Technische Universität München
TUM School of Management

Fakery on Digital Platforms

A Review and Two Empirical Studies of its Extent, Implications, and Countermeasures

Michaela Lindenmayr

Vollständiger Abdruck der von der TUM School of Management der Technischen Universität München zur Erlangung einer Doktorin der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.) genehmigten Dissertation.

Vorsitz:	Prof. Dr. Martin Meißner
Prüfende der Dissertation:	1. Prof. Dr. Jens Förderer
	2. Prof. Dr. Hanna Hottenrott

Die Dissertation wurde am 10.07.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Management am 15.08.2024 angenommen.

Acknowledgements

After more than three years of working on my dissertation, I would like to express my deepest gratitude to several people who supported me throughout the process of completing my doctoral studies.

First and foremost, I would like to thank my supervisor, Prof. Dr. Jens Foerderer, for his consistent guidance throughout the entire journey. His expertise and feedback were invaluable in conceptualizing research ideas, conducting analyses, and deriving contributions conforming to the highest standards of scientific practice. The dissertation would not have been possible without his encouragement and mentorship throughout all stages of the doctoral journey. Thank you for your trust and your belief in my skills that have made this journey a success!

I would also like to extend my gratitude to the members of my dissertation committee, Prof. Dr. Hanna Hottenrott and Prof. Dr. Martin Meißner, for making the last stage of my doctoral studies a great experience.

I am very grateful for all the inspiring colleagues at the Center for Digital Transformation and the TUM Campus Heilbronn in general. A special thanks goes to my close colleagues Alexander, Johannes and Tobias – I could not have imagined a better team to be part of and I will miss our insightful research discussions, joint conference trips, and team lunches. Overall, the diverse environment on campus allowed me to broaden my mind and build personal relationships that will last a lifetime.

On a personal note, I want to express my heartful thankfulness to my family and friends. Thank you for celebrating achievements and milestones with me, but also for believing in me during challenging times. Your mental support was indispensable!

Abstract

Digital platforms give access to a wide variety of content and improve its exchange. However, the ease of content dissemination and lack of control mechanisms generate new dimensions of fakery. Facing fakery on digital platforms as a major problem in information systems research, this dissertation aims to understand why and to what extent it exists, what its implications are, and which countermeasures are effective. It consists of a review of the literature and two follow-up empirical studies driven by public debates around platforms' fake accounts and health misinformation in combination with a lack of scientific evidence. First, a scoping review of the FT50 journals (2016), the journals in the Senior Scholars' Basket of Eight (2011), and the journal Business & Information Systems Engineering provides an overarching understanding of the state-of-the-art research on fakery on digital platforms along the sender-receiver framework. It introduces the research phenomenon and builds the foundation for the empirical studies that follow. Second, fake follower use presents a publicly relevant but research-wise underrepresented topic. A panel data study on Twitter exploits the purge of tens of millions of fake accounts in 2018 to assess the extent of fake follower use by firms, and shareholders' reactions to its revelation. Third, countermeasures for misinformation commonly focus on pruning inaccurate content, with a lack of research on measures that promote credible content. A difference-in-difference-in-differences study on YouTube investigates the implications of a novel intervention, namely the promotion of credible content, as part of the YouTube Authoritative Health Information program introduced in Germany in 2023. This dissertation makes three main contributions. First, it shows rising research interest in platform fakery and identifies open questions. Second, it outlines that fake followers present a minor problem with limited prevalence and impact on shareholder value. Third, it presents the promotion of credible content as a complementary approach to combat fakery, on the account of improving its discovery. Overall, this dissertation informs platform providers, users, and policymakers about the extent, implications, and countermeasures of fakery and helps develop future actions.

Table of contents

List of tables

List of figures

List of abbreviations

AFINN	lexicon-based sentiment classifier named after Finn Årup Nielsen
AMEX	American Stock Exchange
API	application programming interface
BISE	Business & Information Systems Engineering
Bn	billion
CAR	cumulative abnormal returns
CAS	cumulative abnormal sentiment
CEM	coarsened exact matching
CEO	chief executive officer
cont.	continued
COVID-19	coronavirus disease 2019
CRSP	Center for Research in Security Prices
DID	difference-in-differences
DDD	difference-in-difference-in-differences
e.g.	for example (Latin: exempli gratia)
EBIT	earnings before interest and taxes
etc.	and other similar things (Latin: et cetera)
EU	European Union
FF3FM	Fama-French Three Factor model
FFPM	Fama-French Plus Momentum model
FT50	50 journals in Financial Times Research Rank
FTC	Federal Trade Commission
HHI	Herfindahl-Hirschman Index
HML	high minus low
i.e.	that is (Latin: id est)
IS	information systems
k2k	option for coarsened exact matching to ensure the same number of treated and control units
MOM	momentum factor
Max	maximum
Min	minimum
Mn	million
NAICS	North American Industry Classification System
NASDAQ	National Association of Securities Dealers Automated Quotations Stock Market

NYSE	New York Stock Exchange
OLS	ordinary least squares
SMB	small minus big
SD	standard deviation
S&P	Standard & Poor’s
SUTVA	Stable Unit Treatment Value Assumption
TV	television
UK	United Kingdom
U.S.	United States of America
VADER	Valence Aware Dictionary and sEntiment Reasoner
vs.	versus / in contrast to
WRDS	Wharton Research Data Services

1. Introduction

1.1. Research Motivation

Digital platforms, particularly social media platforms, have substantially changed the information exchange (Parker et al. 2016, Shapiro et al. 1998). Digital platforms connect billions of users and enable them to actively engage in discussions, ultimately fostering access to diverse opinions and open exchange (e.g., Asimovic et al. 2021, Chen et al. 2019a, Howard et al. 2011). The value of these platforms, therefore, derives from the network of users that contribute to information dissemination (Evans and Schmalensee 2010, Shapiro et al. 1998).

However, research in information systems indicates that digital platforms enhance both the incentives and abilities to engage in adverse behavior (e.g., Boudreau 2010, Lewandowsky et al. 2012, Parker and Van Alstyne 2018, Suler 2004, Tiwana et al. 2010, Xu et al. 2012). First, online interactions are anonymous and can lead to disinhibition when interacting with others, making them likely to spread unverified and polarizing information (Cohn et al. 2022, Rockmann and Northcraft 2008, Suler 2004, Xu et al. 2012). Also, the relevance of and dependence on online appearance for firms and individuals incentivize them to engage in platform fakery, in particular, if competitors do so and if network effects exist (Dellarocas 2006a, Jin et al. 2023, Katz and Shapiro 1985, Mayzlin et al. 2014). Second, the online setting enhances the ability to engage in fakery. For example, conventional gate-keeping mechanisms or access restrictions are commonly not in place on digital platforms, and monitoring every piece of content is impossible considering the importance of large user networks (Candogan and Drakopoulos 2020, Chen et al. 2023, Germano et al. 2022, Lewandowsky et al. 2012).

