 2.78), which may reflect the higher effort required to create and disseminate AI-generated misinformation. Both two-sided t-tests and Kolmogorov–Smirnov (KS) tests confirm that these differences between AI-generated and non-AI-generated misinformation are statistically significant (each $p < 0.01$).

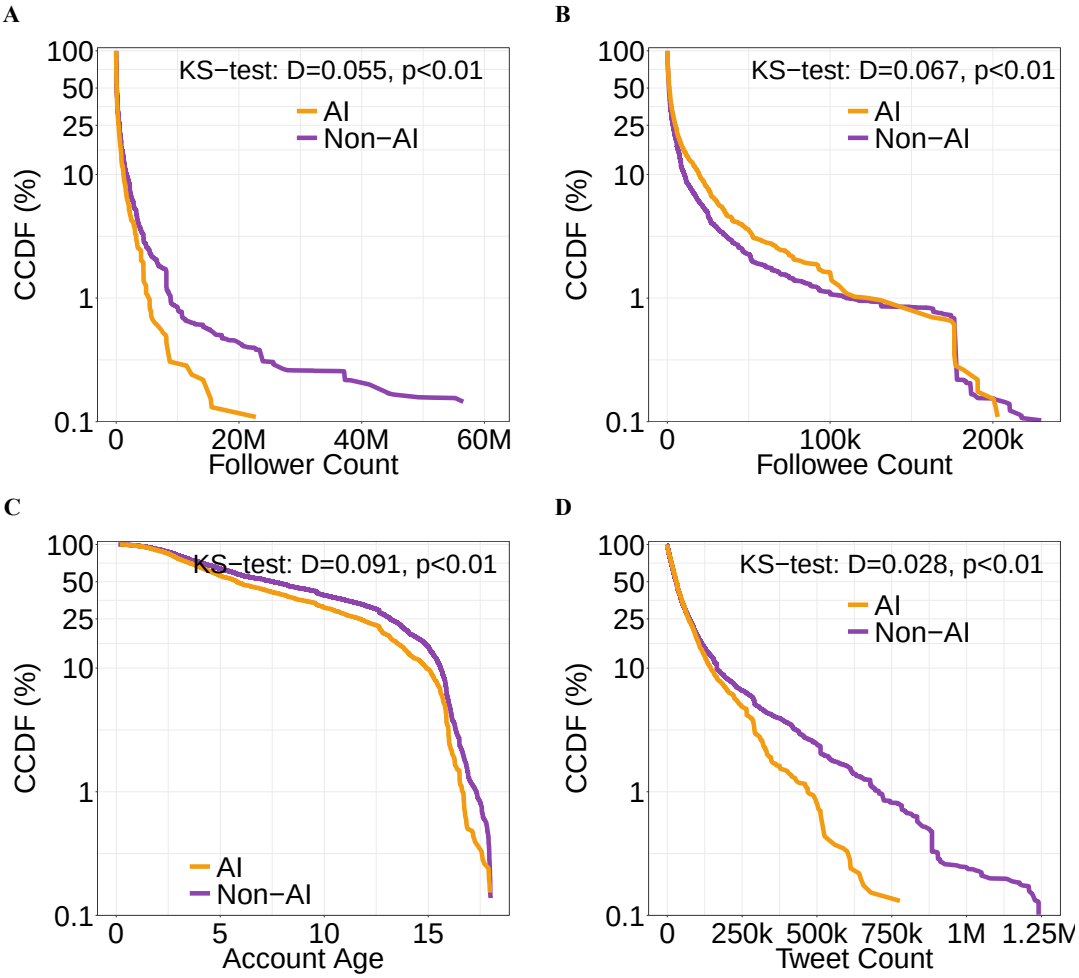


Figure 5.5: Complementary cumulative distribution functions (CCDFs) for AI vs. non-AI generated misleading posts, shown separately for follower counts (**a**), followee counts (**b**), account age (**c**), and the number of posts per author (**d**). $N = 91\,452$.

Partisanship and misinformation exposure: We observe a significantly higher mean partisanship score among authors of AI-generated posts compared to those of non-AI-generated posts (see Fig. 5.6), indicating that authors of AI-generated content tend to be more conservative. The partisanship score ranges from −1 (Democrat) to 1 (Republican), with higher values indicating greater conservatism (Mosleh & Rand, 2022). Specifically, the mean partisanship score for AI-generated posts is 0.272, compared to 0.199 for non-AI-generated posts. While both means fall on the conservative side of the scale, the higher value for AI-generated content suggests a notable shift toward more right-leaning users among those sharing such posts. This difference is statistically significant according to a two-sided $t-$test ($t = -3.43$, $p < 0.01$). In contrast, we find no statistically significant differences in misinformation exposure between the two groups ($t = -1.665$, $p = 0.09$).

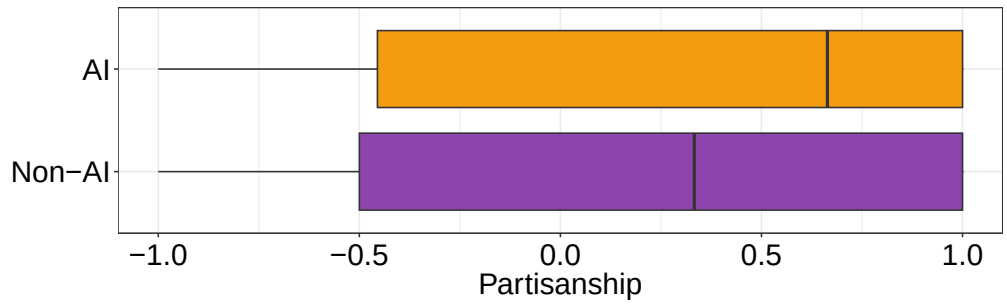


Figure 5.6: Boxplot of partisanship scores for AI vs. non-AI generated misleading posts ($N = 32\,070$). The partisanship scores range from −1 (Democrat) to 1 (Republican), with higher values indicating greater conservatism Mosleh and Rand, 2022.

5.4.3 Virality (RQ3)

The posts in our dataset have generated more than 200 billion impressions, have been reposted more than 137 million times and liked more than 969 million times. On average, AI-generated misinformation received 8.19% more impressions, 20.54% more reposts and 49.42% more likes.

Regression model: To better understand the virality of AI-generated vs. other types of misinformation and account for confounding factors, we implement three negative binomial regression models explaining the number of (i) retweets, (ii) likes and (iii) impressions (e. g., Chuai, Pilarski, et al., 2024; Chuai, Tian, et al., 2024; Rathje et al., 2021). The key explanatory variable is $AIGenerated_i$, i. e., whether the post contains media that was generated using AI (= 1) or not (= 0). Consistent with previous work (e. g., Drolsbach & Pröllochs, 2023b; Stieglitz & Dang-Xuan, 2013; Vosoughi et al., 2018), we control for content characteristics (i. e., media type), and account characteristics (i. e., follower count, followee count, account age, verification status).

Each model takes the following form:

$$\begin{aligned}\log\left(\mathbb{E}[Y_i \mid \mathbf{X}_i]\right) = \; & \beta_0 + \beta_1 \, \textit{AIGenerated}_i \\ & + \beta_2 \, \textit{MediaType}_i + \beta_3 \, \textit{Followers}_i + \beta_4 \, \textit{Followees}_i \\ & + \beta_5 \, \textit{AccountAge}_i + \beta_6 \, \textit{Verified}_i + u_t, \end{aligned} \tag{5.1}$$

where Y_i denotes the outcome variable (retweets, likes, impressions) for post i, $\mathbf{X}_i$ is the vector of explanatory variables, and β_0 the intercept. Furthermore, we include month-year fixed effects u_t, which allow us to control for variation in exposure windows and platform dynamics over time (e. g., Drolsbach & Pröllochs, 2023a, 2023b). In our regression analysis, all continuous variables are z-standardized to aid interpretability.

Coefficient estimates: We find that AI-generated misleading posts are substantially more viral than other forms of misinformation. On average, they receive $e^{0.104} - 1 = 10.81\%$ more retweets (coef: 0.103, $p < 0.01$), 34.16% more likes (coef: 0.294, $p < 0.01$), and 10.32% more impressions (coef: 0.098, $p < 0.01$), compared to non-AI-generated misleading posts (see Fig. 5.7).

Regarding the control variables, we observe that media content plays a significant role in driving user engagement. Compared to posts without media, those containing images receive

30.46% more retweets (coef: 0.266, $p < 0.01$), while posts with videos receive 53.86% more retweets (coef: 0.431, $p < 0.01$). Similar patterns are observed across other engagement metrics, except for the impression count, which shows no significant effect of images. Posts containing images or videos receive significantly higher engagement overall ($p < 0.01$), while GIFs show a significant effect only for retweets – likely due to the small number of posts with GIFs attached in the dataset.

The effects of social influence variables are also statistically significant ($p < 0.01$) and align with findings from prior work (e. g., Drolsbach & Pröllochs, 2023b). Higher follower and followee counts are associated with higher engagement, while a higher account age is associated with lower engagement. Interestingly, the verification status exhibits a strong positive effect on impression count but a negative effect on both retweet count and like count (all $p < 0.01$). This suggests that verified users benefit from greater visibility, likely due to algorithmic amplification on X, which boosts their content in users' feeds. However, after controlling for other account characteristics (e. g., follower count) this increased exposure does not translate into higher user engagement.

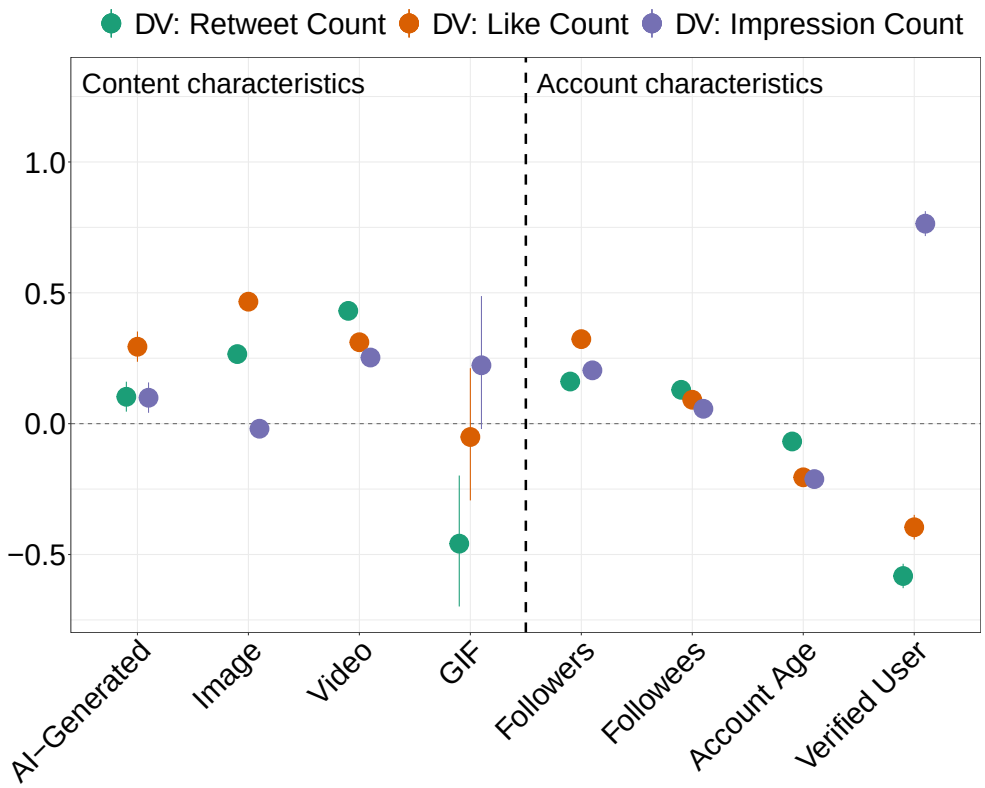


Figure 5.7: Negative binomial regression with the retweet count, like count, and impression count as DVs. The circles show standardized coefficient estimates and the vertical bars represent 99% confidence intervals. Month-year fixed effects are included. $N = 91\,452$.

Model checks: (1) We checked that variance inflation factors as an indicator of multicollinearity are below the critical value of five (Akinwande et al., 2015). (3) We found confirmatory results when estimating separate regressions for AI-generated vs. not AI-generated posts. (iii)

We repeated our regression analysis controlling for the topic and sentiment of the fact-checked posts. Here we used the subset of 3 000 LLM-annotated posts for which we have access to topic and sentiment labels (see Sec. Annotation of post characteristics). We found that AI-generated posts are more viral than other misleading posts even after controlling for differences in topics and sentiment.

5.4.4 Harmfulness & believability (RQ4)

As AI models become increasingly capable of mimicking human communication, their ability to produce more credible and potentially more harmful misinformation grows (Feuerriegel et al., 2023). Understanding characteristics such as believability and harmfulness is therefore critical for assessing the risks associated with AI-generated content and informing effective detection and mitigation strategies (Drolsbach & Pröllochs, 2023a). To address this, we analyzed the LLM-based annotation of post believability and harmfulness (see Section Annotation of post characteristics), comparing AI-generated and non-AI-generated misinformation.

We observe that most posts are perceived as not particularly harmful, with 66% of AI-generated and 62% of non-AI-generated posts rated as low in harmfulness. Ratings of medium and high harmfulness are less common: 23% vs. 27% for medium, and 10% vs. 11% for high (AI vs. non-AI, respectively). In terms of believability, the largest share of posts is perceived as highly believable, with 45% of AI-generated and 42% of non-AI-generated posts rated as such. The remaining distributions are 32% vs. 30% for low believability, and 23% vs. 28% for medium believability (AI vs. non-AI).

Although the absolute differences between AI-generated and non-AI-generated posts in terms of harmfulness and believability are relatively small, Chi-squared tests imply that they are partially statistically significant. For believability, AI-generated posts are more likely to be rated as less believable, while non-AI-generated posts are more likely to be rated as at least medium believable ($\chi = 22.827$, $p < 0.01$). For harmfulness, non-AI-generated posts are slightly more likely to be classified as medium or highly harmful, while AI-generated posts are more likely to be categorized as less harmful ($\chi = 7.317$, $p = 0.03$).

Overall, despite the statistical significance, these differences are quite small. Still, the data suggests a slight tendency that AI-generated posts are both less believable and less harmful.

5.5 Discussion

AI-generated misinformation represents a growing challenge in the digital information landscape. Different from traditional forms of misinformation, it leverages advanced AI technologies to create highly realistic yet deceptive content, making detection increasingly difficult (Feuerriegel et al., 2023; Groh et al., 2024; Hancock & Bailenson, 2021; Vaccari & Chadwick, 2020). Despite these risks, research has largely focused on the societal consequences of AI-generated misinformation (Dobber et al., 2021; Hancock & Bailenson, 2021; Vaccari & Chadwick, 2020) rather than their real-world prevalence. Here, we contribute by conducting a large-scale empirical analysis of AI-generated misinformation circulating on the social media platform X.

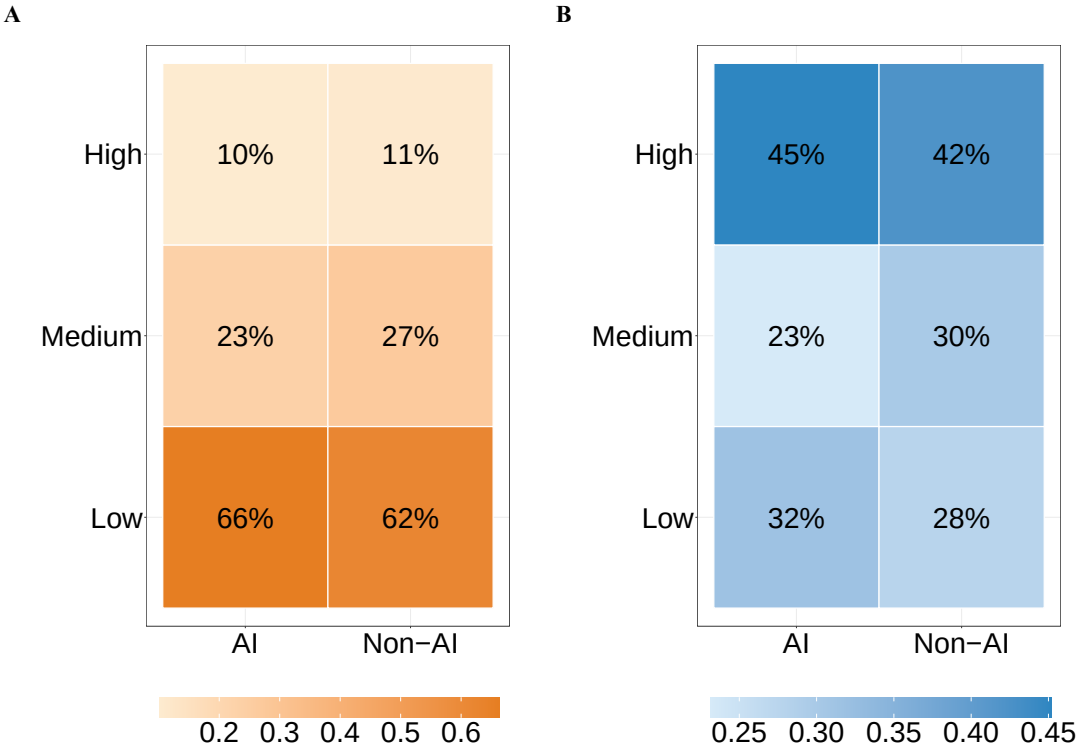


Figure 5.8: Harmfulness (**a**) and believability (**b**) of AI-generated vs. non-AI-generated misleading posts. $N = 3\,000$.

5.5.1 Implications

Our findings highlight the distinct characteristics of AI-generated misinformation and their unique role within the misinformation ecosystem. Compared to other forms of misinformation, AI-generated content is more frequently focused on entertainment and tends to exhibit a more positive sentiment (*RQ1*). Yet, even when accounting for content differences, AI-generated misleading posts are disproportionately more likely to go viral (*RQ2*) – despite frequently originating from smaller accounts (*RQ3*). Moreover, we find that AI-generated misinformation is similarly believable and harmful as traditional forms of misinformation (*RQ4*). This underscores the capacity of generative AI models to produce highly realistic and persuasive misleading content.

These patterns suggest that the AI-generated nature of content should be treated as a distinct and meaningful factor in misinformation research. Its disproportionate virality points to underlying persuasive properties that may not be captured by existing content-based or intent-based categorizations (Tandoc Jr et al., 2018). To account for these properties, future studies should consider incorporating AI-generated content as an explanatory variable in empirical models of misinformation spread and engagement. Our results also motivate further research into the psychological and perceptual mechanisms that make AI-generated misinformation so compelling. Future research should examine the factors driving their spread (e. g., emotional appeal, visual realism, novelty), the demographics most engaged with them, and how they shape user interactions differently from traditional forms of misinformation. types of misleading information.

For platforms, the high virality of AI-generated misinformation highlights the urgency to

develop more effective countermeasures (Feuerriegel et al., 2023). Traditional strategies (e. g., expert-based fact-checking) often focus on high-profile accounts or recurring misinformation themes (Chuai, Zhao, et al., 2025; Greene et al., 2025). However, our results demonstrate that smaller accounts are disproportionately responsible for spreading AI-generated misinformation. This indicates that detection strategies must move beyond account size as a primary signal. Here, community-based fact-checking systems – such as X's Community Notes – can be a promising tool by leveraging the collective judgment of users to identify and flag misleading content that may evade both professional fact-checkers and automated systems (Pilarski et al., 2024). In addition, improved cross-platform coordination could potentially improve consistency and enable more effective mitigation strategies across the broader digital ecosystem.

AI-generated misinformation also presents a growing challenge to user trust and safety online (Feuerriegel et al., 2023; Goldstein et al., 2023; Hancock & Bailenson, 2021). As AI-generated content becomes increasingly indistinguishable from authentic content, users face greater challenges in evaluating the authenticity of the information they encounter and share (Bashardoust et al., 2024; Diel et al., 2024; Groh et al., 2024). This makes media literacy training (Guess et al., 2020; Jones-Jang et al., 2021) more essential than ever. Educational initiatives should help users develop skills to critically assess AI-generated misinformation, recognize manipulation techniques, and understand its broader societal implications (Feuerriegel et al., 2023; Goldstein et al., 2023). Such training is particularly vital for high-stakes domains (e. g., elections, health), where the spread of AI-generated misinformation can have serious real-world consequences (Bär et al., 2023).

5.5.2 Limitations & future research

As any other research, our study is not free of limitations. First, our analysis is observational and cannot establish causal relationships. However, our observational approach allows us to uncover robust patterns based on real-world social media data that would be difficult to study in controlled settings. Second, our data is limited to a single platform (i. e., X), which may constrain the generalizability to other platforms with different user bases or moderation practices. Still, given X's central role in shaping public discourse, it provides a particularly important setting for investing emerging forms of misinformation. Third, we rely on X's Community Notes system to identify misinformation, which inherently only captures content that is flagged by users. As such, some AI-generated misleading posts likely go undetected, potentially biasing our view toward more visible or controversial cases. However, this crowdsourced approach offers a unique advantage in terms of scalability and accuracy, enabling the identification of a wide range of misinformation without relying on manual annotation or automated approaches. Finally, future research should examine how fact-checking effectiveness and visibility differ for AI-generated versus traditional misinformation, and how corrections influence downstream user behaviors (e. g., sharing). Overall, continued research is needed to understand and mitigate the evolving risks posed by AI-generated misinformation on social media.