In addition to the intentional engagement in fakery, users and platforms unintentionally disseminate it (e.g., Calderon et al. 2023, Serra-Garcia and Gneezy 2021, Stanley et al. 2022). As individuals are oftentimes unable to distinguish true from false information, they are likely to believe fakery, consequently disseminating the information further (Chen et al. 2021, Hamby et al. 2020, Serra-Garcia and Gneezy 2021, Stanley et al. 2022). Particularly because fakery often appears as novel and useful, it even spreads further than accurate information (Hamby et al. 2020, London Jr et al. 2022, Vosoughi et al. 2018). Reinforced by algorithms that provide content recommendations and further spread contents across the network, the existence of various forms of fakery leads to misinformed users and, ultimately, undesirable offline behavior (Calderon et al. 2023, Pariser 2011, Ross et al. 2019, Vosoughi et al. 2018). Examples of fakery are manifold, including fake news, fake followers, or social bots during elections on Twitter (Allcott and

Gentzkow 2017, Bessi and Ferrara 2016, Silva and Proksch 2021), fake product reviews on Amazon or Yelp (He et al. 2022b, Luca and Zervas 2016), or fake ads (Wiles et al. 2010).

Recent examples provide evidence for the challenge of fakery on digital platforms. Considering global health concerns, De Beaumont (2021) show that more than 20% of Americans relied on social media sources to obtain information on the COVID-19 pandemic, making them susceptible to misinformation. For example, 41% out of those people believed that the vaccine would make them infertile, and 58% were convinced that it was used as a means to destroy world economies—in both cases, these numbers are more than 15 percentage points higher than the average replies relying on individuals that obtained information from other news sources such as doctors or newspapers. At the same time, the extent of fake accounts on various platforms is substantial—in 2023, Facebook and TikTok both removed around 2.5 billion fake accounts, and even LinkedIn removed more than 120 million accounts (Facebook 2024, LinkedIn 2024, TikTok 2024). Also, around 50% of Americans indicate each year that they have seen fake reviews on Amazon (BrightLocal 2024). These examples show the harm of fakery for platform businesses and society and highlight the need for a better understanding of the overarching topic of fakery on digital platforms to overcome the adverse effects.

Existing research on fakery on digital platforms in information systems and adjacent fields has mainly focused on three streams, i.e., the challenges for platform governance, the implications for social media engagement, and the need for regulatory intervention.

First, existing research investigates quality control in the context of fakery as a new dimension of platform management and governance (e.g., Boudreau 2010, Foerderer 2020, Katz and Shapiro 1985). A platform benefits from a large user base that contributes to the platform and attracts further users (Gandal et al. 2000, Katz and Shapiro 1985). In particular, enhancing supply from complementors enhances the value of a platform (Engert et al. 2023, Foerderer 2020, Foerderer et al. 2018, Lindenmayr and Foerderer 2024, Scholten and Scholten 2012). However, integrating more contributors imposes new challenges for quality control (Boudreau 2010, Candogan and Drakopoulos 2020, Huang et al. 2022b, Lindenmayr and Foerderer 2022). If content cannot be monitored sufficiently, contributors can engage in fakery that disseminates broadly and quickly in social networks (Vosoughi et al. 2018). This shows the trade-off to balance engagement and quality control as a main challenge in the context of platform management and governance.

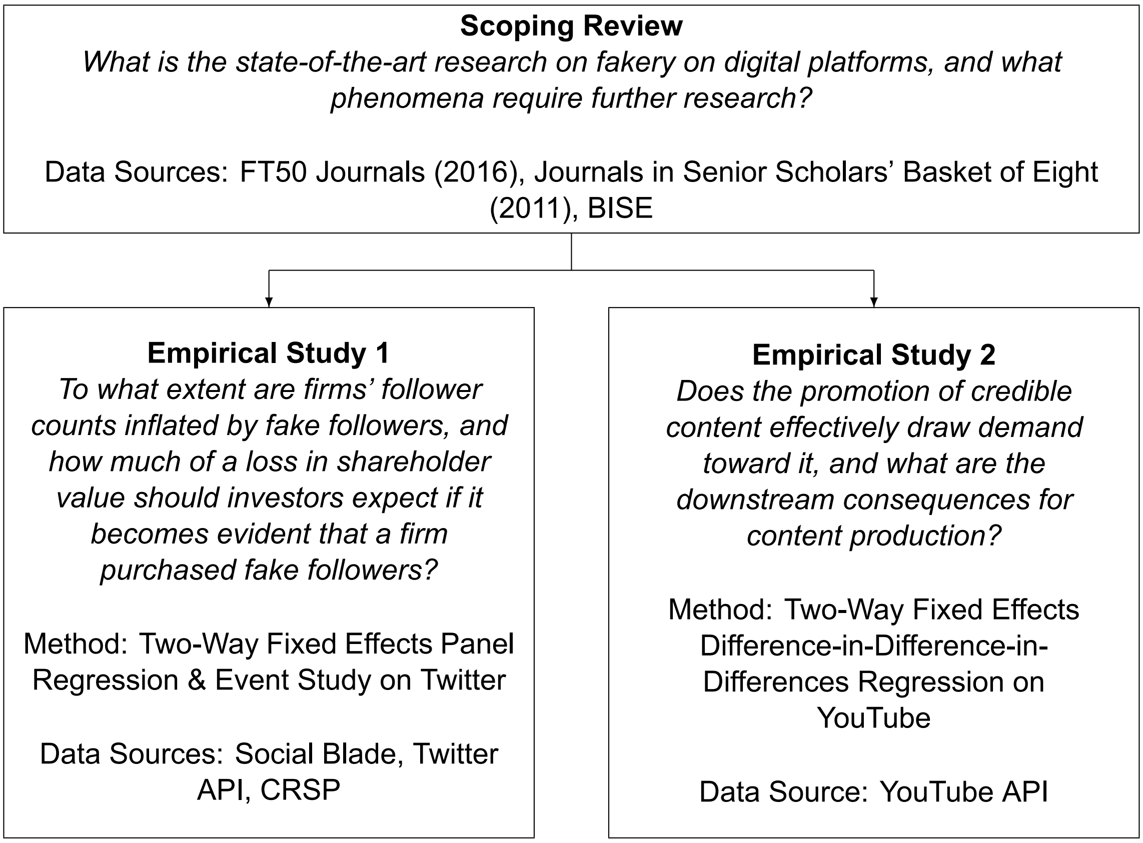
Second, existing literature on social media engagement elaborates on the role of fakery in enhancing supply and demand (e.g., Chen et al. 2015a, 2018, Lee et al. 2018a, Luo et al. 2013, Yan and Tan 2014). Considering the demand perspective, existing studies investigate the incentives of individuals for using social media, e.g., information seeking, entertainment, or community (Khan 2017, Yan and Tan 2014). The engagement with these contents is driven by the influence of peers, but also specific content characteristics (De Oliveira Santini et al. 2020, Lee et al. 2018a, Li and Xie 2020, Mallipeddi et al. 2021, Mousavi and Gu 2019, Stieglitz and Dang-Xuan 2013, Tellis et al. 2019). Ultimately, social media plays a crucial role in customer engagement and brand promotion (Chen et al. 2015a, Hennig-Thurau et al. 2015, Kaplan and Haenlein 2012, Kumar et al. 2016, Wang et al. 2021b) but also in shaping investment decisions (Deng et al. 2018, Lacka et al. 2022, Nofer and Hinz 2015). Considering the effects of fakery, studies show distortion for decision-making for both consumers and investors in line with the strong impact of social media (Clarke et al. 2021, Lappas et al. 2016, Ullah et al. 2014, Vosoughi et al. 2018). At the same time, studies show that content contribution is driven by recognition, influence, or monetary incentives (Chen et al. 2018, Liu and Feng 2021, Sun et al. 2017, Valsesia et al. 2020, Wei et al. 2021). These factors can make the engagement in fakery a rational decision (Chen et al. 2022, Jin et al. 2023, Luca and Zervas 2016, Mayzlin et al. 2014, Qiao and Rui 2023). Overall, this stream of literature highlights that the spread of fakery is due to plausible reasons, leading to challenges for overcoming it.

Third, studies elaborate on the societal impact of platforms and the need for regulatory intervention to control fakery (e.g., Barone and Miniard 1999, Cantarella et al. 2023, Cho et al. 2011, Shi et al. 2022). Online platforms are subject to internal quality control to define the boundary conditions for participation on the platform and to manage the interactions (Boudreau and Hagiu 2009, Huang et al. 2022b, Lindenmayr and Foerderer 2022). However, there is evidence from the literature that platforms might tolerate fakery to benefit from higher engagement and to reduce the efforts associated with content monitoring (Candogan and Drakopoulos 2020, Chatain 2022, Vosoughi et al. 2018). In particular, there is increasing public pressure to reduce the impact of fakery from the viewpoint of policymakers (e.g., Chatain 2022, Colomina et al. 2021, Federal Trade Commission 2023). Therefore, the harm from fakery and the different ways of counteracting, also from a policy perspective, are a major concern in the literature.

1.2. Structure of the Dissertation

To approach the topic of fakery on digital platforms, a scoping review provides an understanding of the state-of-the-art literature after conceptualizing the main research problem. Building upon this state of knowledge, two research gaps are identified that are of societal interest considering current debates. The follow-up studies empirically assess the extent, implications, and countermeasures of fakery. In particular, the first empirical study explores the extent of fake follower use and the implications for the stock market, while the second empirical study looks into the effectiveness of a novel countermeasure to misinformation by promoting credible content. Figure 1.1 summarizes the studies in this dissertation that have the overarching goal of advancing the understanding of the extent, implications, and countermeasures of fakery.

Figure 1.1: Summary of Studies in this Dissertation
Note: The figure summarizes the studies in this dissertation and their main characteristics in terms of content and methodology. The scoping review provides a holistic overview of the state-of-the-art. The two empirical studies build on the literature review by investigating two research gaps using observational data from real-world platforms.



Scoping Review
What is the state-of-the-art research on fakery on digital platforms, and what phenomena require further research?

Data Sources: FT50 Journals (2016), Journals in Senior Scholars' Basket of Eight (2011), BISE

Empirical Study 1
To what extent are firms' follower counts inflated by fake followers, and how much of a loss in shareholder value should investors expect if it becomes evident that a firm purchased fake followers?

Method: Two-Way Fixed Effects Panel Regression & Event Study on Twitter

Data Sources: Social Blade, Twitter API, CRSP

Empirical Study 2
Does the promotion of credible content effectively draw demand toward it, and what are the downstream consequences for content production?

Method: Two-Way Fixed Effects Difference-in-Difference-in-Differences Regression on YouTube

Data Source: YouTube API

The scoping review (Chapter 2) gives a holistic overview of the literature on fakery on digital platforms. The dimensions in which fakery emerges are varied and the implications and consequences are broad. For an understanding of the body of literature in information systems and adjacent fields, the scoping review conceptualizes the research problem as the transmission of fakery in a sender-receiver framework on a two sided digital platform (Parker et al. 2016, Shannon and Weaver 1949). Then, it aggregates findings from prior studies and identifies open research questions (Paré et al. 2015).

Motivated by current public debates, the empirical studies further investigate identified research gaps from an empirical perspective. The studies rely on observational data from real-world platforms to understand the implications of their interventions in the context of fakery. Despite the limitations of low control over the intervention, the research design employed in this dissertation enables the study of fakery in its natural environment. Doing so provides applicable insights into complex real-world platforms. In particular, the design allows the evaluation of the strategies underlying the interventions and enables the analysis of large data sets to retrieve multifaceted findings.

The first empirical study (Chapter 3) investigates the use of fake followers and investors' reactions. Fake followers have obtained attention from the public, including evidence of a firm selling fake accounts and the goal of the Federal Trade Commission to counter fake indicators on social media, including fake followers (Confessore et al. 2018, Federal Trade Commission 2023). However, from the information systems perspective, it is unclear to what extent fake followers are used by firms, what factors moderate their use, and which risks are associated with them. Relying on a set of firms that operate a Twitter account, the study relies on a large-scale intervention of the platform in removing fake accounts in 2018 (Confessore and Dance 2018). Assessing the decline in followers in response to the platform intervention allows to indirectly draw causal conclusions about their existence (Wooldridge 2019). Conducting an event study in line with the standard practice in the finance literature (Brown and Warner 1985, MacKinlay 1997, Sorescu et al. 2017), the study further investigates the effects on the stock market once the fraudulent behavior is revealed. This helps in showing the distortive power of fake accounts toward shareholders.

The second empirical study (Chapter 4) evaluates a novel countermeasure for combating misinformation by promoting credible content, i.e., content produced by credible sources. Public debates have emerged from concerns during the COVID-19 pandemic but also go beyond in terms of fake stories that affect the societal discussions (De Beaumont 2021, Duffy 2022, Naeem et al. 2021). Relying on YouTube as the empirical setting, the study aims to assess the effectiveness of an intervention that promotes credible content in terms of the effects on demand for and the supply of content. The study uses a program introduced by YouTube in Germany in 2023 with the goal of improving health information on the platform. In particular, authoritative health channels such as hospitals, governmental authorities, or doctors are featured in the search results and obtain a quality label for their videos. This unique setting can be exploited as a quasi-experiment that compares German to French health channels to obtain causal estimates for the effects of the program (Angrist and Pischke 2008).

1.3. Contribution to the Research Interests of the Professorship

This dissertation contributes to the research conducted at the Professorship for Innovation & Digitalization led by Prof. Dr. Jens Foerderer since 2019. It integrates the research interests in digital markets and analytics to form a contribution to the business and information systems scholarship.