5.6 Conclusion

Despite growing concerns, the characteristics of AI-generated misinformation on social media are poorly understood. Our work addresses this gap by offering a large-scale empirical analysis

of AI-generated misinformation circulating on X. Drawing on a dataset of 91 452 misleading posts identified via X's Community Notes platform, we show that AI-generated misinformation differs significantly from traditional forms in terms of content attributes, source accounts, and virality. To effectively address this emerging threat, researchers, platforms, and policymakers will need to develop new strategies and countermeasures that account for the unique properties of AI-generated misinformation.

5.7 Ethics Statement

All analyses are based on publicly available data. We declare no competing interests.

Bibliography

Akinwande, M. O., Dikko, H. G., Samson, A., et al. (2015). Variance inflation factor: As a condition for the inclusion of suppressor variable (s) in regression analysis. *Open Journal of Statistics*, *5*(7), 754–767.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393.

Almars, A. M. (2021). Deepfakes detection techniques using deep learning: A survey. *Journal of Computer and Communications*, *9*, 20–35.

Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). New threats to society from free-speech social media platforms. *Communications of the ACM*, *66*(10), 37–40.

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Bashardoust, A., Feuerriegel, S., & Shrestha, Y. R. (2024). Comparing the willingness to share for human-generated vs. ai-generated fake news. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2).

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, *9*(1), tyad011.

Chuai, Y., Pilarski, M., Renault, T., Restrepo-Amariles, D., Troussel-Clément, A., Lenzini, G., & Pröllochs, N. (2024). Community-based fact-checking reduces the spread of misleading posts on social media. *arXiv preprint arXiv:2409.08781*.

Chuai, Y., Sergeeva, A., Lenzini, G., & Pröllochs, N. (2025). Community fact-checks trigger moral outrage in replies to misleading posts on social media. *CHI*.

Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the roll-out of community notes reduce engagement with misinformation on x/twitter? *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), 1–52.

Chuai, Y., Zhao, J., Pröllochs, N., & Lenzini, G. (2025). Is fact-checking politically neutral? asymmetries in how us fact-checking organizations pick up false statements mentioning political elites. *ICWSM*.

Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2024). As good as a coin toss: Human detection of ai-generated images, videos, audio, and audiovisual stimuli. *arXiv preprint arXiv:2403.16760*.

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, *16*, 100538.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & De Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, *26*(1), 69–91.

Drolsbach, C. P., & Pröllochs, N. (2023a). Believability and harmfulness shape the virality of misleading social media posts. *WWW*.

Drolsbach, C. P., & Pröllochs, N. (2023b). Diffusion of community fact-checked misinformation on Twitter. *CSCW*.

Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS nexus*, *3*(7), pgae217.

Feuerriegel, S., DiResta, R., Goldstein, J. A., Kumar, S., Lorenz-Spreen, P., Tomz, M., & Pröllochs, N. (2023). Research can help to tackle ai-generated disinformation. *Nature Human Behaviour*, *7*, 1818–1821.

Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C. E., Rathje, S., Hartmann, J., Mohammad, S. M., et al. (2025). Using natural language processing to analyse text data in behavioural science. *Nature Reviews Psychology*, *4*, 96–111.

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Greene, K. T., Pisharody, N., Carroll, F., & Shapiro, J. N. (2025). Fact-checks focus on famous politicians, not partisans. *PNAS Nexus*, *4*(1), pgae567.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378.

Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. (2024). Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications*, *15*(1), 7629.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences (PNAS)*, *117*(27), 15536–15545.

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 149–152.

He, B., Hu, Y., Lee, Y.-C., Oh, S., Verma, G., & Kumar, S. (2025). A survey on the role of crowds in combating online misinformation: Annotators, evaluators, and creators. *ACM Transactions on Knowledge Discovery from Data*, *19*(1), 1–30.

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? information literacy helps, but other literacies don't. *American Behavioral Scientist*, *65*(2), 371–388.

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11).

Kreps, S. E., & Kriner, D. L. (2022). The COVID-19 infodemic and the efficacy of interventions intended to reduce misinformation. *Public Opinion Quarterly*, *86*(1), 162–175.

Loureiro, D., Barbieri, F., Neves, L., Anke, L. E., & Camacho-Collados, J. (2022). Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2024). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, *19*(2), 477–488.

Montserrat, D. M., Hao, H., Yarlagadda, S. K., Baireddy, S., Shao, R., Horváth, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F., et al. (2020). Deepfakes detection with automatic face weighting. *IEEE/CVPR*, 668–669.

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, *13*(1), 7144.

Pilarski, M., Solovev, K., & Pröllochs, N. (2024). Community notes vs. snoping: How the crowd selects fact-checking targets on social media. *ICWSM*.

Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. *ICWSM*.

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118.

Rivers, C. M., & Lewis, B. L. (2014). Ethical research standards in a world of big data. *F1000Research*, *3*, 38.

Sippy, T., Enock, F., Bright, J., & Margetts, H. Z. (2024). Behind the deepfake: 8% create; 90% concerned. surveying public exposure to and perceptions of deepfakes in the uk. *arXiv preprint arXiv:2407.05529*.

Solovev, K., & Pröllochs, N. (2025). References to unbiased sources increase the helpfulness of community fact-checks. *arXiv preprint arXiv:2503.10560*.

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior*, *149*, 107917.

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, *29*(4), 217–248.

Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153.

Timmerman, B., Mehta, P., Deb, P., Gallagher, K., Dolan-Gavitt, B., Garg, S., & Greenstadt, R. (2023). Studying the online deepfake community. *Journal of Online Trust and Safety*, *2*(1).

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, *6*(1).

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*, 40–53.

Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M., Coleman, K., & Baxter, J. (2022). Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv*. https://arxiv.org/abs/2210.15723

X. (2021). Introducing Birdwatch, a community-based approach to misinformation.

Yan, H. Y., Morrow, G., Yang, K.-C., & Wihbey, J. (2025). The origin of public concerns over ai supercharging misinformation in the 2024 us presidential election. *Harvard Kennedy School Misinformation Review*.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. *CHI*.

Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y.-G. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. *MM*, 2382–2390.

Appendix 5.A Descriptive statistics

An overview of the dataset used in this study is shown in Table 5.1.

Table 5.1: Variable Definitions and Summary Statistics.

Variable	Description	Overall		AI		Non-AI	
		Mean	SD	Mean	SD	Mean	SD
Content Characteristics (N = 91 452)							
AI-generated	Tweet identified as AI-generated (= 1 if true, else 0)	0.120	0.330	1.000	0.000	0.000	0.000
Media Type	Media type in tweet (e.g., image, video, none)			–			
Sentiment	Sentiment score (Twitter RoBERTa-base model)	0.040	0.670	0.010	0.650	0.050	0.680
Content Characteristics (LLM-Annotated, N = 3 000)							
Sentiment (LLM)	Sentiment score (-1 = Negative, $+1$ = Positive)	−0.270	0.775	−0.184	0.801	−0.356	0.738
Topic	Topic category (*Politics*, *Tech*, *Health*, etc.)			–			
Believability	Perceived as believable (= 1 if true)	0.340	0.470	0.580	0.490	0.310	0.460
Harmfulness	Perceived as harmful (= 1 if true)	0.410	0.490	0.660	0.470	0.370	0.480
Account Characteristics							
Followees	Accounts followed (in 1000s)	6.176	24.002	8.152	30.273	6.072	23.621
Followers	Followers (in 1000s)	932.393	7 515.512	585.671	7 622.054	950.660	5 074.501
Account Age	Account age (in years)	8.217	5.075	7.311	4.811	8.322	5.078
Tweet Count	Posts, reposts, quotes & replies (in 1000s)	73.055	141.309	61.047	93.267	73.688	143.357
Partisanship	Political leaning (-1 = Democrat, $+1$ = Republican)	0.202	0.769	0.272	0.782	0.199	0.768
Misinformation Exposure	Exposure to misinformation $[0, 1]$	0.599	0.170	0.606	0.175	0.298	0.170
Virality							
Retweet Count	Number of retweets (in 1000s)	1.469	3.284	1.753	3.570	1.454	3.268
Impression Count	Number of impressions (in 1000s)	2 157.207	7 044.313	2 324.291	6 222.344	2 148.404	7 084.897
Like Count	Number of likes (in 1000s)	10.326	25.370	15.057	32.628	10.077	24.904

Appendix 5.B Prompt: Identification of AI-Generated Posts

To distinguish AI-generated content from other misleading posts, we implemented an OpenAI Assistant (*gpt-4-turbo*). The corresponding prompt is printed below.

"You are a professional annotator specializing on annotating social media content. Your task is to classify social media posts based on Community Notes. The Community Notes have been added for fact-checking purposes to the social media posts, have been found helpful by other users and provide additional context to the social media post. Each post must be annotated along two dimensions: AI-generated, IdentificationConfidence

Your goal is to produce consistent, structured annotations. Follow the definitions and formatting instructions strictly. Return only a valid JSON object with exactly two top-level keys: 'AI-generated' and "IdentificationConfidence".

Definitions and Output Format:

1. *AI-generated*

- *Definition: Identify whether you expect the original post to contain AI-generated content based on the community note provided. AI-generated content refers to any form of text, image, audio, video, or other media that is created with the assistance of artificial intelligence technologies. These technologies include machine learning models—such as large language models (LLMs), image generators, or speech synthesis systems—that can mimic human creativity, language, or perception. Return only the classification (yes or no).*
- *Output format: "Yes" // One of: Yes, No*

2. *Identification confidence*
    - *Indicate how confident you are regarding your identification of AI-generated content as provided in 1.*
    - *Output format: "Medium" // One of: Low, Medium, High*

Return only a valid JSON object using the structure above. Do not include any explanation, comments, or metadata."

Appendix 5.C Prompt: Annotation of post characteristics

We implemented an OpenAI Assistant (*gpt-4-turbo*) to annotate the source posts with a wide variety of content characteristics. The corresponding prompt is printed below.

"You are a professional content reviewer specializing in analyzing social media posts, including both the tweet text and attached media (images or video snapshots). All social media posts for comment were identified as misleading by a community-based fact-checking initiative. Your task is to assess the content across the following four dimensions: Believability, Sentiment, Harmfulness, and Topic. Consider both text and visual media in your assessment. Return only a valid JSON object using the structure and formatting below. Do not include any explanation, comments, or metadata.

Definitions and Output Format:

1. *Believability*
    - *Definition: Rate the believability of the social media post (text + attached media). Believability is defined as the quality of something being convincing or credible enough that people accept it as true or plausible.*
    - *Output format: "believability": "High" // One of: Low, Medium, High*
2. *Sentiment*
    - *Definition: Rate the sentiment of the social media post (text + attached media). Sentiment refers to the underlying emotion or attitude conveyed in a message, reflecting how the author feels about a subject, whether it's positive, negative, or neutral.*
    - *Output format: "sentiment": "Negative" // One of: Positive, Neutral, Negative*
3. *Harmfulness*

- *Rate the harmfulness of the social media post (text + attached media). Harmfulness is defined as the potential of a (misleading) post to cause real-world damage (such as physical harm, social division, or emotional distress) especially when people believe and act on the misinformation.*
- *Output format: "harmfulness": "Medium" // One of: Low, Medium, High*

4. *Topic*
   - *Identify the main topic of the post (text + attached media)*
   - *Output format: "topic": "Politics" // One of: Technology, Health, Politics, Crime, Business, Entertainment, Sports, Education, Satire, Other"*

Chapter 6

Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database

Abstract

The Digital Services Act (DSA) represents a major legislative framework that mandates large social media providers to file Statements of Reasons (SoRs) to the DSA Transparency Database whenever they remove or restrict access to certain content on their platforms in the EU. In this work, we empirically analyze this unique data source and provide an early look at content moderation decisions of social media platforms in the EU. Our empirical analysis based on more than 156 million SoRs reveals significant differences in content moderation practices and how large social media platforms implement their obligations under the DSA. Our findings have important implications for regulators, suggesting the need to lay out more specific rules that ensure common standards on how social media providers handle rule-breaking content on their platforms.

Keywords: Content moderation, social media, online harms, DSA, EU

6.1 Introduction

Social media platforms have become essential channels for accessing information that offer various benefits but also facilitate the dissemination of harmful or illegal content (e. g., hate speech, calls for violence, disinformation). Concerns about such content's impact on elections public health, and safety have grown recently (e. g., Bär et al., 2023; Feuerriegel et al., 2023). In response, platforms employ diverse content moderation systems, i. e., mechanisms that aim to prevent harm by removing or reducing the visibility of rule-breaking content (Grimmelmann, 2015). However, their degree of strictness varies greatly between platforms, with each platform keeping the specifics of how it enacts its moderation decisions largely opaque (Jhaver et al., 2019). Accordingly, there have been increasing calls for legislation to revise the present model of self-regulation, where social media platforms primarily define the rules and procedures of online content moderation (Schlag, 2023; Turillazzi et al., 2023).

The EU, recognizing the need for greater control and oversight, has recently introduced the Digital Services Act (DSA) that establishes a new set obligations for social media providers (Schlag, 2023; Turillazzi et al., 2023). Major platforms can now be held responsible for the risks illegal content on their platforms poses to society. The DSA aims to create a harmonized legal framework that avoids inconsistencies and uncoordinated procedures in content moderation. For this purpose, providers of large social media platforms are required to file *Statements of Reasons* (SoRs) explaining why content was moderated, by reference to the specific legal provision infringed. To ensure scrutiny of content moderation decisions and transparency for both platforms and users, all SoRs are made publicly available by the EU via the DSA Transparency Database (DSA-TDB).

Research goal: Here, we provide a holistic early look at the DSA-TDB. Due to this unique data source, we are, for the first time, able to empirically analyze real-world content moderation decisions of social media platforms in the EU. Specifically, we are interested in what types of content are typically moderated (**RQ1**), what the specific reasons for content moderation are (**RQ2**), to what extent content moderation is automated by the platforms (**RQ3**) and what types of moderation actions the platforms implement (**RQ4**).

Data & Methods: To answer our research questions, we collected all SoRs submitted to the DSA-TDB by major social media platforms within the first two months after its launch on September 25, 2023. Based on over 156 million SoRs, each representing a single content moderation action, we extracted and analyzed a wide variety of variables (e.g., content types, legal grounds) to understand content moderation decisions in the EU. Additionally, we implemented regression analysis to characterize which types of content moderation decisions are more likely to be performed automatically (i. e., without human intervention) by social media platforms.

Contributions: Our work presents the first empirical analysis of the EU's DSA-TDB, yielding the following key findings: (i) There are vast differences in the frequency of content moderation across platforms, with TikTok performing over 350 times more moderation decisions per user than X. (ii) Text and videos are most commonly moderated, while images and other formats undergo less frequent moderation. (iii) Primary reasons for moderation include content outside platform scope, illegal/harmful speech, and pornography, while moderation for misinformation is relatively rare. (iv) Automated methods predominantly detect and decide upon rule-breaking content. (v) Content moderation actions vary substantially across platforms, with some favoring post removal and others visibility reduction. Overall, our study highlights inconsistencies in

how platforms implement their obligations under the DSA, suggesting the need to lay out more specific rules that ensure common standards on how providers handle rule-breaking content.

6.2 Background

Content moderation: Content moderation aims to prevent the spread of illegal and undesirable content in online communities (Grimmelmann, 2015). Social media platforms use various measures like content removal, visibility reduction, labeling, or account actions to address rule-breaking content (Jiang et al., 2023). Effective moderation is crucial for social network functionality, promoting guideline compliance, and reducing incivil behavior (Gillespie, 2018; Horta Ribeiro et al., 2023). Over the last couple of years, content moderation systems on mainstream platforms have become increasingly sophisticated, evolving from relatively small teams overseeing the content to the usage of automated systems that detect and intervene when rule-breaking behavior occurs (Gillespie, 2018). Notably, each social media platform has developed various systems to implement these processes (Gillespie, 2018); yet each platform keeps the specifics of how it enacts its moderation decisions opaque (Jhaver et al., 2019).

Digital Services Act: The Digital Services Act (DSA) stands as a pivotal legislative framework crafted by the EU, which aims to modernize the digital landscape and ensure safer, more open platforms (Schlag, 2023; Turillazzi et al., 2023). Adopted in July 2022 and effective since November 16, 2022, it imposes requirements on online platform providers. All Very Large Online Platforms (VLOPs) with over 45 million users in the EU (corresponding to 10% of the overall population) must meet DSA obligations since August 25, 2023. This includes regular transparency reports detailing content moderation measures, user reports, error rates of automated systems, and team qualifications (Article 15). Furthermore, as mandated by Article 17 of the DSA, VLOPs are required to provide detailed Statements of Reasons (SoRs) for any content moderation activity. The intention is to inform users about content moderation decisions and explain the underlying reasons. In accordance with Article 24(5) of the DSA, all submitted statements are to be collected and made publicly available on the DSA Transparency Database (DSA-TDB). SoRs are to be clear and specific and easily comprehensible and as precise and specific as reasonably possible under the given circumstances.

6.3 Data

We downloaded *all* SoRs from the website of the DSA-TDB, that were submitted between the introduction of the database on September 25, 2023, and November 25, 2023, i. e. within an observation period of two months. In total, more than 550M SoRs were transmitted during the observation period. As we focus on content moderation on social media, we only included SoRs submitted by large social media platforms, namely *Facebook, Instagram, YouTube, TikTok, Snapchat, X, LinkedIn* and *Pinterest*. Furthermore, we excluded SoRs for content published before August 25, 2023, as the obligations associated with the DSA framework became effective from this date. The resulting dataset contains more than 156M SoRs, each including information on one content moderation action.

In our data, the largest number of SoRs was submitted by TikTok (100.15M; 64.09%), followed by Facebook (33.70M; 21.56%), Pinterest (12.45M; 7.97%), YouTube (5.12M; 3.28%), and Instagram (3.95M; 2.53%). Snapchat (0.61M; 0.39%), X (0.27M; 0.17%), and LinkedIn (0.03M; 0.02%) submitted less than one million SoRs during our observation period. Evidently,

TikTok moderates substantially more content than all other platforms, some of which are much more relevant in the European market in terms of user numbers (e. g., YouTube, Facebook and Instagram).

6.4 Empirical Analysis

6.4.1 Content Types (RQ1)

When submitting a SoR, platforms have to indicate the type of content that was moderated (i. e., text, images, videos, etc.). Studying the platform-specific distribution of how the platforms specify the type of the moderated content, we find considerable differences (see Fig. 6.1). While TikTok primarily focuses on **Text** (57.14%) and **Videos** (33.97%), Snapchat moderated a larger share of **Videos** (63.92%) and **Images** (16.96%). The platform X seems to limit its content moderation activities mainly to **Synthetic Media** (99.83%). The remaining platforms each categorized more than 50% of all content moderation actions as content type **Other**, i. e., content types that are not predefined by the DSA.[1] Altogether, we find that the type of content subject to moderation is, in many cases, strongly related to the content that is predominantly published on the respective platform (e. g., Video and Image on Snapchat, Video and Advertisement on YouTube, Pins on Pinterest, job-related content on LinkedIn). It is also worth noting that the content type is specified by the platform (i. e., self-reported), and each content piece is assigned to a single type. However, social media content is often a mix of different types. In a similar vein, AI-generated content may be classified as image/text/video or as synthetic media. Overall, it seems likely that the content type reported by platforms often describes only one dimension of a social media post.

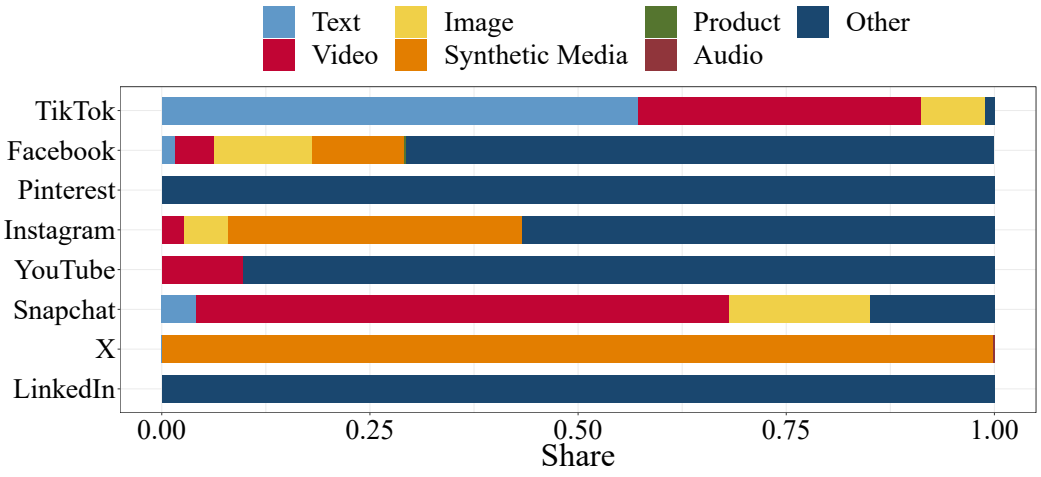


Figure 6.1: Distribution of moderated *content types*.

6.4.2 Reasons for Moderation (RQ2)

Platforms must specify if content is incompatible (Article 17(3)(d) DSA) or illegal (Article 17 (3)(e) DSA) when explaining moderation actions. Incompatibility dominates (99.80%), while

[1]The majority of this content concerns violations that affect entire accounts/profiles, platform-specific content types (e. g., pins or boards on Pinterest), and (job) advertisements (in particular on Youtube, Pinterest, and LinkedIn).

only X focusing more than 99% on illegal content. Additionally, platforms assign one of 14 categories defining the reason for moderation. We find that there is a relatively highly similarity across most platforms (see Fig. 6.2). With the exceptions of X and Pinterest, all platforms moderated a large proportion of content that does not correspond to the **Scope of Platform Service**, **Illegal/Harmful Speech**, and **Violence**. In contrast, Pinterest focused primarily on **Pornography/Sexualized Content** (80.95%). In the case of X, the vast majority of content moderation actions were attributed to **Violence** (41.18%) or **Pornography/Sexualized Content** (44.37%). Snapchat is the platform that moderated content in the widest variety of categories (e. g. 16.69% in **Scams & Fraud** and 10.77% **Unsafe & Illegal Products**).

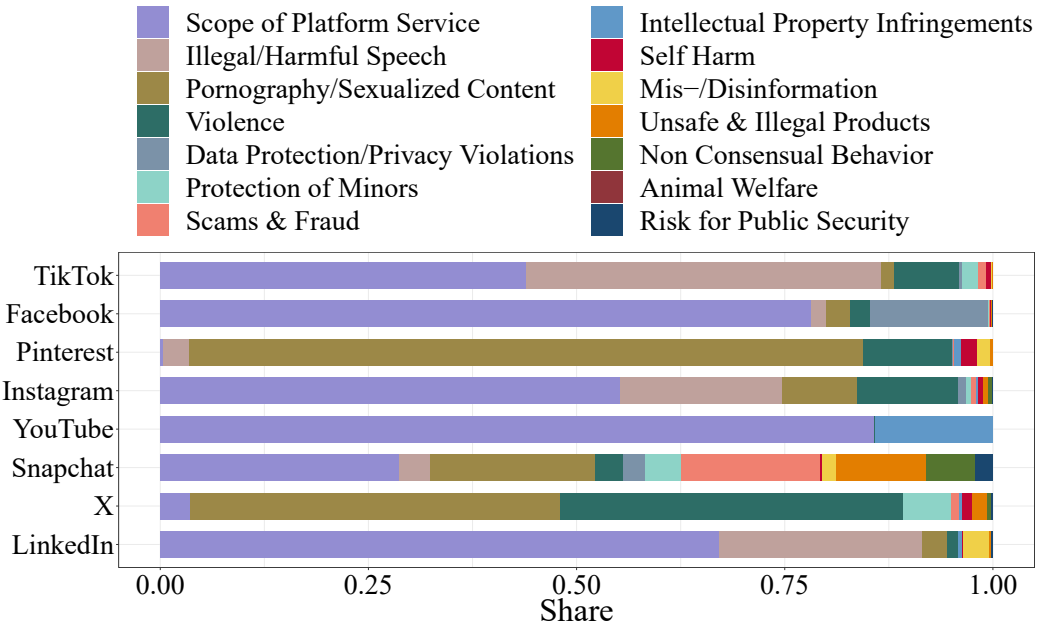


Figure 6.2: Distribution of *reasons for moderation*.

6.4.3 Automation of Content Moderation (RQ3)

Regarding content moderation, platforms must specify if the decision was automated (fully, partially, not), and if the violation was detected automatically (yes, no). Overall, 60.67% of all content moderation decisions were performed fully automated, 31.48% partially automated, and 7.74% not automated. Furthermore, automated detection often led to automated decisions, while manually moderated content was more frequently identified or reported by humans (e.g., users or moderators). Across all fully automated content moderation decisions, over 99% of the content was automatically identified; whereas in the partially automated and not automated categories, this proportion drops to 76.06% and 71.75%, respectively.

Across platforms, we observe further differences (see Fig. 6.3). While TikTok mainly employs **fully automated** decision-making, Facebook, Pinterest and Instagramm tend to combine automated means and human intervention (i. e., **partially automated**). In contrast, YouTube, Snapchat, X and LinkedIn performed the majority of their content moderation **not automated**. X again stands out with a completely manual content moderation (i. e., not automated).

Regression analysis: To better understand situations in which content moderation is more likely to be performed automatically, we implement a logistic regression model estimating

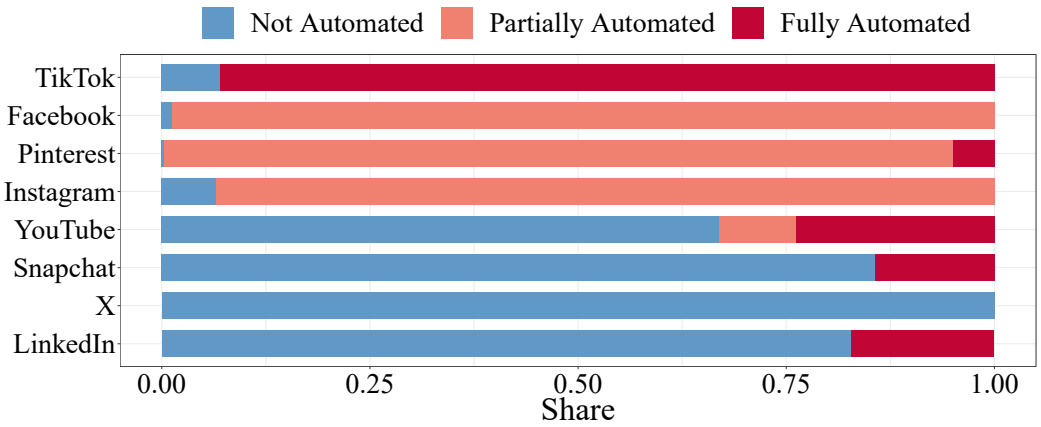


Figure 6.3: Distribution of *automation level* across platforms.

how the likelihood of automated content moderation decision varies across different content types and decision grounds (see Fig. 6.4). We find that the odds of a content moderation decision being performed automatically are $e^{3.470} \approx 32.143$ times higher if the content was detected automatically ($p < 0.01$). Across content types, we find that the odds of a content moderation decision performed automatically are significantly higher for text compared to other content (i. e., *Product*, *Synthetic Media*, *Audio*, *Other*) ($p < 0.01$). At the same time, we find a smaller positive association for *Images* ($p < 0.01$), while *Videos* are less likely to be moderated automatically ($p < 0.01$). With regard to legal grounds, we find that all categories are statistically significantly less likely to be automatically moderated compared to content that does correspond to the *Scope of Platform Service* (each $p < 0.01$). We also observe a small negative association between the time elapsed between the date the content was published on the platform and the moderation date ($p < 0.01$).

6.4.4 Content Moderation Actions (RQ4)

Reported content moderation actions can be grouped into two categories: (i) actions that affect a specific content piece (i. e., **removal**, **labelling**, **disabling**, **demotion** and **age restriction** of content); (ii) actions that affect an account (i. e., **suspension** or **termination** of an account). The platforms describe in the SoRs how the content was moderated, either by selecting one of these predefined actions or by selecting **Other**. In the case of the latter, platforms can provide a short description of the action that was taken. To accommodate such cases, we employed string matching to assign the text descriptions to the predefined action categories.[2] If a description is missing or unclear, the category **Other** is used (0.36%).

Overall, we observe that the most frequent types of content moderation are the **removal of content** (55.15%), the **demotion of content** (25.15%), and the **suspension/termination** (14.96%). The tendency to focus on these actions as the primary choice of content moderation is prevalent across most platforms. However, there are also differences. For instance, Pinterest

[2]The descriptions were assigned to the predefined action categories as follows: "Limited distribution", "not eligible for recommendation", "Bounce" → **Content Demoted**; "Bounce", "Ban" → **Content Disabled** →, "AddTweetAnnotation" → **Content Labelled** (ordered by descending frequency). X describes a large part of the moderated content as "not suitable for work" (NSFW) (71.20%) whereby, according to the platform's own guidelines, either a reduction in visibility (i. e., **demotion**) or **removal** of the content is implemented (see striped area in Fig. 6.5)

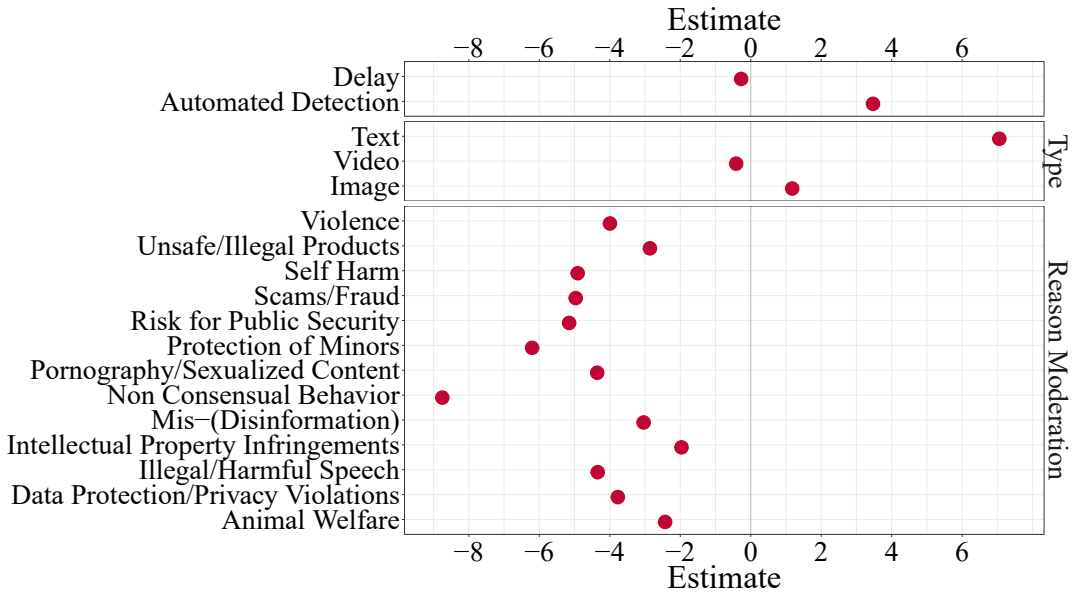


Figure 6.4: Coefficient estimates (circles) and 99% confidence intervals (bars) for a logistic regression model with $\boldsymbol{Automated\ Decision}$ ($=\mathbf{1}$ if yes; otherwise $=\mathbf{0}$) as DV ($\boldsymbol{n}=$ 156 378 199). Monthly and platform fixed effects are included. EVs are $\boldsymbol{AutomatedDetection}$ (*ref. No*), $\boldsymbol{ContentType}$ (*ref.: Other*), $\boldsymbol{Reason\ Moderation}$ (*ref. Scope of Platform Service*), and $\boldsymbol{Delay}$ (days elapsed until moderation).

puts a stronger focus on **content demotion**. Conversely, Snapchat implemented **content disabling** in over 50% of its content moderation actions. X does not distinguish between **removal** and **demotion** in its reporting to the DSA-TDB.

6.5 Discussion & Future Work

Relevance: Effective content moderation is vital for preventing the dissemination of illegal and undesirable content in social networks (Gillespie, 2018). However, social media providers' moderation decisions are often perceived as opaque. Previous research, primarily based on artificial environments or limited observational datasets (e. g., Drolsbach & Pröllochs, 2023; Ling et al., 2023), faced challenges in systematically collecting and analyzing these decisions. To foster transparency and accountability, the DSA mandates major social media platforms in the EU to make key information on their content moderation decisions publicly accessible. Utilizing the EU's DSA-TDB – a unique and previously unavailable data source – we shed light on how major social media providers moderate user-generated content on their platforms.

Implications: Our study reveals significant differences and inconsistencies in the content moderation practices of major social media platforms, with TikTok performing over 350 times more content moderation decisions per user than X. While platforms with higher moderation volumes may or may not encounter rule-breaking content more frequently, our findings indicate that different platforms focus on distinct types of rule-breaking content. Although all platforms consistently address violence and pornography, their approach to other harms like illegal speech and misinformation varies considerably. For instance, X rarely moderates illegal speech and misinformation. Additionally, the actions taken against rule-breaking content differ across platforms, with some removing it frequently and others opting to reduce visibility.

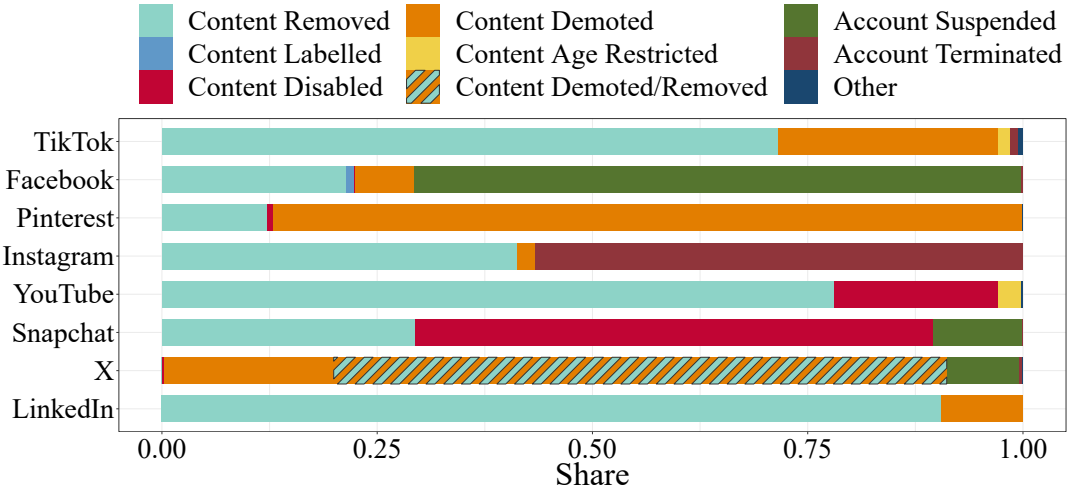


Figure 6.5: Distribution of *decision types* across platforms.

Overall, our study suggests varied interpretations of DSA obligations, leading to a fragmented outcome that the DSA aims to avoid. These findings have implications for regulators, urging clearer guidelines to ensure consistent standards for handling rule-breaking content.

Further, our study contributes to the ongoing debate on the role of automation in content moderation. We find that the majority of rule-breaking content is already identified and processed through automated methods. While some platforms combine automated and human moderation, others (e. g., TikTok), predominantly rely on fully automated content moderation. X stands out by exclusively relying on non-automated means, albeit moderating a significantly lower content volume. Given the massive scale of content moderation in the EU, this suggests a potential need for some level of automation to meet the DSA-imposed regulatory obligations.

Limitations and future work: Our study opens avenues for future research. Firstly, our inferences are restricted to the initial two months after the introduction of the DSA-TDB in the EU, and future work can explore how patterns change over time. Secondly, the evolving nature of content moderation efforts may reach a different steady-state with growing experience and EU rule clarifications. Thirdly, understanding the impact of content removal or visibility reduction on user behavior warrants further investigation. Lastly, analyzing the source content that has been moderated would be insightful, but such data is currently unavailable. Despite these limitations, our study contributes to the understanding of social media content moderation, providing a foundation for future policy improvements. We hope our early work stimulates more research to enhance transparency in content moderation on social media.

Bibliography

Bär, D., Pröllochs, N., & Feuerriegel, S. (2023). New threats to society from free-speech social media platforms. *Communications of the ACM*, *66*(10), 37–40.

Drolsbach, C. P., & Pröllochs, N. (2023). Diffusion of community fact-checked misinformation on Twitter. *CSCW*.

Feuerriegel, S., DiResta, R., Goldstein, J. A., Kumar, S., Lorenz-Spreen, P., Tomz, M., & Pröllochs, N. (2023). Research can help to tackle ai-generated disinformation. *Nature Human Behaviour*, *7*, 1818–1821.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology*, *17*(1).

Horta Ribeiro, M., Cheng, J., & West, R. (2023). Automated content moderation increases adherence to community guidelines. *WWW*.

Jhaver, S., Birman, I., Gilbert, E., & Bruckman, A. (2019). Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction*, *26*(5), 1–35.

Jiang, J. A., Nie, P., Brubaker, J. R., & Fiesler, C. (2023). A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction*, *30*(1).

Ling, C., Gummadi, K. P., & Zannettou, S. (2023). ”learn the facts about Covid-19”: Analyzing the use of warning labels on TikTok videos. *ICWSM*.

Schlag, G. (2023). European union’s regulating of social media: A discourse analysis of the digital services act. *Politics and Governance*, *11*(3), 168–177.

Turillazzi, A., Taddeo, M., Floridi, L., & Casolari, F. (2023). The digital services act: An analysis of its ethical, legal, and social implications. *Law, Innovation and Technology*, *15*(1), 83–106.

Chapter 7

Pass-through of Temporary Fuel Tax Reductions: Evidence from Europe

Abstract

Several European countries have implemented temporary fuel tax reductions in 2022 to relieve the financial burden on their citizens. This paper is the first to provide estimates of the pass-through rates as well as the effect on retail margins for France and Italy. Further, it contributes to the recent literature on the fuel tax reduction in Germany. Using a unique data set containing daily consumer prices at service station chain level for gasoline and diesel, we employ a staggered Difference-in-Differences (DiD) design. Our main results imply that in the aggregate there was a full-shifting of the fuel tax reductions in all three countries. Nevertheless, in an event study design we find that the pass-through rates over time are heterogeneous between the countries and types of fuel. Depending on time, heterogeneous effects imply a full-shifting up to a minor over-shifting of the pass-through rates. These findings also have important implications for the effective design of unconventional fiscal policy as well as for competition policy in the fuel market.

Keywords: Fuel prices, Pass-through, environmental taxes, staggered DiD

7.1 Introduction

More than two years after the beginning of the COVID-19 pandemic many countries worldwide exhibited very high inflation rates. The reasons are a recovering demand in combination with ongoing supply chain problems as well as the war of aggression in the Ukraine. In April 2022, the inflation rate of Germany reached 7.4%, the highest rate since 1981. Other western countries have similar rates: the inflation rate of the whole EU has been 8.1%, the US even had a rate of 8.3% in April 2022. In autumn 2022, inflation in some countries has already risen to around 10%.[1]

In this situation several governments tried to relieve their citizens with tax reductions or transfer payments. On April 27, 2022, the German government announced a (second) stimulus package worth 14-16 billion Euro.[2] Beside new transfer payments and a cheap, nationwide public transport ticket (*€9 ticket*), it also included a temporary reduction of the energy tax rate from June 1 to August 31, 2022 at an estimated cost of 3.15 billion Euro.[3] Since the energy tax is levied on fuel products in Germany, this might also have had an effect on retail fuel prices. However, consumers only benefit from this regulation if the petroleum companies pass-through the tax reduction sufficiently.

Also other countries in the EU, such as France or Italy have implemented temporary measures in 2022. In France, the government introduced a fixed fuel discount between April 1 and August 31, which later has been extended until the end of the year 2022. The Italian government had already approved a subsidy program in March including a fuel tax reduction from March 22 until the end of April. Also this intervention has later been extended until December 31, 2022. Those government actions provide us with ideal exogenous shocks, which we can use as a natural experiment.[4]

In this paper, we estimate the pass-through rate and the effect on the retail margins of the temporary fuel tax reductions in the three largest countries of continental Europe (measured by GDP), namely France, Germany and Italy. Austria, Estonia, Lithuania, and Latvia are being selected as appropriate control countries for the purpose of this analysis because these nations did not introduce any comparable measures during the year 2022. Using a unique panel data set containing daily consumer prices for gasoline and diesel on service station chain level, we compute the pass-through rates and changes in the margins by employing a staggered Difference-in-Differences (DiD) approach.

Our results imply a heterogeneous passing over time of the fuel tax reductions depending on the country as well as on the type of fuel. However, we find the following two key results. First, the average pass-through rates are very high so that there is a full-shifting of the temporary tax reductions, indicating highly competitive markets. Second, the estimated pass-through

[1] See https://www.global-rates.com/de/wirtschaftsstatistiken/inflation/inflation.aspx. (Last accessed: October 19, 2022).

[2] See https://www.bundesfinanzministerium.de/Content/DE/Pressemitteilungen/Finanzpolitik/2022/04/2022-04-27-zweites-entlastungspaket.html. (Last accessed: October 19, 2022). The package was approved by the German parliament on May 13.

[3] See https://www.bundestag.de/dokumente/textarchiv/2022/kw20-de-energiesteuersenkungsgesetz-894664. (Last accessed: October 19, 2022).