The dissertation extends past and ongoing research of the professorship that investigates platform governance from the perspective of complementary innovation and quality control. Earlier studies conducted by the professorship in this field evaluate how platform decisions shape complementary innovation (Foerderer et al. 2018) and show how platform quality is enhanced from knowledge spillovers among complementors (Foerderer 2020). As the integration of complementors can lead to a reduction in quality control, the research at the professorship also delves into various governance mechanisms (Halckenhaeusser et al. 2020) and the assessment of control systems to keep high-quality standards (Lindenmayr and Foerderer 2022). The studies in this dissertation add a new perspective to this research focus by exploring platform governance in the presence of various forms of fakery. While earlier studies touched upon the challenge of low-quality contributions, this series of studies is the first to explicitly address fakery to extend other studies of the professorship. In the context of this dissertation, users contribute to content platforms, thereby actively creating or being subject to adverse behavior and engaging in the spread of fakery. This poses a substantial challenge to governance – the dissertation comprises studies that investigate the extent and implications of such behavior but also potential countermeasures.

Newer studies of the professorship explicitly investigate the challenges associated with online appearance and platform management. A contribution in a book series gives an overview of the challenges of digital platforms in terms of privacy concerns, harmful content, but also competitive dynamics that lead to discrepancies (Lindenmayr et al. 2022). In particular, existing studies of the research team assess how platform incentives or a lack thereof affect contributions in mobile app markets (Foerderer et al. 2021, Kircher and Foerderer 2024). In that context, fakery presents a new dimension to the challenges explored in earlier studies. This perspective helps in getting an extended view of the problems that the platform might encounter, including the incentives to engage in fakery and the risks associated with it. Also, the study of Foerderer and Schuetz (2022) is closely related to fakery in a way that it unveils strategic timing of data breach announcements to mitigate the negative effect on the stock market. Even though not directly related to platforms, this study shows how consumers are potentially deceived by the information they are exposed to – something that is highly relevant in the research on fakery on digital platforms.

In terms of the methodological stance, the dissertation integrates state-of-the-art techniques from data analytics that are central to the professorship. By collecting large data sets from content platforms via APIs or web scraping techniques, the dissertation leverages the potential of engaging with valuable data to derive meaningful conclusions. Statistical analyses in line with the current standards ensure high-quality findings of the studies.

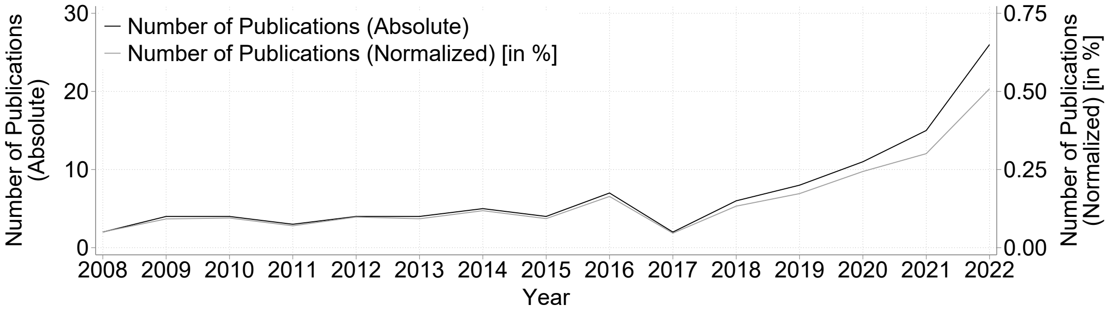
2. Review of the Literature and Open Questions

2.1. Introduction

Over the past years, research interest in fakery experienced a massive growth; yet, at the same time, this growth poses a challenge for the academic conversation in that it loses track of the state-of-the-art. Figure 2.1 shows that the number of publications in the FT50 journals, the journals in the Senior Scholars' Basket of Eight, and the journal Business & Information Systems Engineering – absolute and normalized by journal space – have increased substantially in recent years. While 53 papers on fakery were published between 2008 and 2019, almost the same amount of papers, 52, were published only between 2020 and 2022. From 2008 to 2022, there was an increase in publications by 1,300%. Although attempts have been made to organize specific substreams of research such as fake news consumption (e.g., Baptista and Gradim 2020, Pennycook and Rand 2021) or fake news during the pandemic (Awan et al. 2022), there is no overarching categorized understanding of the state-of-the-art.

Figure 2.1: Distribution of Publications over Time

Note: The figure shows the distribution of publications on fakery over time, considering the last 15 years, in the FT50 journals (2016), the journals in the Senior Scholars' Basket of Eight (2011), and the journal Business & Information Systems Engineering. The left y-axis (black line) shows the absolute number of publications. The right y-axis (grey line) denotes the normalized number of publications by journal space.



Against this backdrop, the goals of this scoping review are twofold: First, it seeks to organize the state-of-the-art around fakery on digital platforms. Second, it aims to identify areas in the academic conversation that require further investigation, thereby deriving avenues for future research.

The study departs from conceptualizing fakery along two theoretical lines. First, it conceptualizes fakery using a sender-receiver framework (Shannon and Weaver 1949). In this framework, fakery represents a noisy signal transmitted by a sender to deliberately mislead a receiver. Second, it embeds the framework into a multi-sided platform setting. This allows the structure of the literature to provide a holistic overview of how fakery emerges, how it expresses for senders and receivers, and what downstream consequences and countermeasures are. Fakery on digital platforms is a core topic in information systems but has also been of particular interest in adjacent fields, including marketing and strategy. Therefore, the review includes the FT50 journals, the journals in the Senior Scholars' Basket of Eight, and the journal Business & Information Systems Engineering as the basket of interest (Paré et al. 2015). Using a search string of different synonyms of fakery, the Web of Science Core Collection provides the basis for relevant papers to build a framework that presents the current knowledge on platform fakery.

This chapter is structured as follows. First, Section 2.2 conceptualizes fakery along the sender-receiver framework and derives guiding questions for the inquiry. Second, Section 2.3 describes the methodology and procedure of the literature review. Third, in alignment with the conceptual framework, Section 2.4 summarizes and reflects on the research findings.

2.2. Conceptual Framework

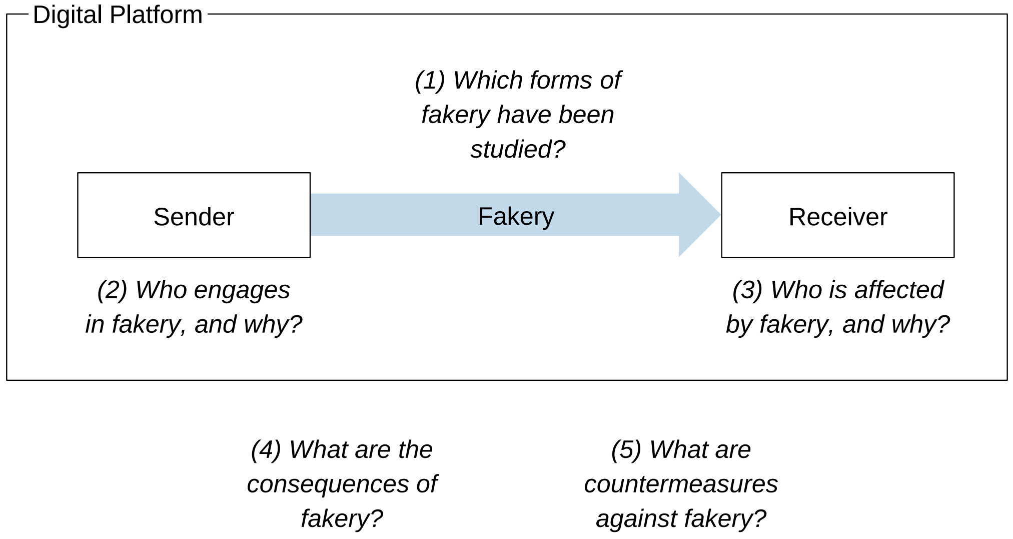
The dissertation conceptualizes fakery on digital platforms along two theoretical lines to structure the inquiry as illustrated in Figure 2.2.