[4] Due to the circumstance that the interventions were introduced to ease the burden to consumers and with the intention to lower consumer prices, potential endogeneity issues will be discussed at the end of Chapter 7.5.

rates are on average higher for gasoline than for diesel,[5] which may result from the special situation on the European energy markets in 2022 following the Russian invasion of Ukraine and a relating high demand of (heating) diesel.

The results of our paper have implications for the effective design of unconventional fiscal policy and are also relevant for competition policy. Unconventional fiscal policy can only be effective in stimulating demand if consumers expect tax reductions to be passed through by firms. Besides, such fuel tax reductions also have distributional- and climate-economical effects. While the discount may act like a redistribution from bottom to top as particularly high-income consumers with large cars are benefiting, it is generally questionable whether subsidizing fossil fuels is a good idea in times of climate change.

The rest of the paper is structured as follows. Section 7.2 presents related literature, followed by a description of retail fuel markets in Section 7.3. We present our data set and descriptive statistics in Section 7.4. In Section 7.5, we explain our empirical strategy and then present the estimation results in Section 7.6. We conclude in Section 7.7 by discussing policy implications and limitations.

7.2 Related Literature

Since gasoline markets are typically characterized by a very specific cyclical pricing pattern, academia as well as competition authorities are highly interested in analyzing this industry sector. The leading theory to explain price cycles in gasoline markets are Edgeworth price cycles. This theory has been formalized by Maskin and Tirole (1988) and assumes a dynamic oligopoly game where firms compete in prices and sell homogeneous goods. Starting at a supra-competitive price, firms undercut each other until the price reaches marginal costs. Given that there is no gain to lowering prices further, firms play a war of attrition. After one firm relents the price back to a high level, the other follow and the cycle begins anew (see M. D. Noel et al., 2011).

In contrast to the literature mentioned above, other authors discuss the possibility of tacit collusion in gasoline markets. Since petrol stations can easily observe and monitor price changes as well as learn the price setting behavior of their competitors, an explicit agreement is not necessary to establish such an behavior. Evidence for collusion in gasoline markets has been found for Australia (Byrne & De Roos, 2019) and Norway (Foros & Steen, 2013). With respect to Germany, Dewenter et al. (2017) show that the introduction of the 'Markttransparenzstelle für Kraftstoffe' (market transparency unit for fuels, MTS-K)[6] in 2013 has increased both gasoline and diesel prices. Assad et al. (2023) find that algorithmic pricing has a significant effect on competition in the German gasoline market.

Another strand of the literature analyzes the effects of changes in the crude oil price on refined petroleum products. Here, most of the papers are focused on the oil-gasoline relationship. It has been shown that downstream prices seem to respond to increases in upstream prices more rapidly than their responses to decreases in upstream prices, so that there is a potentially

[5] Note: Our hypothesis tests indicate that there is no statistically significant difference between them (see Chapter 7.6.1).

[6] The MTS-K is an independent unit of the German competition authority. All petrol stations in Germany are legally bound to inform the MTS-K about price changes in real time (see https://www.bundeskartellamt.de/EN/Economicsectors/MineralOil/MTU-Fuels/mtufuels_node.html;jsessionid=0E947D4936B3B12872C630A4005CED95.2_cid378).

asymmetric pass-through of increasing and decreasing costs ('rockets and feathers') (e.g., Grasso & Manera, 2007; M. Noel, 2009; M. D. Noel, 2015). In this context, similar studies explore the causes for this asymmetric relation between crude oil and gasoline. They identify refinery utilization rates and inventories as a main driver of those asymmetries (e.g., Kaufmann & Laskowski, 2005; Perdiguero-Garcia, 2013)

Recent papers also analyze the pass-through of taxes and excise duties on fuel prices. In general, pass-through rates depend on consumer behavior as well as on competition parameters (e.g., Montag et al., 2021; Genakos & Pagliero, 2022; Harju et al., 2022). The effect of tax changes on market prices primarily depends on supply and demand elasticities (Edgeworth, 1897). In a perfectly competitive market, the pass-through rate increases in the elasticity of supply and decreases in the elasticity of demand. However, if competition is not perfectly competitive, the pattern of tax incidence becomes more complex and several degrees of tax shifting are possible: under-, full- and over-shifting to consumers (see 7.A). Besides, not only the horizontal market structure but also vertical market power has to be considered (Fuest et al., 2020).

Some empirical results indicate that the coefficient associated with taxes on gasoline prices is not statistically different from one (or slightly less than one) (e.g., Marion & Muehlegger, 2011; Bello & Contín-Pilart, 2012; Li et al., 2014). In contrast, other studies find that a higher percentage of a tax increase is passed to consumers than a tax reduction (Doyle Jr & Samphantharak, 2008; Silvia & Taylor, 2014) or identify state-specific rates of pass-through (Kaufmann, 2019). Regarding the fuel tax reduction in Germany in 2022, results range between a partial pass-through to a full-pass-through (e.g., Bernhardt et al., 2023; Dovern et al., 2023; Fuest et al., 2022; Kahl, 2023; Schmerer & Hansen, 2023; Seiler & Stöckmann, 2023). Here we explicitly contribute for the case of Germany but also other not yet examined countries (France and Italy) utilizing a staggered difference-in-difference design.

7.3 The Retail Fuel Market

The fuel market is characterized by a vertical structure, with refineries producing fuels from crude oil in the upstream market and selling them to fuel stations, which in turn distribute the fuels to end customers (downstream market). In our study, we focus on the analysis of retail prices on the service station chain level, however, an understanding of the upstream sector is still relevant, especially for the calculation of margins. During fuel production a barrel (42 gallons) of crude oil can be refined into 19 gallons of gasoline, 12 gallons of diesel and 13 gallons of other products.[7] In addition to crude oil, refineries also add other oils and liquids to the finished products that are sold to the petrol stations.

After significant increases in the European retail fuel prices at the beginning of 2022, several countries adopted measures with the aim of relieving consumers. For our analysis, we focus on the three largest economies in continental Europe that have introduced reductions of excise duties on fuel or similar measures, explicitly Germany, France, and Italy. To choose appropriate control countries for our staggered DiD approach, we need to find countries of the European Union (EU) which have not implemented any regulations in the fuel market in 2022. Table 7.1 presents an overview of policies introduced in all member states of the EU. It is obvious,

[7] See https://www.eia.gov/energyexplained/oil-and-petroleum-products/refining-crude-oil-inputs-and-outputs.php (Last accessed: October 19, 2022).

that there are numerous overlaps in timing (i.e., measures came into force on the same day), which prevent a comparison. Apart from that, there are several countries that have chosen VAT reductions or price caps as policy measures, which also reduces comparability (due to varying magnitude of actual discounts over time). The consideration of all countries shows that by these criteria the majority of all EU countries are not eligible as control countries for our analysis.[8] Yet, Austria, Estonia, Latvia, and Lithuania, as countries that have not introduced any regulations, are considered suitable for comparison.

Country	Country Code	Type of Measure	Date (mm/dd/yy)	Tax reduction		in Sample?
				E5	Diesel	
Austria	AT	–	–	–	–	Control
Belgium	BE	VAT reduction + Fuel tax reduction	02/01 + 03/19/22	15% + 17.5ct/l	15% + 17.5ct/l	–
Bulgaria	BG	Fixed discount	07/09/22	13ct/l	13ct/l	–
Croatia	HR	Price cap	10/17/21			–
Cyprus	CY	–	–	–	–	–
Czech Republic	CZ	Fuel tax reduction	06/01/22	1.5CZK/l	1.5CZK/l	–
Denmark	DK	–	–	–	–	–
Estonia	EE	–	–	–	–	Control
Finland	FI	–	–	–	–	–
France	FR	Fixed discount	04/01/22	15ct/l	15ct/l	Treatment
Germany	DE	Fuel tax reduction	06/01/22	29.55ct/l	14.04ct/l	Treatment
Greece	GR	–	–	–	–	–
Hungary	HU	Price cap	11/11/21			–
Ireland	IE	Fuel tax reduction	03/10/22	20ct/l	15ct/l	–
Italy	IT	Fuel tax reduction	03/22/22	25ct/l	25ct/l	Treatment
Latvia	LV	–	–	–	–	Control
Lithuania	LT	–	–	–	–	Control
Luxembourg	LU	Fuel tax reduction	03/31/22	7.5ct/l	7.5ct/l	-
Malta	MT	–	–	–	–	-
Netherlands	NL	Fuel tax reduction	04/01/22	17.3ct/l	11.1ct/l	–
Poland	PL	VAT reduction	02/01/22	15%	15%	–
Portugal	PT	”Autovoucher” (limited to 50l/month)	11/01/21	10ct/l	10ct/l	–
Romania	RO	–	–	–	–	–
Slovenia	SI	Price cap	03/15/22	–	–	–
Slovakia	SK	–	–	–	–	–
Spain	ES	Fixed discount	04/01/22	20ct/l	20ct/l	–
Sweden	SE	Fuel tax reduction	06/01/22	17ct/l	17ct/l	–

Table 7.1: Overview of fuel tax reductions in all EU member states. In the case of fuel tax reductions given values are excl. associated VAT reduction. Sources: https://www.bruegel.org/dataset/national-policies-shield-consumers-rising-energy-prices (last accessed on 07/08/2023).

The retail fuel markets in all countries of our sample are characterized by an oligopoly. Those oligopolists operate nationwide, while there are also smaller suppliers with a single or small number of service stations that operate on a regional basis. For instance, in Germany five firms (Shell, BP/Aral, Esso, Total, and Jet) combine for a market share of 67%. In the other countries, the market shares of the oligopolists are within a comparable range (see Table 7.2). Differences in the total number of service stations are primarily related to country size and population.

In contrast, the upstream markets in the individual countries of our sample have larger differences. In Austria, for example, there is only one refinery, and the majority of fuel is imported

[8] Belgium, Croatia, Hungary, Poland, Portugal, and Slovenia have introduced regulations other than a fixed tax reduction/discount. Bulgaria, Czech Republic, Ireland, the Netherlands, Spain, and Sweden were excluded as additional treated countries for timing reasons. Due to their specific geographical location, data unavailability and/or a currency other than the Euro we decided not to consider Cyprus, Denmark, Finland, Greece, Malta, Romania, and Slovakia.

from Germany. France also has a relatively small number of refineries and refining capacity in relation to the market size, resulting in a more inelastic supply side compared to Germany and Italy. Estonia, Latvia, and Lithuania do not have any (or only one) refinery and are therefore also strongly dependent on fuel imports. However, we incorporate these observable differences between countries by including the refinery utilization, imports of crude oil and petroleum products, and the number of gas stations per chain as control variables into our empirical analysis (see Sections 7.4 and 7.5).

		Austria	Estonia	France	Germany	Italy	Latvia	Lithuania
Downstream	Number of fuel stations	2,759	515	11,040	14,452	21,700	600	765
	Oligopoly members	BP, ENI, Jet, OMV, Shell	Alexela Oil, Neste, Circle K, Olerex, Saare Kütus	Shell, Aral, Esso, Total, Jet	Shell, Aral, Esso, Total, Jet	Eni, Q8, Esso, Tamoil	Circle K, Neste, Viada, Virsi-A	Viada, Circle K, Neste, Baltic Petroleum
	Market share of oligopolists	67%	≈54%	62%	67%	49%	≈52%	≈51%
Upstream	Number of refineries	1	0	6	11	10	0	1
	Refinery capacity (in Mt/a)	9.80	0	58.20	100.90	83.90	0	9.60

Table 7.2: Overview of relevant market characteristics in all countries considered. When market share values were not publicly available, they were approximated based on the stations in our dataset relative to all stations (denoted by ≈). **Source:** Statistical Report 2023, FuelsEurope, https://www.fuelseurope.eu/publications/publications/statistical-report-2023.

Considering retail fuel prices, it becomes clear that the price of crude oil accounts for an important share of prices and their fluctuations. Yet, taxes and other duties account for the largest share. Table 7.3 summarizes the excise duties on gasoline and diesel for the countries in our data set. All countries levy a lower excise duty on diesel than on gasoline, with Germany having the largest diesel privilege (at least without taking into account the temporary fuel tax reductions). Without considering any temporary tax reductions, Austria has the lowest excise duties for fuel and Italy has the highest ones. All countries also levy an additional Value-added tax (VAT) on gasoline and diesel.[9] In Germany, an additional fuel carbon tax of 7.2 cents (8.03 cents) on gasoline (diesel) and an additional fuel storage fee of 0.27 cents (0.30 cents) on gasoline (diesel) are levied.

In Germany, the excise duty on fuel ("energy tax") has been lowered by 29.55 cents per liter for gasoline (35.20ct incl. VAT) and by 14.04 cents per liter for diesel (16.70ct incl. VAT) for the period between June 1 and August 31, 2022.[10] With this reduction, Germany has lowered the excise duty on fuel to the minimum level permitted in the EU. In Italy, a reduction of the excise duty on gasoline and diesel by 25 cents per liter (30.50ct incl. VAT) has been introduced from March 22, 2022 on.[11] This measure was initially limited until April 30, but was extended shortly after and ultimately lasted until the end of 2022. The French government has passed a law that introduced a discount for all important fuel products by 15 cents per liter (18.00ct incl. VAT) from April 1, 2022 on.On September 1, the fuel discount has even been increased

[9] VAT rates are as follows: 19% in Germany, 20% in Austria, Estonia, and France, 21% in Latvia and Lithuania, and 22% in Italy. To calculate margins and pass-through rates we include VAT reductions associated with the energy tax reductions/discounts to consider the overall reductions.

[10] See: https://www.bundestag.de/dokumente/textarchiv/2022/kw20-de-energiesteuersenkungsgesetz-894664 (Last accessed: July 10, 2023).

[11] See: https://www.gazzettaufficiale.it/eli/id/2022/03/21/22G00032/sg and https://www.loc.gov/item/global-legal-monitor/2022-05-31/italy-new-law-reduces-excise-taxes-and-vat-on-fuels-to-ameliorate-financial-crisis-caused-by-war-in-ukraine/ (Last accessed: July 10, 2023).

from 15 to 25 cents per liter and in addition has been extended until December 31, 2022.[12] This discount was paid as a subsidy for quantities sold to the distributor at the second-last distribution level. Based on the termination of the tax reduction in Germany on August 31, 2022 and the simultaneous change of the discount in France, we have chosen an observation period until August 31, 2022.

Even though technically the introduced discount in France is different compared to the tax reductions in Germany and Italy, basically it has a similar effect on the costs of the retailers (i.e., service stations). For this reason, it is referred to as a reduction of excise duties paid on the retail-level in the following. In our empirical analysis (see Section 7.5 and 7.6), we compare our estimated coefficients with the overall tax reductions (also including the associated VAT changes), which are also given in Table 7.3. With regard to the implemented measures, it is important to note that these represented a one-time reduction in all treated countries. In Austria, Estonia, Latvia, and Lithuania, the tax rate remained constant throughout the whole observation period (see Figure 7.8 in 7.C).

		Gasoline (E5)				Diesel			
Country	**Treatment**	**Pre**	**Post**	**Diff.**	**Incl. VAT**	**Pre**	**Post**	**Diff.**	**Incl. VAT**
Austria	–	48.00	48.00	–	–	40.00	40.00	–	–
Estonia	–	42.28	42.28	–	–	39.29	39.29	–	–
France	04/01/2022	68.29	53.29	-15.00	-18.00	59.40	44.40	-15.00	-18.00
Germany	06/01/2022	65.45	35.90	-29.55	-35.20	47.04	33.00	-14.04	-16.70
Italy	03/22/2022	72.84	47.84	-25.00	-30.50	61.74	36.74	-25.00	-30.50
Latvia	–	41.12	41.12	–	–	33.30	33.30	–	–
Lithuania	–	43.44	43.44	–	–	33.02	33.02	–	–

Table 7.3: Excise taxes on Gasoline (E5) and Diesel in cents per liter before and after treatment (where applicable). "Incl. VAT" indicates estimated changes when VAT is applied. Source: https://ec.europa.eu/taxation_customs/tedb/

7.4 Data and Descriptive Statistics

7.4.1 Data Collection

Our analysis is based on five different data sources. First, we scraped data on daily average gasoline (E5) and diesel consumer prices on a service station chain level from the information platform *Fuelo*. These prices on *Fuelo* are the basis of our analysis. *Fuelo* uses official sources as well as information from consumers, publishes this on its website and displays historical information on a daily average level. Real-time price updates are not considered relevant for our analysis as the platform only provides the historical price averages at the service station chain level.[13] The data from *Fuelo* also provides information on the number of fuel stations per service chain. Incorporating this measure serves a dual purpose. Firstly, it helps control for

[12] See https://www.connexionfrance.com/article/French-news/How-the-French-government-fuel-discount-will-change-from-September-1 (Last accessed: July 12, 2023)

[13] Example for German prices from February 2, 2022, https://de.fuelo.net/prices/date/2022-2-2?lang=en. (Last accessed: July 11, 2023). Statement from Fuelo on their sources: https://de.fuelo.net/prices/last_updated?lang=de. (Last accessed: July 11, 2023).

variations in the number of stations across different countries. Secondly, it takes into account the fact that the average fuel prices displayed on the Fuelo website are constructed based on different numbers of chain stations in each country.[14]

Second, we use data on the crude oil price Brent from *Onvista* and exchange rates from Dollar into Euro by the Federal Reserve Bank of St. Louis (FRED) to highlight the relation between consumer prices and the Brent price.[15] The Brent price is also a crucial part to determine retail margins.

Third, we incorporate data on refinery capacities and convert them into a measure of refinery utilization, which indicates how efficiently the refineries are utilizing their maximum capacity. It is crucial to control for refinery utilization in our analysis, since gasoline and diesel can either be produced domestically within the country or imported from other countries. To assess this, we utilize data from both *Concawe* and *Eurostat*.[16] *Concawe* provides annual national-level data on refinery capacities, measured in mega tonnes per annum (Mt/a). With the assistance of *Eurostat* data on the supply (and transformation) of oil and petroleum products, we convert these capacities into a measure of refinery utilization.

Exact calculation of the utilization rate needs some clarification. Before the utilization rate can be determined the domestic production must be calculated from several variables, i.e. the stream of raw oil, loss from refining the crude oil, changes in stock, releases of strategic reserve or inflow from marine bunkers.[17]

The capacities provided by *Concawe* and *Eurostat* are available on a yearly basis, while the data for refinery utilization is required on a monthly level. To bridge this gap, the yearly capacities are converted into monthly capacities by dividing them by 12. Dividing the monthly domestic production by monthly available capacity determines the utilization of the refinery on a monthly level. Controlling for these refinery utilization possesses the opportunity to rule out differences at the supply side from the local refinery level, i.e. from breakdown in the refinery or loss of access to crude oil.

Fourth, to account for variations in the total imports of oil and petroleum products, we incorporate national-level data from *Eurostat* specifically related to the imports of these products. These imports are measured in thousands of tonnes. By controlling for changes

[14] See, 7.B, Table 7.9 for the distribution of number of service chain stations in the data. Note: Data from Fuelo has a large market coverage. Example: Germany had 14,452 stations in 2022. Our scraped data covers 10,600 stations.

[15] See historical Brent prices, https://www.onvista.de/rohstoffe/db-Oelpreis-Brent-26262975 and exchange rates from FRED, https://fred.stlouisfed.org/series/DEXUSEU. (Last accessed: July 11, 2023).

[16] See information from Concawe, https://www.concawe.eu/refineries-map/ and Eurostat https://ec.europa.eu/eurostat/databrowser/view/nrg_cb_oilm/default/table?lang=en. (Last accessed: July 11, 2023).

[17] *Eurostat* provides information from the Monthly Oil and Gas questionnaire at page 10, No. 11 on how the gross inland deliveries are determined and we will use this to rearrange this equation for domestic production (https://ec.europa.eu/eurostat/documents/38154/42198/MOS_v2012.1.pdf/f4a7a75c-b0d1-4370-802a-560ca5f86f4d#:~:text=Gross%20inland%20deliveries%20(Observed)%3A,.%20to%20the%20inland%20market. (Last accessed: July 11, 2023)). From the data of *Eurostat* we calculate domestic production for petroleum products with the formula: Domestic Production = Gross inland deliveries - Primary product receipts - Recycled products + Refinery fuel - Imports + Exports + International marine bunkers - Interproduct transfers + Products transferred + Stock changes. Note: Refinery gross output denotes what we call domestic production. Statistics from *Eurostat* regarding what they refer to as *indigenous production* is not available.

in imports, we aim to capture another aspect of supply-side changes that are likely to be significantly influenced by the outbreak and ongoing war in Ukraine.[18]

7.4.2 Descriptive Statistics

Our final data set includes consumer price data in Euro per liter for seven European countries Germany, France, Italy, Austria, Lithuania, Estonia and Latvia on a service station chain level during the period from January 3 to August 31, 2022.

Table 7.8 in 7.C presents the summary statistics divided by countries. To calculate the margins, we simply subtract taxes and duties as well as the share of the crude oil price (Brent price) attributable to the production of diesel and gasoline from the gross consumer prices.[19] Even though these margins still contain different cost types (e.g., cost of refining, transportation costs), with the crude oil price we can eliminate the main source of input cost variation.

The data set contains 12,515 observations on a service station chain level, i.e. we have a panel data set including price information on 52 unique country-service-station-chain-pairs for 241 days.[20] For each country we observe a different number of chains present in the data (see Table 7.9 in 7.C).[21] For instance, Germany has 15 different service station chains present in the data, whereas Austria has six. Overall, there are 30 chains in the treated country's data and 22 chains in the non-treated data.[22] Countries display variations in terms of refinery utilization, the number of stations per chain, and total imports of oil and petroleum products. To illustrate this point, a comparison of Germany and Austria serves as an examples. When comparing these two countries, we observe differences in the magnitude of imports of oil and petroleum products (mean: 10,191 v. 987; measured in thousand tonnes), refinery utilization (mean: 0.91 v. 0.55; represented by decimal units) and the number of fuel stations operated per chain (mean: 706 vs. 196). By incorporating these covariates into our analysis, we aim to account for and capture differences in the pre-existing trends and characteristics across countries. These factors help us consider the unique features and dynamics of each country's fuel market.[23]

[18] See: *Eurostat*, imports of oil and petroleum products by partner country, https://ec.europa.eu/eurostat/databrowser/view/NRG_TI_OILM__custom_6837161/default/table?lang=en. (Last accessed: July 11, 2023).

[19] An important note is that our measure of retail margins includes the refinery margin, the station margin, as well as different cost types such as the cost of refining or the cost of transportation. For a detailed description on the calculation of margins see 7.B.

[20] The data set is slightly unbalanced.

[21] Fuelo's market coverage per service station chain varies and total market coverage is different across countries. However, it is worth mentioning that the geographic coverage within countries comprises almost their entirety, which can be substantiated through visual inspection. Nevertheless, the goal of this study is to analyze the overall pass-through rate. In this respect, our identification strategy relies on the comparison of the evolution of country-wide large chains average prices, such as most important players Shell, Esso, or Total rather than analyzing the entire market.

[22] The estimation will utilize the not-yet-treated characteristics of the data. During the time periods when part of the data is not yet treated these chains will be used as a comparison group.

[23] Estonia and Latvia do not have a refinery. For regression purpose the values are set to zero. For E5 in Lithuania (Estonia) there are 9 (2) observations missing which are filled by the last available value of the respective chain to complete the series. For some months in Lithuania refinery utilization is sometimes slightly larger than 1. This probably comes from data accuracy and calculations from an annual capacity to a monthly levels.

	Country	Austria	Estonia	France		Germany		Italy		Latvia	Lithuania
				Pre	Post	Pre	Post	Pre	Post		
Fuel Price	E5	1.783	1.857	1.916	2.038	1.947	1.861	1.948	1.983	1.803	1.738
	Diesel	1.815	1.759	1.886	2.065	1.881	1.968	1.840	1.960	1.753	1.726
Fuel Margin	E5	0.273	0.392	0.257	0.392	0.204	0.352	0.228	0.372	0.347	0.270
	Diesel	0.386	0.342	0.323	0.505	0.326	0.465	0.253	0.467	0.387	0.367
Relative Fuel Margin/Lerner-Index	E5	0.260	0.345	0.278	0.330	0.220	0.310	0.260	0.320	0.318	0.262
	Diesel	0.337	0.310	0.324	0.392	0.309	0.374	0.277	0.374	0.338	0.329

Table 7.4: Summary statistics of fuel prices and margins by country before (pre) and after (post) the tax decrease (Numbers in Euro per liter, except the relative margins). Averages are based on the country and service station chain pairs.

Despite these differences across countries and service station chains, Table 7.4 reports that the average price level of diesel and gasoline is very similar across the seven countries, although prices are smaller in Austria, which is mainly driven by the low fuel taxes in this country. Concentrating on the treated countries (France, Germany, Italy), we mostly observe higher average consumer prices after the fuel tax reductions (compare *Pre* and *Post* in Table 7.4).[24] Even though this seems to be counterintuitive at first glance, this is mainly driven by the increasing price for crude oil during our observation period (see development of the Brent prices in Figure 7.1), which has mostly overcompensated the decreased fuel taxes. Table 7.4 also shows that the absolute as well as the relative retail margins for diesel and gasoline have increased in Italy and Germany after the fuel tax reductions, while they started to decrease in France after the introduction (on average).[25]

Figure 7.1 visualizes the development of the average median consumer prices and Figure 7.2 shows the daily average retail margins. The figures are divided into sub figures to point out the development of gasoline (upper) and diesel (middle) of the seven European countries during our observation period. The Brent price (lower) is also depicted to highlight the strong link to the market price for crude oil. The vertical lines reflect the introduction of the respective tax reductions in Italy (March 22, yellow), France (April 1, blue), and Germany (June 1, red). In fact, the prices as well as the margins in the seven countries tend to follow the same trend before the policy changes. In all countries, there is also a noticeable increase in both, prices and margins, at the end of February when the war in the Ukraine has started.

With respect to the diesel and gasoline consumer prices, Figure 7.1 shows that both have decreased in the first phase after the respective fuel tax cuts in the three treated countries. However, they tend to increase again after a while which is mainly driven by the price increase for crude oil (depicted in dashed grey in the lower sub figure).[26].

[24] It is worth mentioning that the definition of the *Pre* and *Post* periods is distinct for the three treated countries due to the different implementation dates of the fuel tax reductions.

[25] Relative retail margin reflects the simple Lerner-Indices formula, dividing the absolute margins by net prices.

[26] To have a better understanding of the individual consumer price curves, we additionally present the gasoline and diesel price development for the seven countries separately in Figure 7.7 of 7.C.

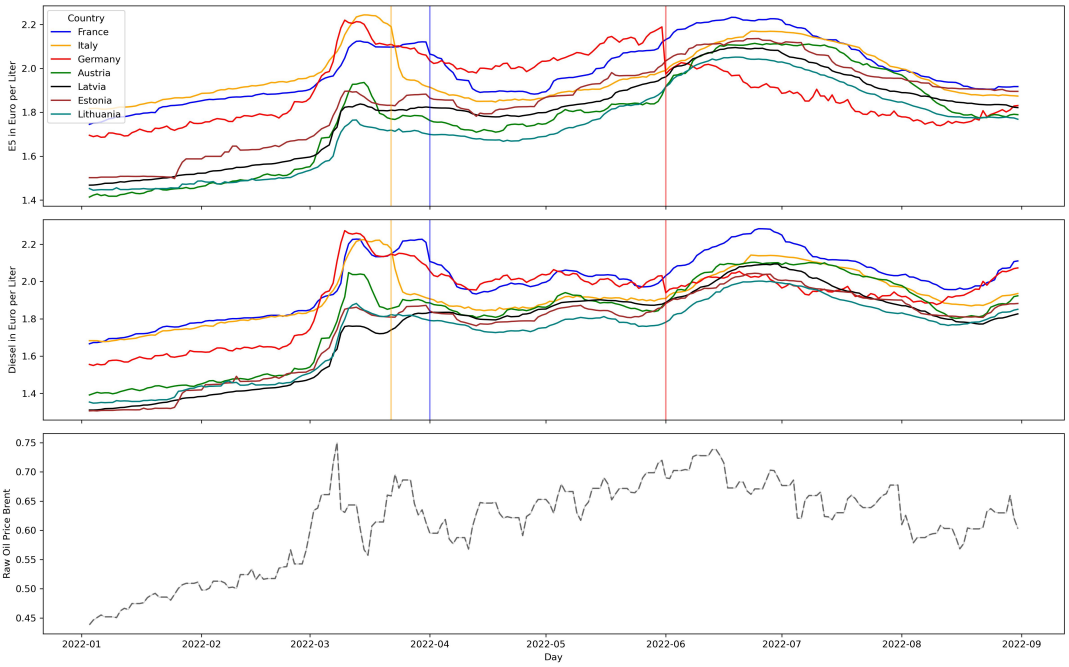


Figure 7.1: Development of average consumer prices for gasoline (upper) and diesel (middle). The vertical lines reflect the introduction of the respective tax reductions in Italy (March 22, yellow), France (April 1, blue), and Germany (June 1, red). Brent prices (lower) in Euro per Liter is denoted in dashed grey.

Simultaneously, the absolute (and also the relative) effect on retail margins exhibits similar trends between the three treated countries (see Figure 7.2). In addition, the margins reveal a slight difference between the countries. While the margins have increased in Germany for diesel and gasoline, in France and Italy they immediately started to decrease after the temporary tax reduction. In relative terms these margins (see Figure 7.10 in 7.B) reflect the Lerner-Indices (Giocoli, 2012; Lerner, 1934) which range from 0 (no market power) to 1 (monopoly market power). Interpretation of this crude measure of market power is problematic, especially without deep knowledge about the exact cost structure on all parts of the vertical chain within the fuel market and should be carried out with caution (Elzinga & Mills, 2011). Therefore, the large increase and long-term shift in the margins seen in the mid of March 2022 may be prominent but the exact cause cannot be determined without detailed market information on costs and is not part of this paper. In this regard, our empirical analysis will show that these changes in margins are on average not affected by the tax reductions.

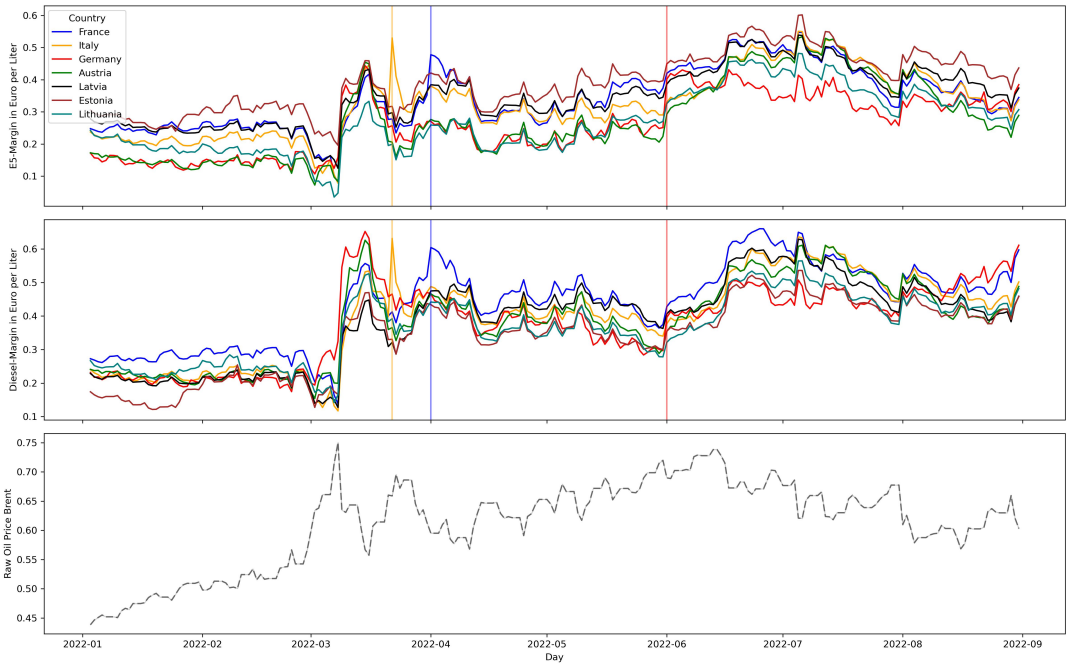


Figure 7.2: Development of average retail margins for gasoline (upper) and diesel (middle). The vertical lines reflect the introduction of the respective tax reductions in Italy (March 22, yellow), France (April 1, blue), and Germany (June 1, red). Brent prices (lower) in Euro per Liter is denoted in dashed grey. See relative retail margins in Figure 7.10 in 7.B.

7.5 Methodology

In our empirical analysis, we estimate the impact of the temporary fuel tax reductions on fuel prices and retail margins. In order to do this, we compare the evolution of consumer prices and retail margins at fuel stations in Germany, France, Italy, Austria and the Baltic States, before and after the reductions of the fuel taxes.

We apply a staggered Difference-in-Differences (DiD) design to causally estimate the effect of the temporary fuel tax reduction on fuel prices and retail margins. In contrast to the canonical DiD setup, the staggered design allows to estimate the unbiased average treatment effect on the treated (ATT) when there are more than two time periods and variation in timing of the treatment. This design is more credible and robust than the canonical DiD with a single treatment period because including multiple treatments plausibly alleviates concerns that contemporaneous trends drive the observed treatment effects (see, e.g., Baker et al., 2022). Goodman-Bacon (2021) shows that time-varying treatment effects can create a bias in the static two-way fixed effects (TWFE) DiD estimate since earlier-treated units act as effective

controls for later-treated units so that the resultant DiD estimates could reflect differences in treatment effects over time between different treatment groups.

Hence, more recent papers propose alternative DiD estimators that do not suffer from the pitfalls associated with TWFE described above (Callaway & Sant'Anna, 2021; De Chaisemartin & d'Haultfoeuille, 2020; Sun & Abraham, 2021). We follow the recent DiD methodology developed by Callaway and Sant'Anna (2021) as it allows to estimate a time-varying and cohort-specific ATT using not-yet-treated or never-treated as clean controls. Specifically, the estimation strategy follows the stylized regression:[27]

$$y_{ijt} = X'\beta + \tau_{it} \cdot TAX_{it} + \eta_{ij} + \lambda_t + \epsilon_{ijt}, \tag{7.1}$$

where y_{ijt} denotes the consumer price (or retail margin) of gasoline or diesel sold by gas station chain j in country i at date t, and TAX_{it} is a dummy variable that equals one when country i implements a temporary fuel tax reduction at date t (note that France, Italy and Germany implemented there reductions at different dates, see Section 7.3). The vector X' contains our control variables refinery utilization rate, total imports of oil and petroleum imports and number of gas stations.[28] The variable η_{ij} corresponds to country service station chain fixed effects and controls for any time-invariant differences between the countries in our dataset. Finally, λ_t gives the day fixed effects, which capture the transitory shocks that identically affect the individual countries, such as fluctuations in the price of crude oil or the conflict in the Ukraine.

Let us further assume that G_i contains i different states treated at different points of time and C_i is a set of never treated states. Then, under the parallel trend and anticipation assumptions (J. Wooldridge, 2021) we can estimate the ATT for a treatment-timing group g at a point in time

[27] Depicting this equation's purpose is to intuitively highlight the estimation strategy. Exact estimation will rely on Equations (7.2) and (7.3) as well as Footnote 28.

[28] Note that we can add time-invariant variables when using the approach from Callaway and Sant'Anna (2021) because those variables are interacted with the day fixed effects. Thus, the covariates are not collinear with our state fixed effects, but act more like state-specific time trends. Technically, we use an inverse probability weighting (IPW) to rebalance the distribution of covariates and estimate reweighted ATTs (Abadie, 2005).

as the group-time average treatment effect using never-treated (7.2) or not-yet-treated (7.3) units as controls by using the R package as provided by Callaway and Sant'Anna (2021):[29]

$$ATT(g,t) = E[Y_t - Y_{g-1}|G = g] - E[Y_t - Y_{g-1}|C = 1]. \tag{7.2}$$

$$ATT(g,t) = E[Y_t - Y_{g-1}|G = g] - E[Y_t - Y_{g-1}|D_t = 0, G \neq g]. \tag{7.3}$$

In the Equations (7.2) and (7.3), t indexes the time in days, g gives the period in which country i is treated and Y_{it} is the fuel price or retail margin of country i.

Finally, we can average the $ATT(g,t)$ over all countries:

$$\Theta_S(g) = \frac{1}{T-g+1} \sum_{t=2}^{T} 1\{g \leq t\} ATT(g,t). \tag{7.4}$$

Equation (7.4) then gives the time-average for each group and the overall average respectively. As already mentioned above, we use the fuel prices of seven different European countries to causally identify the effect of the temporary fuel tax reductions. Thereby, Germany, Italy, and France are the treated countries and Austria as well as the Baltic States are the never-treated countries in our staggered DiD approach.

We also want to estimate the treatment effect heterogeneity over time as the effect of the temporary fuel tax reductions on the retail prices might be dynamic. Using an event study design we can prove the process of tax pass-through over time to check whether there is an effect, how many periods it takes to have an effect, and how long it lasts. Moreover, we can test the parallel trend assumption checking the pre-treatment estimators. Hence, based on Equations (7.2) and (7.3) we provide an event study including pseudo-ATTs for the pre-period and ATTs for the post-period.

To perform the described analysis, various requirements for an unbiased and exogenous estimation have to be satisfied. In general, we can assume that the countries in our data set are very comparable. They are all members of the European Single Market, which implies

[29] See https://bcallaway11.github.io/did/index.html.

harmonized border checks, common customs policy, and identical regulatory procedures on the movement of goods within the European Union (EU). Beyond, the seven countries are relatively similar in their geographic location and have highly correlated public and school holidays. In our observation period, also the travel restrictions put in place due to the COVID-19 crisis were similar and no major reforms, which could also affect fuel prices, were implemented.

Furthermore, to causally identify an unbiased ATT of the temporary fuel tax reductions on fuel prices, there should also be no other transitory shocks that would differently affect fuel prices in the individual countries before and after the tax reduction. Due to their geographic proximity the petroleum companies in the seven countries procure most of their crude oil from similar sources. Finally, we also focus on a relatively narrow window around the tax reductions, which should alleviate concerns on transitory shocks differently affecting the seven countries.

Moreover, requirements have to be considered in the context of government actions tackling high energy prices observed due to the start of the Ukraine war (see Chapter 7.1). The fuel tax reductions were implemented as a measure to counter high inflation rates which may induce the idea of potential endogeneity between tax reductions and price changes, i.e. price changes are affected by the tax reduction and vice versa. There are two major arguments why endogeneity does not pose a problem in our analysis. First, depending on time intervals of tax changes, i.e. tax changes every month, frequent changes might pose a problem for identifying an unbiased effect, especially if long-run relationships are examined (Kaufmann, 2019). However, excise duty reductions by the European governments do not occur on a regular basis,[30] and the interventions in 2022 were implemented at short notice. The energy taxation in Europe is more rigid compared with e.g. the taxation in the US (Kaufmann, 2019) and the interventions in 2022 were a reaction on an exogenous shock, namely the war in the Ukraine. Second, we utilize a panel data set on a service station chain level and a (staggered) DiD design to compare the decisions of countries that implemented a tax reduction to countries that decided against introducing such an intervention. All European countries were equally

[30] Excise duties on gasoline and diesel have not been changed in Germany since 2006, in Italy since (at least) 2016, and in France since 2018 (prior to the reduction in 2022). For the legal basis, see: https://ec.europa.eu/taxation_customs/tedb/index.html. See also Figure 7.8 that highlights exact tax changes in 2022.

affected by high fluctuations of the crude oil price (see Figure 7.1), which is accounted for by time fixed effects. This design mitigates a bias for the ATT as long as the control groups are not affected by a similar intervention and generally provide a reasonable comparison group to construct an appropriate counterfactual, i.e. the parallel trend assumption must be satisfied (J. Wooldridge, 2021). Our analysis meets this requirement, especially because the never-treated part of the control countries did not receive any fuel tax reductions during our observation period. Figures provided in Chapter 7.6.2 (Figure 7.3 and Figure 7.4) provide evidence in form of event studies, highlighting that the assumption of a common trend can be assumed to be satisfied.

7.6 Results

7.6.1 Baseline Results

Table 7.5 presents the results of estimating regression equation (7.4) using the consumer price for gasoline and diesel as outcome variables. The coefficients in columns (I) and (II) correspond to the average treatment effect of the temporary fuel tax reductions on gasoline and diesel in France, Italy and Germany without any other control variables. Columns (III) and (IV) show the effect on consumer prices when we additionally control for the supply side parameters refinery utilization, number of gas stations and the total imports of oil and petroleum products.

Table 7.5: Staggered DiD, referring to estimates of (7.2) and (7.3) that are averaged by means of (7.4). Approach with consumer prices as outcome variable. Bootstrapped (robust) standard errors provided in parentheses are clustered on the country and service station chain level.

	(I)	(II)	(III)	(IV)
	Gasoline	Diesel	Gasoline	Diesel
Italy	-0.32*** (0.02)	-0.30*** (0.02)	-0.35*** (0.02)	-0.30*** (0.02)
France	-0.19*** (0.02)	-0.18*** (0.01)	-0.20*** (0.02)	-0.19*** (0.01)
Germany	-0.36*** (0.02)	-0.15*** (0.01)	-0.39*** (0.02)	-0.18*** (0.01)
Simple Weighted Avg.	-0.31*** (0.01)	-0.20*** (0.01)	-0.33*** (0.01)	-0.21*** (0.01)
Pass-Through Italy	106.09%	98.67%	114.27†	100.00%
Pass-Through France	104.65%	99.91%	111.37%	105.99%
Pass-Through Germany	103.52%	89.73%	110.32%	106.69%
Time FE	Yes	Yes	Yes	Yes
Country and Chain FE	Yes	Yes	Yes	Yes
Supply Parameters	No	No	Yes	Yes
Observations	12,515	12,515	12,515	12,515

Note: H_0: No effect. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; H_0: 100% Pass-Through: $^{\dagger}p < 0.05$, $^{\ddagger}p < 0.01$

In general, the results in Table 7.5 show that the fuel tax reductions led to a statistically significant decline in the average consumer prices of both fuel types for all three countries treated ($p < 0.001$). For our model without any covariates, in Germany the average price for diesel decreases by 15 cents per liter after the fuel tax reduction (column (I)), whilst the average price for gasoline decreases by about 36 cents per liter (column (II)). Also for France the price decrease for diesel (-18 cents per liter) is slightly lower compared to the one for gasoline (-19 cents per liter). With a price drop of more than 30 cents per liter for diesel and 32 cents per liter for gasoline, also the estimated pass-through rates in Italy are very high. Including additional control variables (columns (III) and (IV)) only quantitatively changes our estimation results, even though we apparently underestimate the average treatment effects without controlling for the supply side.