The first premise is to conceptualize fakery in a sender-receiver framework following Shannon and Weaver (1949). In this framework, information is transmitted from a sender to a receiver via a certain medium. In this dissertation, this medium is a digital platform. Based on this conceptualization, fakery represents what has been described as a *noisy signal* (Koohikamali and Sidorova 2017, Li et al. 2022, Shannon and Weaver 1949). Noisy signals are defined as pieces of information that are "intentionally and verifiably false and could mislead" (Allcott and Gentzkow 2017, p. 213), mostly in a "deliberate attempt" (DePaulo et al. 2003, p. 74). Both the sender and receiver of fakery can take different forms, such as representing an individual, a group, an organization, or further. The sender-receiver conceptualization is particularly adequate and helpful for an inquiry into fakery because it allows the distinction between the sender of fakery and the receiver of a fakery attempt, as well as the information.

The second premise of the framework is to embed the sender-receiver framework into a two-sided platform setting. Based on the literature, a digital platform "uses technology to connect people, organizations, and resources in an interactive ecosystem in which [...] value can be created and exchanged." (Parker et al. 2016, p. 3). The platform acts as

Figure 2.2: Conceptual Framework of Fakery and Guiding Questions for the Literature Review

Note: The figure illustrates the conceptualization of fakery along a sender-receiver framework. Questions (1) to (5) are derived from this conceptualization and guide the literature review.


Digital Platform
(1) Which forms of fakery have been studied?
Sender
Fakery
Receiver
(2) Who engages in fakery, and why?
(3) Who is affected by fakery, and why?
(4) What are the consequences of fakery?
(5) What are countermeasures against fakery?

an intermediary that brings different platform participants together to enhance their exchange. Examples of digital platforms include marketplace platforms such as Amazon, social media platforms like YouTube or Twitter, and software development platforms such as Apple iOS.

The following research questions emerge around the framework: (1) Which forms of fakery have been studied?, (2) Who engages in fakery, and why?, (3) Who is affected by fakery, and why?, (4) What are the consequences of fakery?, and (5) What are countermeasures against fakery? These questions give a holistic overview and categorize the literature accordingly.

2.3. Methodology of the Literature Review

2.3.1. Search and Analysis

The scoping literature review takes a concept-centric focus following Webster and Watson (2002) and Paré et al. (2015). This approach is well-suited due to the (1) novel nature of the research phenomenon, (2) the large number of publications, and (3) the goal to achieve a holistic understanding. This method helps structure the existing body of knowledge, identify contradictions, and uncover missing evidence for future studies.

Regarding publication types, the review exclusively relies on articles published in scientific journals. To account for the interdisciplinary nature of the topic, it uses all journals in the FT50 journal list (Ormans 2016) and complements it with the journals in the Senior Scholars' Basket of Eight (Association for Information Systems 2011) and the journal Business & Information Systems Engineering. Even though the topic of fakery on digital platforms emerges from the field of information systems, merely focusing on related journals might omit causes or consequences extending beyond information systems or arising from offline interactions. Fakery can, for instance, affect investors and consumers (Clarke et al. 2021, He et al. 2022b, Jia et al. 2020, Lappas et al. 2016, Rao 2022), making it relevant to disciplines such as finance and marketing. The FT50 list provides a reference list to retrieve high-quality journals for standard disciplines in the business field. At the same time, the FT50 does not capture the breadth of the information systems field as it includes only three information systems journals. This is why the journals in the Senior Scholars' Basket of Eight and the journal Business & Information Systems Engineering complement the list of journals.

Regarding databases, the Web of Science Core Collection is used to extract all relevant publications from the journals. The Web of Science Core Collection offers a standardized way to obtain relevant papers from many journals. As all journals in the basket of interest are part of the database, the process of obtaining the articles is comparable among all journals.

Regarding the search string and the search fields, the keywords are extracted from the research question by identifying the main concepts and obtaining synonyms or related terms (Kitchenham and Charters 2007, Xiao and Watson 2019). To start, synonyms for fakery are obtained from Thesaurus.com and Merriam-Webster.com. The most relevant synonyms are combined with truncation symbols to consider different writings according to the various parts of speech they can take. Eventually, the decision on the search string also involves a decision on the number of results. In this regard, the broader the search string, the more articles are returned, and therefore, the likelihood of capturing relevant work increases. However, at the same time, this procedure requires a greater effort of individually assessing the articles. The decision on this search string specifically, seeks to use a relatively broad scope with the consequences of yielding many articles. This way, it delivers all relevant papers, which allows to derive meaningful conclusions from the review. The search string is formulated as follows: Fake OR False OR Inaccurate OR Incorrect OR Mislead* OR Spam OR Artificial OR Decept* OR Fraud* OR Manipulat* OR Fictit* OR Suspici* OR Fabricat*. According to the Web of Science Core Collection, the searched fields are title, abstract, and keywords.

Following Webster and Watson (2002) and to reduce the risk of a too-narrow search string, the search strategy is complemented by a backward- and forward search to complete the set of relevant literature. This strategy helps obtain further relevant papers that the main keywords in the string might not directly find.

Based on evaluation criteria following Vom Brocke et al. (2009), the filtering ensures that the obtained literature investigates fakery based on the definition in the framework. This is done based on the assessment of the titles, abstracts, and full texts. Papers irrelevant to the research question are removed from further investigation. This includes papers that use the keywords in a different semantic or do not refer to the concept of fakery at all (i.e., that represent false positives). Examples are, for instance, papers on voting manipulation, academic fraud, or purely algorithmic biases. To exclude subjectivity in the process of selecting the papers following, three independent researchers read and coded the papers (Kitchenham and Charters 2007). The initial agreement rate in the selection of papers was 96.51%. An explicit agreement on the relevance of the papers resulted in a definite inclusion vs. exclusion. Coders discussed their choices for all remaining papers and agreed on a solution.

Overall, the final set consists of 142 articles. This set of articles is retrieved as follows: 4,215 articles are obtained from the initial search, which is reduced to a relevant set of 129 articles after filtering by relevance. 9 articles are added from the backward search, and four further articles from the forward search. The final list of papers included in this literature review is displayed in Appendix Table A2.1.

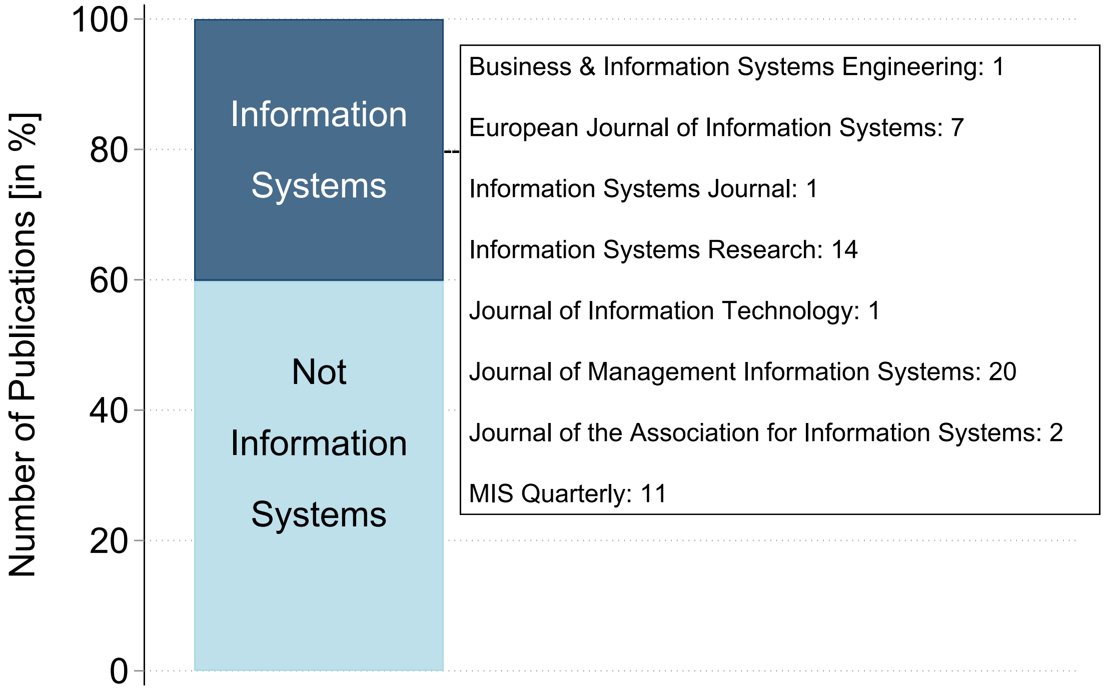
2.3.2. Characteristics of the Obtained Literature

Before discussing the results, the review gives an overview of the characteristics of the literature across their publication outlets and the methods used.

Figure 2.3 differentiates between papers in information systems journals and non-information systems journals. Among all 142 papers in the basket of interest, 40% of the papers have been published in information systems journals and 60% in other disciplines. However, only considering more recent publications within the last 15 years (since 2008), 55 relevant papers have been published in information systems and 64 in all other disciplines. This shows that, in particular, the more recent publications with a clear focus on the digital world mainly stem from information systems despite the relevance for other areas.

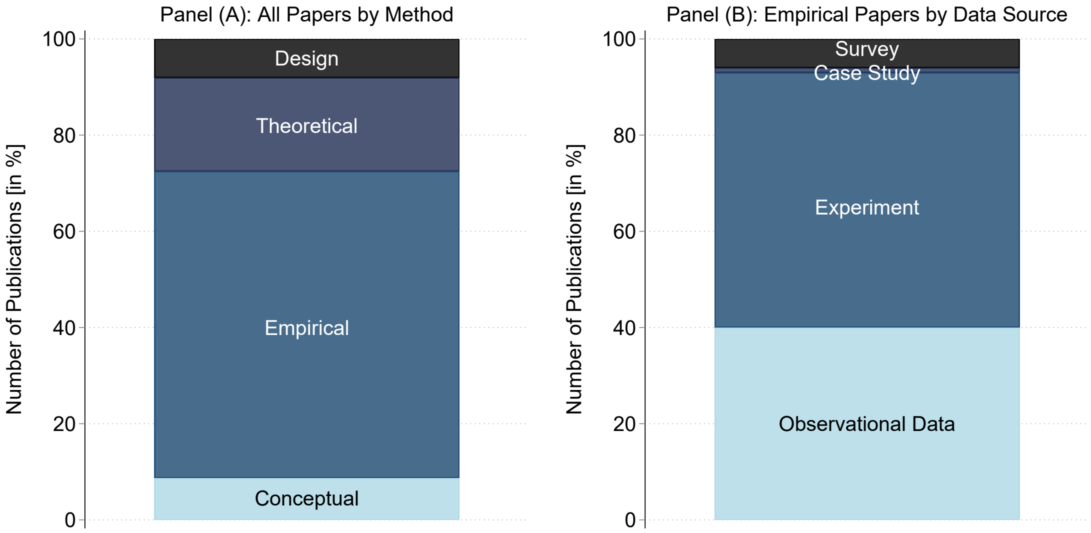
Figure 2.4 provides an overview of the obtained literature regarding the methods used. From the journals of the basket, empirical research is dominant. Design science studies

Figure 2.3: Distribution of Publications by Publication Outlets
Note: The figure describes the distribution of papers by information systems vs. non-information systems journals in the FT50 journals (2016), the journals in the Senior Scholars' Basket of Eight (2011), and the journal Business & Information Systems Engineering. For the information systems journals, the figure provides the split by journals.



are used to understand approaches to combat fakery, i.e., the design of detection systems or algorithms. Regarding the data sources used by the empirical papers, observational data is dominant. It mostly comes from a relatively small set of platforms, namely Twitter (Jia et al. 2020, London Jr et al. 2022), Sina Weibo (Ng et al. 2021, Wang et al. 2021a), Facebook (Cantarella et al. 2023, Harrison 2018), Seeking Alpha (Clarke et al. 2021, Kogan et al. 2023), Yelp (Luca and Zervas 2016, Siering and Janze 2019), Expedia (Mayzlin et al. 2014), TripAdvisor (Lappas et al. 2016), and Amazon (He et al. 2022b, Kokkodis et al. 2022). Also, experimental studies mimic popular platforms (e.g., Moravec et al. 2019, 2020, 2022, Pennycook et al. 2020) and almost all of these studies are conducted within, not across, platforms.

Figure 2.4: Distribution of Publications by Method
Note: The figure describes the publications included in the sample along (A) the research methods and (B) the data sources for empirical studies. For (A), conceptual studies include literature reviews, taxonomies, and frameworks developed from existing research. Empirical studies use various forms of data. Theoretical studies develop a formal mathematical model. Design studies refer to systems design, e.g., machine learning classifiers that learn from certain data and predict output variables. For (B), observational data is collected in a natural environment to passively observe relationships. For experimental data, variables are manipulated to create a treatment and control group to measure the effects. Case studies are built around one or a few research objects to understand certain behaviors in depth. Survey data is obtained via questionnaires to get self-reported data on certain characteristics. Coding is depicted in Table A2.2.