In a next step, we can calculate the average pass-through rates of the fuel tax reductions. Therefore, we divide the estimated coefficients by the actual tax reductions in the three

countries.[31] The estimated pass-through rates in Table 7.5 mostly imply a full- or even an over-shifting of the temporary fuel tax reductions. In our baseline estimations (columns (I) and (II) in Table 7.5), there is an over-shifting for gasoline and almost a full-shifting for diesel. With an estimated pass-through rate of approx. 106%, Italy has the highest passing on of the temporary fuel tax reduction for gasoline and France the highest one for diesel (approx. 100%). Overall, the estimated rates are very similar in the three countries, even though the estimated pass-through rate for diesel is slightly lower in Germany (approx. 90%). Including the control variables for the supply side (columns (III) and (IV) in Table 7.5) into our regression model generally increases the estimated pass-through rates so that we also find a full- or over-shifting for diesel now. In general, the high pass-trough rates might be explained by the inelastic demand for fuel products and particularly by the high public awareness as well as the threat of policymakers to pursue antitrust measures. The 2022 fuel tax reductions had great political and economic implications so that there was a high attention in the public debate (Kahl, 2023).

However, testing whether the average pass-through rates are statistically significant different from a full pass-through (100%) shows that all but one are not different to 100% (see Table 7.5). Only the pass-through rate of gasoline in Italy in column (III) is statistically significantly higher than 100% ($p < 0.05$).

Beside the general high average pass-through rates, a second interesting finding is that the effects of the tax reductions are mostly higher for gasoline compared to diesel in our estimations. This is in sharp contrast to the literature that finds a more inelastic demand for diesel compared to gasoline (Ajanovic et al., 2012; Fridstrøm & Østli, 2021; Karagiannis et al., 2011). However, the Russian invasion of Ukraine led to a high uncertainty of consumers in the energy markets in 2022, which was combined by an unusually high demand for heating diesel in spring and summer 2022. Households increased their heating diesel stocks out of fear of continuously rising prices, because they expected even higher prices in the future. This phenomena was particularly present in Germany.[32] As heating diesel is a close substitute for diesel (whereas gasoline is not), this might explain the lower pass-through rates for diesel in our results.

[31] For instance, in our baseline estimation for diesel (column (I) of Table 7.5) the pass-through rate for Germany can be calculated as follows: $passthrough = \frac{EstCoeff}{TaxReduction} = \frac{15}{16.7} = 0.8973 = 89.73\%$.

[32] See https://www.dw.com/en/german-residents-make-plans-amid-fears-of-a-winter-gas-shortage/a-62482737.

However, scrutinizing differences of the average gasoline estimates against average diesel estimates in each country by means of a hypothesis test reveals no statistically significant differences. Table 7.6 shows the p-values for each of the tests, with none rejecting the null hypothesis based on common significance levels.[33]

	H_0: (I) = (II)	H_0: (III) = (IV)
Italy	0.4451	0.2392
France	0.7679	0.7403
Germany	0.2582	0.7104

Table 7.6: T-test of gasoline versus diesel pass-through rates from Table 7.5. Values are p-values. See Footnote 33 for details on hypothesis construction.

7.6.2 Pre-Treatment Trends and Dynamic Effects

To check whether the estimated results are causal effects and to highlight evolution of pass-through rates over time, we will present an event study design next. The crucial assumption to interpret the results as causally is the parallel trends assumption. Even though this assumption is not directly testable, the event study design does lead to a formal test of pre-treatment trends. With this approach, we can also observe the treatment effects of the fuel tax reductions over time.

Figure 7.3 presents the group-time average treatment effects from Equation (7.3) for gasoline in the three treated countries.[34] We use the regression model including all control variables and compute bootstrapped 95% confidence intervals. Moreover, we apply a varying base period which means that a pseudo-ATT is computed in each treatment period by comparing

[33] Hypothesis test is constructed by each model and country where the t-statistic is then given by $t = \frac{\frac{EstCoeff_{Gasoline}}{TaxReduction_{Gasoline}} - \frac{EstCoeff_{Diesel}}{TaxReduction_{Diesel}}}{\sqrt{(\frac{StdError_{Gasoline}}{TaxReduction_{Gasoline}})^2 + (\frac{StdError_{Diesel}}{TaxReduction_{Diesel}})^2}}$, following J. M. Wooldridge, 2015, Chapter. 4-4. Note: The test usually contains a covariance term in the denominator between the two variables. In our comparison the pass-through rates are independent because the underlying estimation is run separately, thus there is no covariance term between gasoline and diesel. Hence, from an econometricians' viewpoint most average pass-through rates are neither statistically different from a complete pass-through nor do the average pass-through rates differ between gasoline and diesel. Overall, the average pass-through rates are mostly in line with findings of the literature on Germany 2022 which imply a full pass-through (Bernhardt et al., 2023; Dovern et al., 2023; Kahl, 2023; Schmerer & Hansen, 2023; Seiler & Stöckmann, 2023).

[34] In Figure 7.11 of 7.C, we present the same dynamic analysis but without including any covariates. The results there are qualitatively very similar.

the changes in outcomes for a particular group relative to the comparison group in the pre-treatment periods.[35] This just means that we compute changes in the pre-treatment periods from period $t-1$ to period t, but repeatedly change the value of t (Callaway & Sant'Anna, 2021). The pre-treatment coefficients are close to zero and mostly insignificant in all three countries providing supportive evidence for the common trend assumption. An exception is the time of the beginning of Russia's invasion in the Ukraine, which leads to a short divergence in the pre-trends for Italy and France. However, the pre-trends converge back to the zero line and are statistically insignificant shortly before the exogenous shocks of the tax cuts in all three countries.

In Figure 7.3 we can also observe that the treatment effects over time are negative and mostly statistically different from zero. In Germany, there is an immediate drop at the day of the fuel tax reduction with almost full pass-through (red dashed horizontal line). There is a similar development in Italy, where full pass-trough is already reached three days after the tax cut. In the following, there is an over-shifting in both countries. On the contrary, in France it takes almost two weeks until there is a full pass-through. This is in line with our theoretical predictions, since France has a more inelastic supply side compared to the two other treated countries (see Section 7.3). Overall, Figure 7.3 suggests that the treatment effects are relatively stable over time in all three countries.

[35] Pseudo-ATT means that we estimate the effect of participating in the treatment if the treatment had occurred in that period (instead of when it acutally occurred).

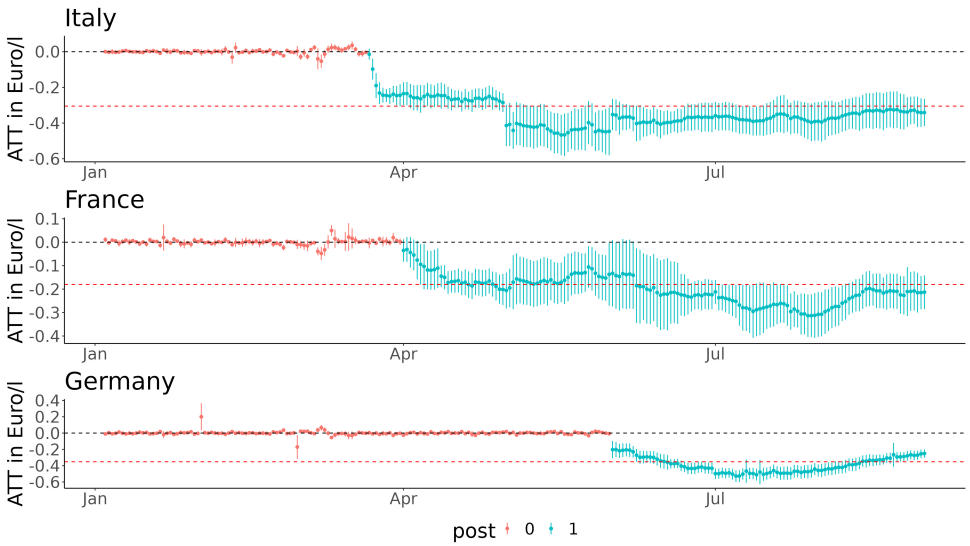


Figure 7.3: Event Study of prices with gasoline (E5) and all covariates. Bootstrapped (robust) standard errors are clustered on the country and service station chain level. Error bars represent 95% confidence intervals. Red dashed horizontal line depicts the value for a full-pass-through.

Figure 7.4 shows the group-time average treatment effects for diesel from Equation (7.3).[36] The pre-treatment coefficients are again close to zero and mostly insignificant (except during the start of the Ukraine conflict for Italy and France). The pattern of the treatment effects over time is very similar compared to the event study for gasoline in Figure 7.3. While Italy has a relatively fast full pass-through again, it takes some time in France until there is a significant effect and even longer for a full-pass through. In Germany, we again observe an immediate drop in the treatment effects at the day of the fuel tax reduction. Again, the treatment effects are relatively stable over time, even though the effects get insignificant for Germany in the end of August. This can be explained by the drought in Germany throughout the summer of 2022. The very high temperatures led to exceptional low water levels in German rivers which, in turn, raised the transportation costs for diesel imports (Dovern et al., 2023).

[36] In Figure 7.12 of 7.C, we present the same dynamic analysis but without including any covariates. The results there are again qualitatively very similar.

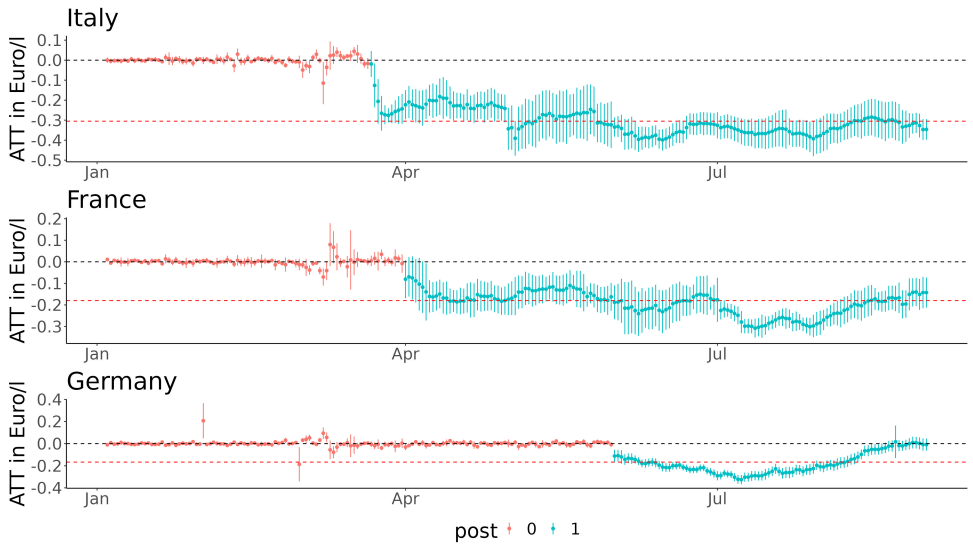


Figure 7.4: Event Study of prices with diesel and all covariates. Bootstrapped (robust) standard errors are clustered on the country and service station chain level. Error bars represent 95% confidence intervals. Red dashed horizontal line depicts the value for a full-pass-through.

7.6.3 Retail Margins

Table 7.7 shows the results of estimating regression equation (7.3) averaged w.r.t equation (7.4) using the retail margins for gasoline and diesel as outcome variables. The results indicate that the reduction in fuel taxes had no significant effect on the average margins in the three countries. This is in line with our results from Section 7.6.1 as we mostly find a full-shifting of the temporary fuel tax reductions which, on average, should not have an effect on the retail margins.

Table 7.7: Staggered DiD, referring to estimates of (7.2) and (7.3) that are averaged by means of (7.4). Approach with retail margins as outcome variable. Bootstrapped (robust) standard errors provided in parentheses are clustered on the country and service station chain level.

	(I) Gasoline	(II) Diesel	(III) Gasoline	(IV) Diesel
Italy	-0.02 (0.01)	0.00 (0.01)	-0.01 (0.02)	0.02 (0.01)
France	-0.01 (0.02)	0.00 (0.01)	-0.01 (0.02)	-0.00 (0.01)
Germany	-0.01 (0.02)	0.02 (0.01)	-0.02 (0.02)	0.00 (0.01)
Simple Weighted Avg.	-0.01 (0.01)	0.01 (0.01)	-0.01 (0.01)	0.00 (0.01)
Time FE	Yes	Yes	Yes	Yes
Country and Chain FE	Yes	Yes	Yes	Yes
Supply Parameters	No	No	Yes	Yes
Observations	12,515	12,515	12,515	12,515

Note: H_0: No effect. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

However, performing an event study design for the outcome variable retail margins implies that there are some positive margins for diesel as well as gasoline in the first days after the fuel tax reductions. Figure 7.5 presents the event study results for gasoline in the three treated countries. The margins are significantly positive in the first days after the tax cuts for Italy and France, but insignificantly for Germany. This is in line with our findings in Figure 7.3 because it takes some days in Italy and France until the tax reduction is passed through to consumers. Since those daily average treatment effects get insignificantly after a few days, the overall average treatment effects in Table 7.7 are still insignificant. In contrast, there is an immediate drop in Germany in the gasoline prices, which results in the insignificant margins also in the first days after the tax cut there.

In the following weeks, we even find some negative average treatment effects in Figure 7.5. For instance, in Germany there is a drop in the estimated ATTs in June and July. Again, this corresponds to our estimated pass-through rates for gasoline (see Figure 7.3) because for this months we find an over-shifting of the tax reduction in Germany. Since this means that the German petroleum companies passed through more than 100% of the temporary tax reduction to the consumers at this time, their retail margins are lower compared to the counterfactual scenario where there had been no tax cut.

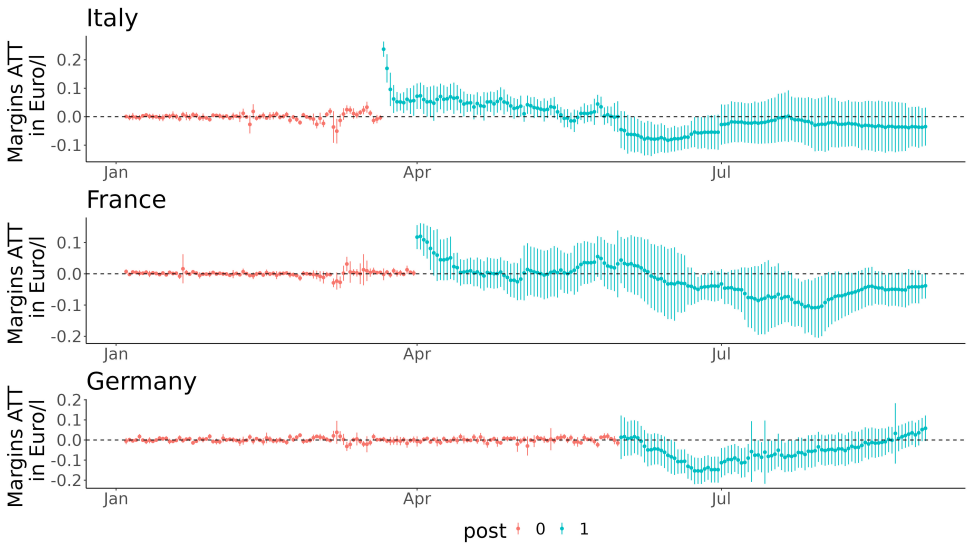


Figure 7.5: Event Study of Margins with gasoline (E5) and all covariates. Bootstrapped (robust) standard errors are clustered on the country and service station chain level. Error bars represent 95% confidence intervals.

Figure 7.6 shows the equivalent results for the retail margins of diesel. Beside the positive retail margins in Italy and France, we also find some positive diesel margins in the first days after the tax cut for Germany now. During the time of the over-shifting of the fuel tax reductions (see Figure 7.4), we even find some significant negative average treatment effects here. This is similar to the estimated retail margins for gasoline (see Figure 7.5) and again relates to the fact that the petroleum companies passed through more than 100% of the temporary fuel tax reductions to the consumers at these times leading to lower retail margins compared to the counterfactual scenarios. Overall, those effects over time cancel each other out so that we have no significant average effect in Table 7.7. This also relates to our main estimations in Table 7.5 where we find that the temporary tax reductions are mostly passed to consumers on a one-to-one basis on average, which should not lead to any significant changes in the average retail margins.

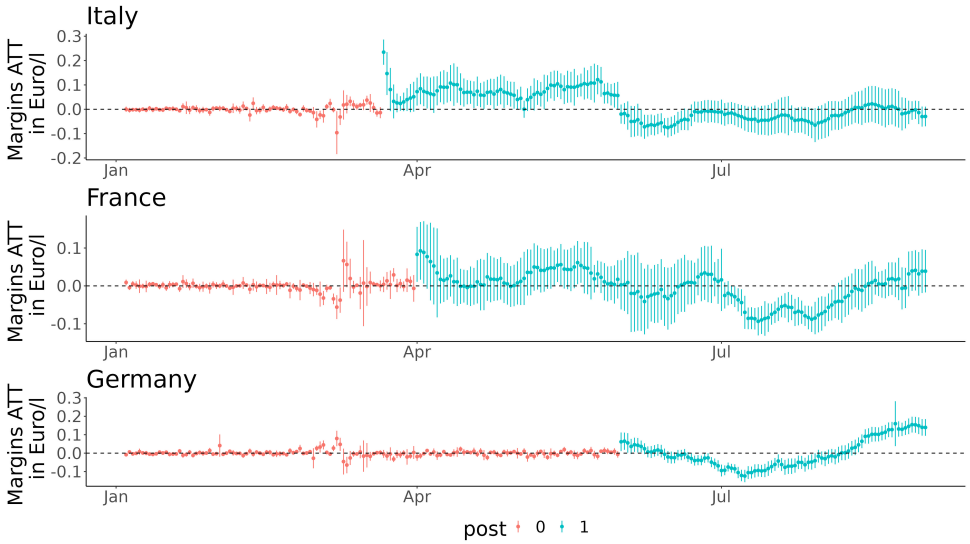


Figure 7.6: Event Study of Margins with diesel and all covariates. Bootstrapped (robust) standard errors are clustered on the country and service station chain level. Error bars represent 95% confidence intervals.

7.7 Conclusion and Policy Implications

This paper provides empirical evidence on the pass-through of temporary fuel tax reductions in the three largest European economies. The governments in Italy, France and Germany introduced relief packages to mitigate the effects of increasing energy prices in the course of post-pandemic economic recovery and the Russian aggression towards the Ukraine. As a part of those packages, the three countries reduced the fuel taxes (introduced a discount on fuel) for several months in 2022. Since the individual measures have taken place at different points of time, we apply a staggered DiD design to causally estimate pass-through rates as well as changes in retail margins.

Our results imply a heterogeneous pass-through over time of the fuel tax reductions depending on the country and type of fuel. We mostly find a full-shifting of the temporary fuel tax reductions meaning the estimated average pass-through rates are close to 100%. This identifies the fuel markets in the three countries as highly competitive, where the consumers enjoy all of the tax reliefs. High pass-through rates can be explained by the general inelastic demand for fuel products and particularly by the high public awareness as well as the threat of policymakers to pursue antitrust measures during the 2022 tax cuts.

A second finding of our paper is that the average pass-through rates are generally higher for gasoline compared to diesel. However, hypothesis tests indicate that, from a statistical perspective, there is no discernible distinction between the pass-through rates for gasoline and diesel. Besides this statistical perspective, the average findings are in contrast to the previous literature, which finds a more inelastic demand for diesel compared to gasoline, this might be explained by the unusual market situation in 2022. The Russian invasion of Ukraine led to a high uncertainty of consumers in the European energy markets, which (among others) resulted in a higher demand for heating diesel, a close substitute for diesel.

Analyzing dynamics associated with time within the framework of an event study reveals differences with regard to the development of the ATTs between countries and types of fuel. The period of time until a full pass-through is reached for the first time differs, and, in addition, different periods of over- and full-shifting are observable.

With respect to the margins, we find no significant effect different from zero on the average retail margins in the three countries. This is in line with the estimated pass-through rates because the tax reductions were mostly passed on to the consumers one-for-one, which does not change the retail margins of the petroleum companies. However, performing an event study design suggests that the petroleum companies have made some positive retail margins at least in the first days after the fuel tax reductions as it has taken some days until the tax reduction has been fully passed-through to the consumers.

A key takeaway from our paper for policymakers is that temporary fuel tax reductions seem to be a suitable measure to lower consumer prices for diesel and gasoline, even though it may take some time until full pass-through is reached. Hence, the primary goal of the governments to relieve their citizens by achieving lower consumer prices for petroleum products has been met. Whether the corrective goal of a Pigouvian tax or subsidy can be achieved generally depends on whether the consumers also bear the incidence of the measure. In this context, the fuel markets in the three countries seem to be competitive enough so that environmental taxes are passed on to the consumers. However, due to the distributional- and climate-economical shortcomings as well as the relatively high fiscal burden of fuel tax reductions it is debatable whether a temporary fuel tax reduction is an suitable intervention at all.

From a competition policy perspective, our results hardly allow any conclusions to be drawn about whether there are competition restrictions in the fuel market at all. However, the estimated pass-through rates in the three countries imply that the alleged restrictions can at least not hinder a high pass-through of the tax reductions. In general, comprehensive sector analyses by the competition authorities to find the mildest means of competition policy seem to be more appropriate than short-term government interventions in the fuel market.

Apart from already mentioned limitations regarding policy implications, data limitations do not allow to make any statements on welfare effects, as we cannot observe the traded volumes. Furthermore, due to the aggregated price data at service chain level, it is not possible to look at regional effects within individual countries. However, the geographic location of the service stations included in the dataset shows that we observe a balanced geographic coverage, which implies that the average effects within countries are robust. With regard to our observation period and the design of the measures studied, only temporary effects are analyzable. For further studies, it would be interesting to extend the period and also examine the end of the measures and the associated tax increases under the subject of asymmetric pass-through of increasing and decreasing costs, i.e. rockets and feathers. Overall, it is crucial to emphasize that the obtained results are not readily transferable or applicable to other industries. The retail fuel market (and any other market) is characterized by unique features and therefore an own empirical assessment of the pass-through of tax reductions in other industries would be necessary.

Nevertheless, our work provides new and important insights into the transmission of tax reductions in a dynamic and much studied industry, using the most recent methods.

Bibliography

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The review of economic studies*, *72*(1), 1–19.

Ajanovic, A., Dahl, C., & Schipper, L. (2012). Modelling transport (energy) demand and policies—an introduction. *Energy Policy*, *41*(2012), iii–xiv.

Assad, S., Clark, R., Ershov, D., & Xu, L. (2023). Algorithmic pricing and competition: Empirical evidence from the german retail gasoline market. *Journal of Political Economy*, *, forthcoming*.

Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics*, *144*(2), 370–395.

Bello, A., & Contín-Pilart, I. (2012). Taxes, cost and demand shifters as determinants in the regional gasoline price formation process: Evidence from spain. *Energy Policy*, *48*, 439–448.

Bernhardt, L., Breiderhoff, X., & Dewenter, R. (2023). The impact of the tax reduction on fuel prices in germany–a synthetic difference-in-differences approach. *Review of Economics*, *74*(2), 141–160.

Byrne, D. P., & De Roos, N. (2019). Learning to coordinate: A study in retail gasoline. *American Economic Review*, *109*(2), 591–619.

Callaway, B., & Sant'Anna, P. H. (2021). Did: Difference in differences [R package version 2.1.2]. https://bcallaway11.github.io/did/

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

De Chaisemartin, C., & d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964–96.

Dewenter, R., Heimeshoff, U., & Lüth, H. (2017). The impact of the market transparency unit for fuels on gasoline prices in germany. *Applied Economics Letters*, *24*(5), 302–305.

Dovern, J., Frank, J., Glas, A., Müller, L. S., & Perico Ortiz, D. (2023). Estimating pass-through rates for the 2022 tax reduction on fuel prices in germany. *Energy Economics*.

Doyle Jr, J. J., & Samphantharak, K. (2008). $2.00 gas! studying the effects of a gas tax moratorium. *Journal of public economics*, *92*(3-4), 869–884.

Edgeworth, F. Y. (1897). The pure theory of taxation. *The Economic Journal*, *7*(25), 46–70.

Elzinga, K. G., & Mills, D. E. (2011). The lerner index of monopoly power: Origins and uses. *American Economic Review*, *101*(3), 558–564.

Foros, Ø., & Steen, F. (2013). Vertical control and price cycles in gasoline retailing. *The Scandinavian Journal of Economics*, *115*(3), 640–661.

Fridstrøm, L., & Østli, V. (2021). Direct and cross price elasticities of demand for gasoline, diesel, hybrid and battery electric cars: The case of norway. *European Transport Research Review*, *13*(1), 1–24.

Fuest, C., Neumeier, F., & Stöhlker, D. (2020). *The pass-through of temporary vat rate cuts: Evidence from german retail prices* (tech. rep.). ifo Working Paper.

Fuest, C., Neumeier, F., & Stöhlker, D. (2022). Der tankrabatt: Haben die mineralölkonzerne die steuersenkung an die kunden weitergegeben? *Perspektiven der Wirtschaftspolitik*, *23*(2), 74–80.

Genakos, C., & Pagliero, M. (2022). Competition and pass-through: Evidence from isolated markets. *American Economic Journal: Applied Economics*, *14*(4), 35–57.

Giocoli, N. (2012). Who invented the lerner index? luigi amoroso, the dominant firm model, and the measurement of market power. *Review of Industrial Organization*, *41*, 181–191.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*(2), 254–277.

Grasso, M., & Manera, M. (2007). Asymmetric error correction models for the oil–gasoline price relationship. *Energy Policy*, *35*(1), 156–177.

Harju, J., Kosonen, T., Laukkanen, M., & Palanne, K. (2022). The heterogeneous incidence of fuel carbon taxes: Evidence from station-level data. *Journal of Environmental Economics and Management*, *112*, 102607.

Kahl, M. P. (2023). *Was the german fuel discount passed on to consumers?* Leuphana Universität Lüneburg, Institut für Volkswirtschaftslehre.

Karagiannis, S., Panagopoulos, Y., & Vlamis, P. (2011). Symmetric or asymmetric interest rate adjustments? evidence from southeastern europe. *Review of Development Economics*, *15*(2), 370–385.

Kaufmann, R. K. (2019). Pass-through of motor gasoline taxes: Efficiency and efficacy of environmental taxes. *Energy policy*, *125*, 207–215.

Kaufmann, R. K., & Laskowski, C. (2005). Causes for an asymmetric relation between the price of crude oil and refined petroleum products. *Energy policy*, *33*(12), 1587–1596.

Lerner, A. P. (1934). The concept of monopoly and the measurement of monopoly power. *The Review of Economic Studies*, *1*(3), 157–175. Retrieved 2023-07-12, from http://www.jstor.org/stable/2967480

Li, S., Linn, J., & Muehlegger, E. (2014). Gasoline taxes and consumer behavior. *American Economic Journal: Economic Policy*, *6*(4), 302–42.

Marion, J., & Muehlegger, E. (2011). Fuel tax incidence and supply conditions. *Journal of public economics*, *95*(9-10), 1202–1212.

Maskin, E., & Tirole, J. (1988). A theory of dynamic oligopoly, ii: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica: Journal of the Econometric Society*, 571–599.

Montag, F., Sagimuldina, A., & Schnitzer, M. (2021). Does tax policy work when consumers have imperfect price information? theory and evidence.

Noel, M. (2009). Do retail gasoline prices respond asymmetrically to cost shocks? the influence of edgeworth cycles. *The RAND Journal of Economics*, *40*(3), 582–595.

Noel, M. D., et al. (2011). Edgeworth price cycles. *New Palgrave Dictionary of Economics. Palgrave Macmillan*.

Noel, M. D. (2015). Do edgeworth price cycles lead to higher or lower prices? *International Journal of Industrial Organization*, *42*, 81–93.

Perdiguero-Garcia, J. (2013). Symmetric or asymmetric oil prices? a meta-analysis approach. *Energy policy*, *57*, 389–397.

Schmerer, H.-J., & Hansen, J. (2023). Pass-through effects of a temporary tax rebate on german fuel prices. *Economics Letters*, *227*, 111104.

Seiler, V., & Stöckmann, N. (2023). The impact of the german fuel discount on prices at the petrol pump. *German Economic Review*, *24*(2), 191–206.

Silvia, L., & Taylor, C. T. (2014). Tax pass-through in gasoline and diesel fuel: The 2003 washington state nickel funding package increase.

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175–199.

Wooldridge, J. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Appendix 7.A Appendix A

Economic theory implies that the elasticities of demand and supply as well as the competitive situation in a market determine the level of pass-through. Following Weyl and Fabinger (2013), we denote p as the retail price and t as the quantity tax rate, so that the pass-through rate is given by $\rho = \frac{dp}{dt}$. We further define the elasticity of demand ($\epsilon_D \equiv -(D'p/Q)$) and supply ($\epsilon_S \equiv S'p/Q$). In this framework, Weyl and Fabinger (2013) postulate that the solution of the firm maximization problem can be described by the conduct parameter $\theta = (p - mc(q))/p\epsilon_D$. θ maps the degree of competition in a market. For instance, θ is equal to 0 in perfect and Bertrand competition, equal to 1 in a monopolistic market, and equal to $1/n$ in Cournot competition. Then, the pass-through rate ρ is independently of the specific model given by

$$\rho = \frac{1}{1 + \frac{\theta}{\epsilon_\theta} + \frac{\epsilon_D - \theta}{\epsilon_S} + \frac{\theta}{\epsilon_{ms}}}. \tag{7.5}$$

Aside from the conduct parameter θ, formula (7.5) implies that the pass-through of a marginal cost increase also depends on the elasticity of demand ϵ_D, the elasticity of the inverse marginal cost curve (the elasticity of supply)[37] ϵ_S, the curvature of the demand function ϵ_{ms}[38], and the variation of θ in changes of production ϵ_θ[39].

Even though formula (7.5) suggests that the sign and magnitude of the pass-through is ambiguous, we can simplify the expression for ρ in some special cases. If there is perfect competition in a market ($\theta = 0$), then $\rho = \frac{1}{1+(\epsilon_D/\epsilon_S)}$ so that the pass-through only depends on the ratio of demand and supply elasticity. More generally, if the marginal cost were constant, demand were linear, and θ were constant, expression (7.5) would simplify to $\rho = 1/(1 + \theta)$. A rise in the conduct parameter θ (less competition) would lead to lower pass-through in this situation (Genakos & Pagliero, 2022). For instance, in a monopolistic market ($\theta = 1$) the pass-through would be lower ($\rho = 0.5$) compared to a market with perfect competition ($\theta = 0$) where we would have full pass-through ($\rho = 1$).

However, in general, the sign of the effect of an increase in the conduct parameter θ on the pass-through remains ambiguous. This is especially the case for an oligopolistic market, which should be the most appropriate market form to model the fuel industry in Europe. The impact of the conduct parameter on the pass-through can either be positive or negative, depending on the actual market situation. Under certain assumptions also pass-through rates larger than one are possible. Hence, the impact of the intensity of competition on the pass-through rate in an oligopolistic market remains an empirical problem (Genakos & Pagliero, 2022).

[37] The monopolist determines the price based on demand and its costs, there is, just like in an oligopoly, no supply curve and accordingly, no supply elasticity in the sense of perfect competition.

[38] Given by $\epsilon_{ms} = \frac{ms}{ms'q}$, where ms is the negative of the marginal consumer surplus ($ms = -p'q$). If demand is linear, then $\epsilon_{ms} = 1$, if concave, $\epsilon_{ms} < 1$, and if convex, $\epsilon_{ms} > 1$ (and the opposite is also true) Genakos & Pagliero, 2022.

[39] Given by $\epsilon_\theta = (\theta/q(d\theta/dq))$.

Appendix 7.B Appendix B

To compute the daily average retail margins for the five countries in our data set, we subtract a fuel share of the crude oil price (major input cost) as well as the country-specific taxes and duties (see Montag et al., 2021). For each country in our raw data set, we observe a daily average gross consumer price. In a first calculation step, we calculate the average consumer prices without VAT taxes for every day and country.[40] To get the daily average net price, we then also subtract the excise duties for the individual countries (see Table 7.3). Thereby, for the treated countries we have to differentiate between the period before and after the fuel tax reductions.

In a final step, we have to subtract the input cost of crude oil (Brent) from the daily net price. Therefore, we use the information that around 54% of the Brent oil price per barrel corresponds to the production of 19 gallons of gasoline and around 34% to the production of 12 gallons of diesel.[41] We further transform these measures into the input cost per liter of gasoline and diesel. The retail margins of gasoline and diesel are then computed as the average gross consumer price per liter adjusted to VAT taxes and excise duties minus the share of crude oil price per liter of a corresponding fuel product.

[40] The VAT taxes are very heterogeneous in the five countries: 22% in Italy, 20% in Austria and France, 19% in Germany, and 7.7% in Switzerland.

[41] See https://www.eia.gov/energyexplained/oil-and-petroleum-products/refining-crude-oil-inputs-and-outputs.php. (Last accessed: October 19, 2022)

Appendix 7.C Appendix C

Country	Statistic / Variable	count	mean	std	min	25%	50%	75%	max
	Diesel	1,446	1.81	0.23	1.28	1.67	1.87	1.99	2.15
	E5	1,446	1.78	0.22	1.29	1.61	1.79	1.95	2.16
	Margin of Diesel in Euro/l	1,446	0.39	0.13	0.06	0.26	0.40	0.48	0.67
	Margin of E5 in Euro/l	1,446	0.27	0.13	-0.01	0.18	0.25	0.37	0.56
Austria	Number of Stations per Chain	1,446	196.17	104.28	8.00	117.00	229.50	281.00	312.00
	Relative Margin/Lerner-Index of Diesel	1,446	0.34	0.07	0.08	0.29	0.34	0.39	0.50
	Relative Margin/Lerner-Index of E5	1,446	0.26	0.08	-0.01	0.21	0.25	0.32	0.43
	Total Imports of Oil and Petroleum Products	1,446	987.21	212.86	745.05	779.82	887.39	1,156.51	1,328.22
	Utilization of Capacity	1,446	0.55	0.27	0.25	0.26	0.37	0.78	0.93
	Diesel	1,190	1.76	0.22	1.26	1.67	1.83	1.90	2.06
	E5	1,188	1.86	0.19	1.45	1.74	1.90	1.98	2.16
	Margin of Diesel in Euro/l	1,190	0.34	0.11	0.06	0.24	0.37	0.43	0.55
	Margin of E5 in Euro/l	1,188	0.39	0.09	0.13	0.32	0.40	0.46	0.61
Estonia	Number of Stations per Chain	1,190	55.78	30.58	9.00	34.00	59.00	79.00	95.00
	Relative Margin/Lerner-Index of Diesel	1,190	0.31	0.07	0.09	0.27	0.32	0.36	0.43
	Relative Margin/Lerner-Index of E5	1,188	0.35	0.05	0.13	0.32	0.35	0.38	0.45
	Total Imports of Oil and Petroleum Products	1,190	135.61	24.04	103.00	120.00	131.00	147.00	190.00
	Utilization of Capacity	0							
	Diesel	1,687	2.00	0.18	1.52	1.86	2.01	2.13	2.51
	E5	1,687	1.99	0.16	1.60	1.88	1.97	2.11	2.50
	Margin of Diesel in Euro/l	1,687	0.44	0.14	0.05	0.34	0.45	0.54	0.84
	Margin of E5 in Euro/l	1,687	0.34	0.12	0.00	0.25	0.33	0.42	0.74
France	Number of Stations per Chain	1,687	426.29	216.77	97.00	197.00	419.00	685.00	736.00
	Relative Margin/Lerner-Index of Diesel	1,687	0.37	0.07	0.06	0.33	0.37	0.42	0.52
	Relative Margin/Lerner-Index of E5	1,687	0.31	0.07	0.00	0.27	0.31	0.36	0.49
	Total Imports of Oil and Petroleum Products	1,687	6,749.68	454.79	6,041.00	6,395.00	6,965.00	6,987.00	7,486.00
	Utilization of Capacity	1,687	0.73	0.09	0.60	0.69	0.71	0.85	0.90
	Diesel	3,615	1.91	0.19	1.51	1.81	1.97	2.03	2.49
	E5	3,615	1.91	0.16	1.65	1.77	1.90	2.04	2.40
	Margin of Diesel in Euro/l	3,615	0.38	0.12	0.01	0.26	0.40	0.47	0.96
	Margin of E5 in Euro/l	3,615	0.26	0.10	-0.07	0.17	0.26	0.34	0.84
Germany	Number of Stations per Chain	3,615	706.67	683.04	13.00	188.00	458.00	980.00	2,597.00
	Relative Margin/Lerner-Index of Diesel	3,615	0.33	0.06	0.01	0.28	0.33	0.38	0.57
	Relative Margin/Lerner-Index of E5	3,615	0.25	0.06	-0.09	0.20	0.25	0.31	0.53
	Total Imports of Oil and Petroleum Products	3,615	10,191.52	321.37	9,391.28	10,161.25	10,305.77	10,483.69	10,507.72
	Utilization of Capacity	3,615	0.91	0.04	0.84	0.89	0.92	0.95	0.96
	Diesel	1,928	1.92	0.15	1.62	1.82	1.89	2.03	2.37
	E5	1,928	1.97	0.13	1.75	1.86	1.94	2.07	2.35
	Margin of Diesel in Euro/l	1,928	0.40	0.14	0.07	0.31	0.40	0.50	0.80
	Margin of E5 in Euro/l	1,928	0.33	0.11	0.04	0.24	0.31	0.39	0.71
Italy	Number of Stations per Chain	1,928	2,109.50	1,511.06	176.00	1,017.50	1,973.00	3,054.25	4,437.00
	Relative Margin/Lerner-Index of Diesel	1,928	0.34	0.07	0.07	0.30	0.35	0.39	0.52
	Relative Margin/Lerner-Index of E5	1,928	0.30	0.07	0.04	0.26	0.29	0.35	0.49
	Total Imports of Oil and Petroleum Products	1,928	6,446.63	540.48	5,701.19	5,993.60	6,614.20	6,986.43	7,182.83
	Utilization of Capacity	1,928	0.86	0.09	0.70	0.82	0.92	0.93	0.97
	Diesel	1,444	1.75	0.23	1.28	1.51	1.82	1.90	2.13
	E5	1,444	1.80	0.19	1.44	1.61	1.83	1.93	2.13
	Margin of Diesel in Euro/l	1,444	0.39	0.13	0.06	0.24	0.42	0.47	0.64
	Margin of E5 in Euro/l	1,444	0.35	0.09	0.03	0.27	0.35	0.41	0.57
Latvia	Number of Stations per Chain	1,444	64.63	22.20	30.00	41.00	70.00	88.00	89.00
	Relative Margin/Lerner-Index of Diesel	1,444	0.34	0.07	0.06	0.28	0.35	0.39	0.47
	Relative Margin/Lerner-Index of E5	1,444	0.32	0.05	0.03	0.29	0.32	0.35	0.44
	Total Imports of Oil and Petroleum Products	1,444	184.64	49.18	125.43	154.16	180.90	202.69	298.02
	Utilization of Capacity	0							
	Diesel	1,205	1.73	0.20	1.29	1.55	1.79	1.85	2.04
	E5	1,196	1.74	0.19	1.38	1.58	1.74	1.88	2.10
	Margin of Diesel in Euro/l	1,205	0.37	0.10	0.10	0.28	0.38	0.45	0.59
	Margin of E5 in Euro/l	1,196	0.27	0.10	-0.02	0.19	0.25	0.34	0.53
Lithuania	Number of Stations per Chain	1,205	81.40	25.40	44.00	74.00	76.00	91.00	122.00
	Relative Margin/Lerner-Index of Diesel	1,205	0.33	0.06	0.10	0.29	0.33	0.38	0.45
	Relative Margin/Lerner-Index of E5	1,196	0.26	0.07	-0.02	0.22	0.26	0.31	0.41
	Total Imports of Oil and Petroleum Products	1,205	747.42	245.31	327.60	438.80	764.90	1,059.60	1,061.10
	Utilization of Capacity	1,205	0.81	0.35	0.20	0.24	1.02	1.06	1.07

Table 7.8: Summary statistics of Austria, Estonia, France, Germany, Italy, Latvia, and Lithuania at the service station chain level.