2.4. Results: State-of-the-Art and Questions for Future Research

2.4.1. Which Forms of Fakery Have Been Studied?

State-of-the-Art

Existing research mainly clusters around specific forms of fakery, namely social bots, fake reviews, fake news, and fake ads.

First, the research studies *social bots*, in terms of user accounts that are created with false or misleading information about the identity of the user and with the goal to pretend popularity of people or contents (e.g., Benjamin and Raghu 2023, Cho et al. 2011, Ross et al. 2019). They imitate human online behavior and provide a misleading picture of the online conversation (Ferrara et al. 2016).

The second type of fakery studied are *fake reviews*, defined as “non-authentic online reviews” typically posted on behalf of third parties (Hu et al. 2012, p. 674). Fake reviews have been studied on various platforms, including Yelp and Expedia (e.g., Anderson and Simester 2014, Luca and Zervas 2016, Mayzlin et al. 2014).

The third type is *fake news* (e.g., Allcott and Gentzkow 2017, Wardle and Derakhshan 2017). It includes misinformation that results from “an honest mistake” (Hernon 1995, p. 134), and disinformation, which is “a deliberate attempt to deceive or mislead” (Hernon 1995, p. 134).

The final type are *fake ads* (e.g., Gardner 1975, Nikitkov and Bay 2008, Park et al. 2023, Sher 2011, Xiao and Benbasat 2011). Fake ads are advertisements that lead to misconceptions driven by false claims, omission of important facts, or misrepresentation (Sher 2011).

Open Questions

In contrast to fake accounts that manipulate online discussions qualitatively, so far fake followers that manipulate quantitative metrics have not been studied. Fake followers inflate users’ follower counts and make them perceived as more popular (Caruccio et al. 2018). It is unclear whether and to what extent fake followers distort users’ online followerships, but also how dangerous they are in manipulating other users. As this is a major open question, it will be covered in the empirical study in Chapter 3 of this dissertation.

Recent technological advancements pose two further questions regarding the forms of fakery. First, the advancements surrounding generative artificial intelligence lead to the question of how the different forms of fakery evolve. For example, so-called *deepfakes* appear like authentic media and are used for manipulation. The use of artificial intelligence creates such videos, photos, or audios, that are extremely difficult to detect (e.g., Khan et al. 2022b, Mohammed and Salam 2021, Vasist and Krishnan 2022). Second, the emerging concern about identity theft concerns using bots to engage in public discussions. For example, research deals with phenomena such as phishing or fake websites (Abbasi et al. 2010, Herzberg and Jbara 2008, Wang et al. 2017) that can be used to steal personal data. However, the question as to what extent individuals use a fake identity to engage in fakery remains to be answered. In particular, fake accounts using others’ identities have not been considered in the past and require closer investigation. The likelihood of detecting such individuals declines when individuals are perceived as natural and contents are believed to stem from a credible source (Cheung et al. 2012, Stanley et al. 2022), making it essential to understand whether such behavior exists, how other users react to it, and what platform firms can do to combat.

2.4.2. Who Engages in Fakery, and Why?

State-of-the-Art

Substantial research has documented that a wide variety of market actors engage in fakery, including firms (Jin et al. 2023, Lee et al. 2018b, Luca and Zervas 2016, Mayzlin 2006), incentivized consumers or third parties (Chen et al. 2022, He et al. 2022b, Qiao and Rui 2023), politicians (Cantarella et al. 2023), and public organizations (Cho et al. 2011). Different motivations have been observed to cause market actors to engage in fakery: profit motives, competitive pressure, low visibility, low reputation, and sender perception. Table 2.1 summarizes the main results.

Profit Motives: One finding is that senders engage in fakery to obtain a financial benefit (Jin et al. 2023, Keppo et al. 2022, Khan et al. 2022a, Mullainathan and Shleifer 2005). Economic and power-related motivations drive engagement in fakery on social media platforms (George et al. 2021). First, firms can increase their payoffs (Keppo et al. 2022) or boost their search ranking by fake orders, also known as brushing (Jin et al. 2023). Second, other stakeholders, e.g., newspapers, are incentivized to engage in fakery to align with readers' preferences (Mullainathan and Shleifer 2005). Third, direct payoffs for third parties make these willing to engage in fakery, e.g., from Facebook groups to seek fake reviewers (He et al. 2022b), incentivized Vine-reviews on Amazon (Qiao and Rui 2023), or conditional-rebate strategies after writing a positive review (Chen et al. 2022). In contrast to these incentives, Mostagir and Siderius (2023b) highlight that bribes are only effective if consumers are unable to recognize bribed reviewers, and Anderson and Simester (2014) mention that it is highly likely that reviews without a purchase are written by loyal customers who want to give feedback.

Competitive Pressure: One factor that has repeatedly been confirmed as a predictor of fakery is market competition. This is supported by different studies on movies using Twitter (Lee et al. 2018b), hotels using TripAdvisor and Expedia (Mayzlin et al. 2014) as well as some further hotel platforms (Nie et al. 2022), restaurants on Yelp (Luca and Zervas 2016), or products on Amazon (He et al. 2022b) that provide evidence for an effect of industry competition. Based on Nie et al. (2022), competition affects firms to different degrees depending on their similarity. They find that the effect of Airbnb as a new entrant in the lodging market does not affect low-end hotels but enhances self-promotion for high-end hotels. The model of Pu et al. (2022) suggests that all firms are incentivized to engage in fakery in terms of fake sales, reviews, or posts, and Dellarocas (2006a) conclude that firms must engage in fakery resulting from a competitive "rat race"; otherwise, perceptions will be biased against them.

Low Visibility: Another factor that has been observed as a predictor of fakery is low visibility (e.g., Luca and Zervas 2016, Mayzlin et al. 2014). Fakery enables firms to acquire wider reach and signal greater popularity and relevance (Sher 2011). For example, less visible, independent, small firms are more likely to engage in review fakery than larger firms (He et al. 2022b, Luca and Zervas 2016, Mayzlin et al. 2014). This is supported by Lee et al. (2018b), who observe increased sentiment manipulation on Twitter for independently produced movies in contrast to major studios and high-budget movies.

Low Reputation: Fakery can potentially mitigate negative perceptions. Studies show that firms are likely to engage in fakery if they were previously exposed to negative attitudes, e.g., low-ranked health inspection results (Siering and Janze 2019) or weak reputation in terms of reviews (Chen and Papanastasiou 2021, He et al. 2022b, Luca and Zervas 2016, Mayzlin 2006). However, market policies and expectations can make it attractive for firms of all quality levels to engage in fakery (Dellarocas 2006a, Pu et al. 2022). In contrast, the incentives for fakery can also be moderated by reputation. Park et al. (2023) show that the number of reviews moderates the incentive for firms to engage in price increases in combination with the introduction of a list price—their incentive to do so is higher if they have more and better reviews. In this sense, a better reputation enhances engagement in fakery.