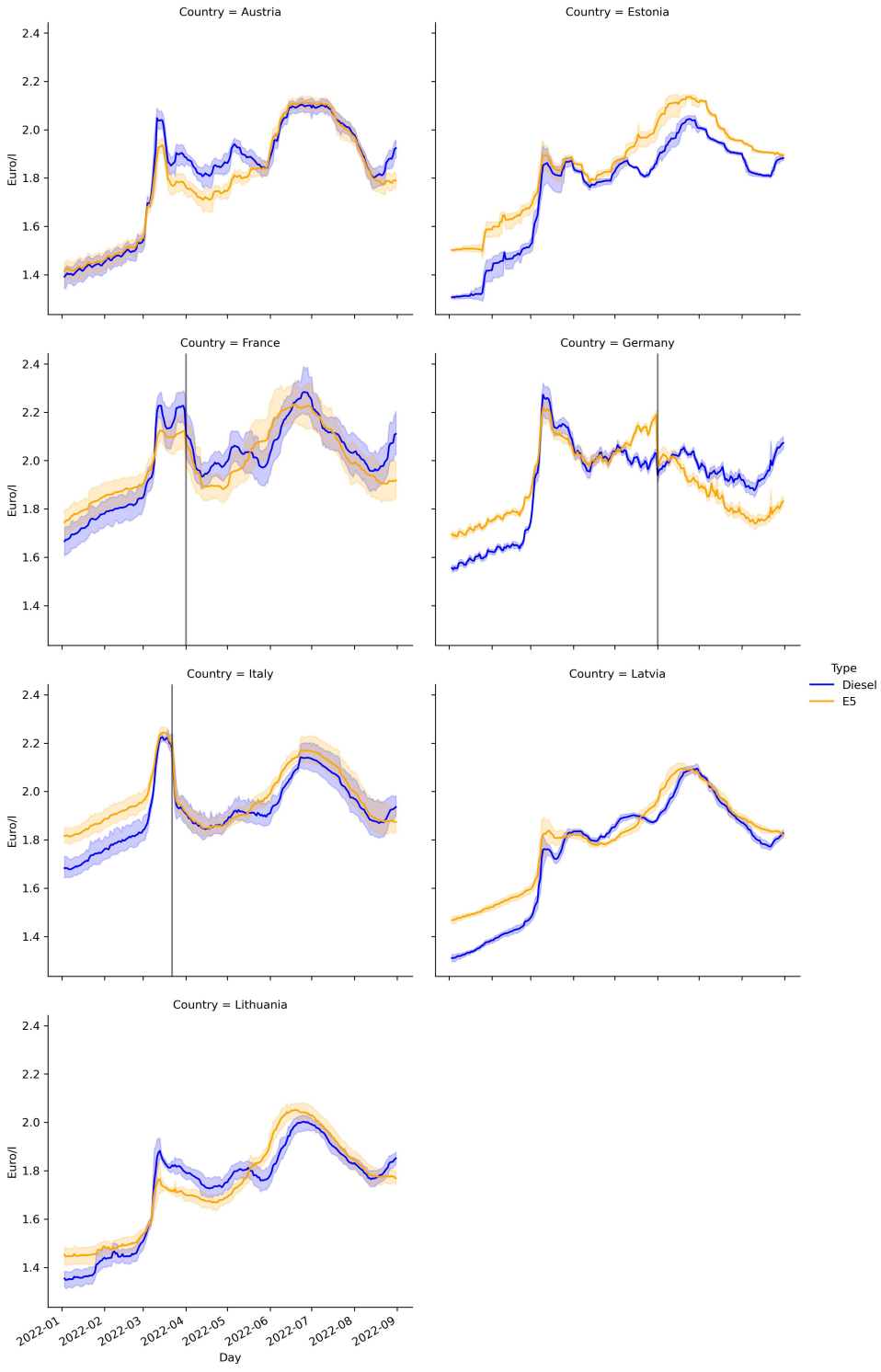


Figure 7.7: Development of gasoline and diesel consumer prices for the seven countries in our data set. The vertical lines reflect the introduction of the respective tax reductions. Confidance band is shown to highlight that data varies on the service station chain level.

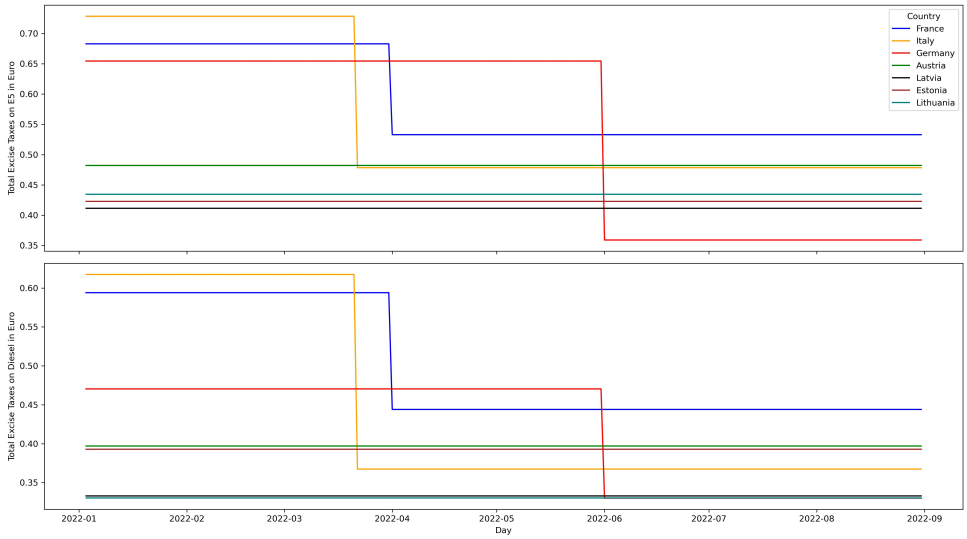


Figure 7.8: Development of gasoline and diesel excise duties (inclusive of further duties) for the seven countries in our data set.

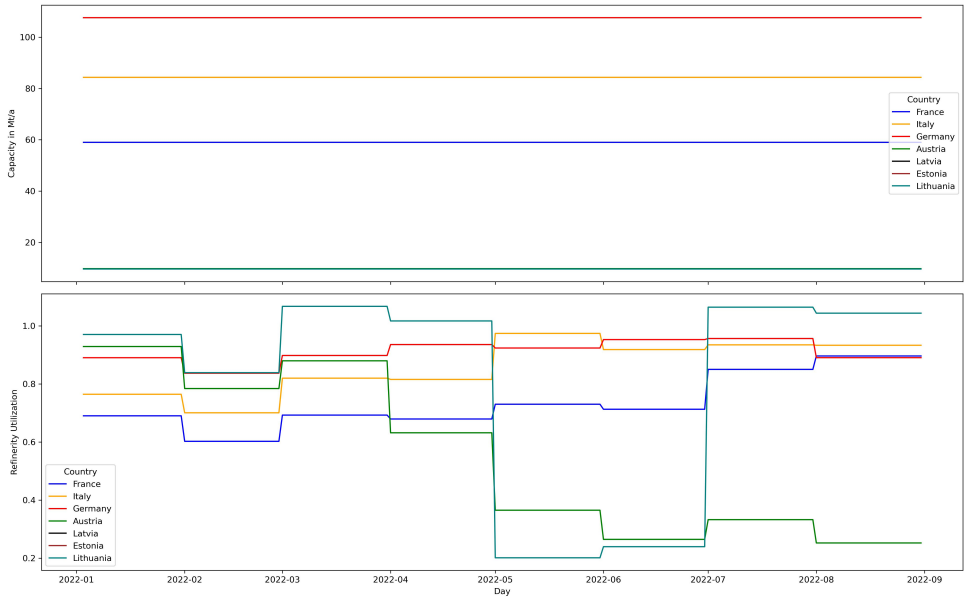


Figure 7.9: Development of Capacity and Refinery Utilization.

Country	Statistic / Variable	count	mean	std	min	25%	50%	75%	max
	Diesel	1,446	1.81	0.23	1.28	1.67	1.87	1.99	2.15
	E5	1,446	1.78	0.22	1.29	1.61	1.79	1.95	2.16
	Margin of Diesel in Euro/l	1,446	0.39	0.13	0.06	0.26	0.40	0.48	0.67
	Margin of E5 in Euro/l	1,446	0.27	0.13	-0.01	0.18	0.25	0.37	0.56
Austria	Number of Stations per Chain	1,446	196.17	104.28	8.00	117.00	229.50	281.00	312.00
	Relative Margin/Lerner-Index of Diesel	1,446	0.34	0.07	0.08	0.29	0.34	0.39	0.50
	Relative Margin/Lerner-Index of E5	1,446	0.26	0.08	-0.01	0.21	0.25	0.32	0.43
	Total Imports of Oil and Petroleum Products	1,446	987.21	212.86	745.05	779.82	887.39	1,156.51	1,328.22
	Utilization of Capacity	1,446	0.55	0.27	0.25	0.26	0.37	0.78	0.93
	Diesel	1,190	1.76	0.22	1.26	1.67	1.83	1.90	2.06
	E5	1,188	1.86	0.19	1.45	1.74	1.90	1.98	2.16
	Margin of Diesel in Euro/l	1,190	0.34	0.11	0.06	0.24	0.37	0.43	0.55
	Margin of E5 in Euro/l	1,188	0.39	0.09	0.13	0.32	0.40	0.46	0.61
Estonia	Number of Stations per Chain	1,190	55.78	30.58	9.00	34.00	59.00	79.00	95.00
	Relative Margin/Lerner-Index of Diesel	1,190	0.31	0.07	0.09	0.27	0.32	0.36	0.43
	Relative Margin/Lerner-Index of E5	1,188	0.35	0.05	0.13	0.32	0.35	0.38	0.45
	Total Imports of Oil and Petroleum Products	1,190	135.61	24.04	103.00	120.00	131.00	147.00	190.00
	Utilization of Capacity	0							
	Diesel	1,687	2.00	0.18	1.52	1.86	2.01	2.13	2.51
	E5	1,687	1.99	0.16	1.60	1.88	1.97	2.11	2.50
	Margin of Diesel in Euro/l	1,687	0.44	0.14	0.05	0.34	0.45	0.54	0.84
	Margin of E5 in Euro/l	1,687	0.34	0.12	0.00	0.25	0.33	0.42	0.74
France	Number of Stations per Chain	1,687	426.29	216.77	97.00	197.00	419.00	685.00	736.00
	Relative Margin/Lerner-Index of Diesel	1,687	0.37	0.07	0.06	0.33	0.37	0.42	0.52
	Relative Margin/Lerner-Index of E5	1,687	0.31	0.07	0.00	0.27	0.31	0.36	0.49
	Total Imports of Oil and Petroleum Products	1,687	6,749.68	454.79	6,041.00	6,395.00	6,965.00	6,987.00	7,486.00
	Utilization of Capacity	1,687	0.73	0.09	0.60	0.69	0.71	0.85	0.90
	Diesel	3,615	1.91	0.19	1.51	1.81	1.97	2.03	2.49
	E5	3,615	1.91	0.16	1.65	1.77	1.90	2.04	2.40
	Margin of Diesel in Euro/l	3,615	0.38	0.12	0.01	0.26	0.40	0.47	0.96
	Margin of E5 in Euro/l	3,615	0.26	0.10	-0.07	0.17	0.26	0.34	0.84
Germany	Number of Stations per Chain	3,615	706.67	683.04	13.00	188.00	458.00	980.00	2,597.00
	Relative Margin/Lerner-Index of Diesel	3,615	0.33	0.06	0.01	0.28	0.33	0.38	0.57
	Relative Margin/Lerner-Index of E5	3,615	0.25	0.06	-0.09	0.20	0.25	0.31	0.53
	Total Imports of Oil and Petroleum Products	3,615	10,191.52	321.37	9,391.28	10,161.25	10,305.77	10,483.69	10,507.72
	Utilization of Capacity	3,615	0.91	0.04	0.84	0.89	0.92	0.95	0.96
	Diesel	2,667	1.93	0.19	1.51	1.84	1.98	2.06	2.36
	E5	2,667	1.98	0.16	1.65	1.89	1.98	2.09	2.37
	Margin of Diesel in Euro/l	2,667	0.31	0.13	-0.02	0.21	0.31	0.41	0.67
	Margin of E5 in Euro/l	2,667	0.32	0.11	-0.03	0.24	0.32	0.40	0.70
Italy	Number of Stations per Chain	2,667	1,508.27	988.63	19.00	756.50	1,819.00	2,319.00	2,774.00
	Relative Margin/Lerner-Index of Diesel	2,667	0.27	0.07	-0.01	0.23	0.27	0.31	0.46
	Relative Margin/Lerner-Index of E5	2,667	0.27	0.06	-0.01	0.23	0.27	0.31	0.46
	Total Imports of Oil and Petroleum Products	2,667	6,385.49	437.68	5,594.38	6,008.47	6,245.89	6,746.58	7,558.76
	Utilization of Capacity	2,667	0.68	0.09	0.55	0.62	0.65	0.76	0.85
	Diesel	728	1.74	0.21	1.25	1.66	1.81	1.89	2.00
	E5	726	1.77	0.20	1.37	1.67	1.83	1.92	2.06
	Margin of Diesel in Euro/l	728	0.33	0.11	0.04	0.24	0.36	0.41	0.56
	Margin of E5 in Euro/l	726	0.37	0.10	0.11	0.30	0.39	0.45	0.61
Latvia	Number of Stations per Chain	728	65.45	36.49	13.00	36.00	63.00	93.00	106.00
	Relative Margin/Lerner-Index of Diesel	728	0.30	0.07	0.08	0.26	0.31	0.35	0.44
	Relative Margin/Lerner-Index of E5	726	0.33	0.06	0.10	0.30	0.34	0.38	0.46
	Total Imports of Oil and Petroleum Products	728	215.13	39.28	168.00	174.00	199.00	253.00	290.00
	Utilization of Capacity	0							
	Diesel	831	1.77	0.22	1.29	1.66	1.83	1.92	2.06
	E5	831	1.84	0.20	1.44	1.71	1.86	1.96	2.16
	Margin of Diesel in Euro/l	831	0.33	0.11	0.04	0.24	0.36	0.42	0.55
	Margin of E5 in Euro/l	831	0.37	0.10	0.10	0.30	0.39	0.44	0.60
Lithuania	Number of Stations per Chain	831	68.00	31.79	13.00	41.00	63.00	90.00	106.00
	Relative Margin/Lerner-Index of Diesel	831	0.30	0.07	0.08	0.26	0.31	0.35	0.44
	Relative Margin/Lerner-Index of E5	831	0.32	0.06	0.09	0.29	0.33	0.36	0.45
	Total Imports of Oil and Petroleum Products	831	287.36	43.15	234.00	241.00	281.00	327.00	372.00
	Utilization of Capacity	831	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 7.9: Summary statistics for number of service stations per chain present in the data. Overall coverage: Austria $1177/2748 \approx 43\%$, Germany $10600/14500 \approx 73\%$, France $2984/11151 \approx 27\%$, Italy $16876/21700 \approx 78\%$, Estonia $276/491 \approx 56\%$, Latvia $388/605 \approx 64\%$ and Lithuania $407/718 \approx 56\%$. Visual inspection of the stations displayed on the map provided by Fuelo reveals the extent of geographical coverage within the national markets. Source: Fuelo.net, https://de.fuelo.net/gasstations?lang=en.

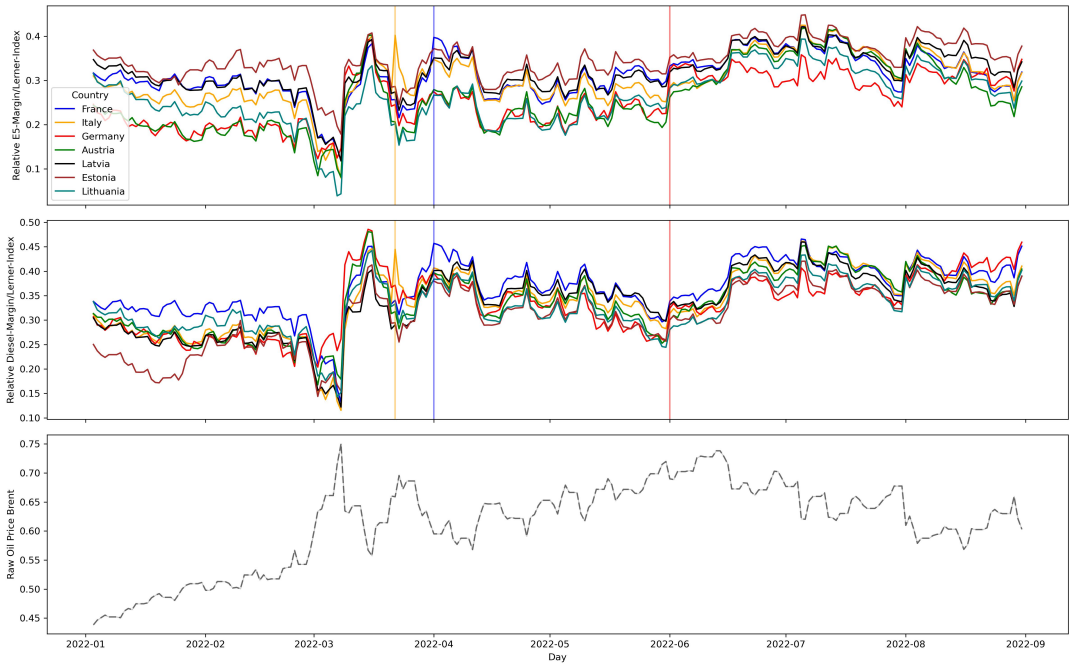


Figure 7.10: Development relative retail margins for gasoline (upper) and diesel (middle). The vertical lines reflect the introduction of the respective tax reductions in Italy (March 22, yellow), France (April 1, blue), and Germany (June 1, red). Brent prices (lower) in Euro per Liter is denoted in dashed grey.

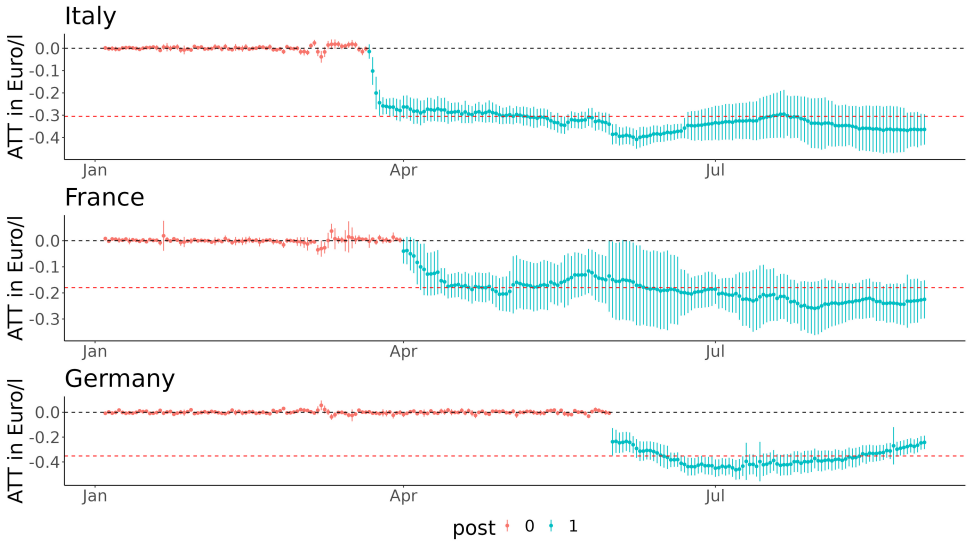


Figure 7.11: Event Study of prices with gasoline (E5) and no covariates. Bootstrapped (robust) standard errors are clustered on the country and service station chain level. Error bars represent 95% confidence intervals.

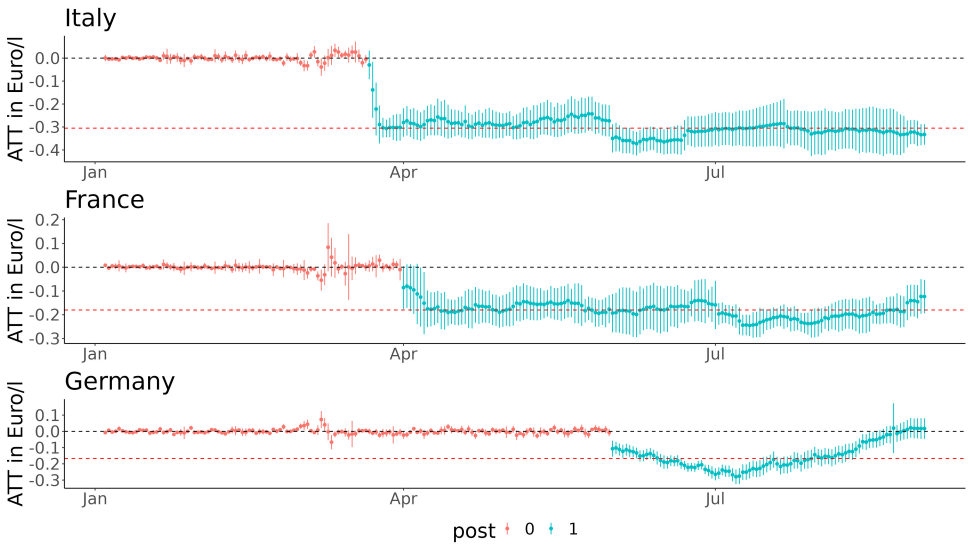


Figure 7.12: Event Study of prices with diesel and no covariates. Bootstrapped (robust) standard errors are clustered on the country and service station chain level. Error bars represent 95% confidence intervals.

Declaration of Authorship

I hereby declare that I completed the papers submitted and listed hereafter independently and only with those forms of support mentioned in the relevant paper. When working with the authors listed, I contributed no less than a proportional share of the work. In the analysis that I have conducted and to which I refer in the papers, I have followed the principles of good academic practice, as stated in the Statute of Justus Liebig University Giessen for ensuring good scientific practice.

Ort, Datum	Unterschrift