Sender Perception: The medium and the ease of fakery are related to the tendency to engage in fakery. A lower cue multiplicity of a particular medium, e.g., text in comparison to face-to-face, makes senders more likely to engage in fakery—this also makes individuals more likely to engage in fakery online in contrast to offline (Xu et al. 2012). However, affective-based trust, i.e., the building of a relationship, has been shown to mediate the effect of lower media richness on increased deception to overcome the higher level of deception in a computer-mediated environment (Rockmann and Northcraft 2008). Also, characteristics such as anger are shown to directly influence empathy, influencing how likely senders are to engage in fakery (Yip and Schweitzer 2016). However, more research is required to validate these findings and provide evidence in more contexts.

Open Questions

In the area of senders of fakery, there are open questions regarding individual characteristics and the role of platforms in engaging in and tolerating fakery.

First, the individual characteristics of senders have not been studied. Research on individual characteristics can complement the insights on institutional characteristics and allows an understanding of the actual sender instead of only the institution behind it. Understanding the individual characteristics would enable researchers to design better

Table 2.1: Summary of Results for the Sender of Fakery
Note: The table shows the main findings for the sender of fakery on digital platforms.

	Definition	Relevant Publications
Profit Motives	• Direct economic or power-related motivations • Fakery to adapt to customer preferences • Payoffs for third parties that produce fakery on behalf of the profiting party	Chen et al. (2022), George et al. (2021), He et al. (2022a), Jin et al. (2023), Keppo et al. (2022), Mullainathan and Shleifer (2005), Qiao and Rui (2023)
Competitive Pressure	• Increased fakery under higher competition	Luca and Zervas (2016), Mayzlin et al. (2014), Nie et al. (2022)
Low Visibility	• Increased fakery for small, independent market actors	He et al. (2022b), Luca and Zervas (2016), Mayzlin et al. (2014)
Low Reputation	• Mitigation of reputational damage • Pushing of products of low quality • “Rat race” with competitors • Higher likelihood for fakery under higher consumer trust (higher prior reputation)	Dellarocas (2006a), He et al. (2022b), Luca and Zervas (2016), Mayzlin (2006), Park et al. (2023), Siering and Janze (2019)
Sender Perception	• Higher fakery with lower cue multiplicity in medium • More fakery with certain emotions of the sender	Rockmann and Northcraft (2008), Xu et al. (2012), Yip and Schweitzer (2016)

control mechanisms and reduce fakery, e.g., by allocating control instances to individuals more prone to fakery. For example, studies of corporate misconduct—not online fakery—have repeatedly investigated individual-level characteristics. For example, it has been observed that personality traits of firms’ CEOs can be associated with fraud or sexual misconduct (Van Scotter and Roglio 2020, Zahra et al. 2005). Future research should investigate such personal characteristics, e.g., of leaders of firms or individuals in the context of platform fakery.

Second, what remains particularly understudied is the role that the platform plays in fakery. Research studies attribute some importance to the platform. Platforms require a substantial user base (*critical mass)* to be successful due to network effects and face a cold-start problem (Evans and Schmalensee 2010, Katz and Shapiro 1985). This makes fakery, such as fake accounts or fake content, an important strategy for platforms to appear larger, thereby solving their cold-start problem (Huang et al. 2018). At the same time, they can increase their own profits, e.g., via click fraud (Edelman 2009, Wilbur and Zhu 2009). So far, these incentives remain little understood and require further studies.

Third, it is unclear whether platforms are incentivized to indirectly tolerate fakery (Candogan and Drakopoulos 2020). On the one hand, tolerating fakery can help platforms appear more popular and prominent and also allows for increased engagement. On the other hand, quality deficiencies can have negative implications. There is a broad understanding of fakery on digital platforms. However, it is unclear whether platform firms tolerate it deliberately, because they are missing the required resources to combat it, or because regulations do not allow them to remove borderline content. In particular, future research should understand whether platform firms are incentivized to remove or tolerate fakery and to what extent this decision is affected by reputational damage.

2.4.3. Who is Affected by Fakery, and Why?

State-of-the-Art

Research has observed different factors that increase the likelihood of receivers falling for fakery. These factors include age, literacy, partisanship, cultural differences, cognitive biases, the ease of content dissemination and algorithmic bias, sender credibility, message characteristics, and receiver perception. Table 2.2 summarizes the main results.

Age: Age seems to play a relevant role in the tendency of individuals to fall for fakery. Studies find that older individuals are more likely to fall for fakery based on memory deficiencies. For instance, older adults are more likely to remember false claims as accurate (Skurnik et al. 2005), and are more likely to fall for repetition-induced fakery if younger adults can double-check claims (Algarni et al. 2017). Nevertheless, Algarni et al. (2017) outline that younger adults are more susceptible to social engineering measures in online social networks.

Literacy: Literacy can reduce the likelihood of falling for fakery (e.g., Gaeth and Heath 1987, George et al. 2021, Johar 2022). A large body of research confirms the positive effects of literacy, analytical thinking, and education on the ability to discern fakery from truth (Algarni et al. 2017, Gaeth and Heath 1987, George et al. 2021, Johar 2022). For example, in the context of social engineering, security knowledge is shown to be negatively related to susceptibility (Algarni et al. 2017). Darke et al. (2010) further show that distrust from unfulfilled expectations carries over to a more general skepticism toward marketing practices, particularly for rather unrealistic advertising claims, showing how experience shapes literacy on subsequent behavior. However, not all studies find positive effects of literacy (Miller et al. 2024). This inconsistency requires future research to investigate how literacy is measured, considering the challenge of biased self-reports and how it is correlated with individual characteristics that potentially lead to a divide.

Partisanship: Partisanship has been associated with falling for fakery for information biased toward one political direction. Findings show that aligned news changes emotions, predicting interaction behavior (Horner et al. 2021) and increasing credibility and sharing bias (Turel and Osatuyi 2021).

Cultural Differences: Cultural differences have been associated with false beliefs. Mostagir et al. (2022) provide model-based evidence that moderate societies relying on their own and others' views are most susceptible to fakery and prevention-focused people are less likely to fall for fakery (Kirmani and Zhu 2007). Also, more dense networks are less susceptible to fakery as fewer knowledgeable individuals are required to spread accurate information than spare networks that require more well-connected spreaders of truth (Mostagir et al. 2022). However, George et al. (2008) do not find general differences in individuals' ability to detect fakery depending on different cultural backgrounds presumed that cultural expectations are irrelevant.

Cognitive Biases: Several research papers have studied cognitive biases influencing fakery effectiveness. First, people believe any claims in the first place as they generally assume that people would tell the truth (Stanley et al. 2022). Despite the truth bias typically leading to low engagement with fakery based on the default pattern of non-interaction, political alignment with the news sender and issue involvement positively moderate this effect (Miller et al. 2024). Second, individuals are unable to differentiate between fakery and accuracy. Individuals are unable to distinguish true from false news because the cues on which they rely (e.g., emotions, speech characteristics) do not reliably predict the truth (Serra-Garcia and Gneezy 2021) or because missing data points lead to misleading conclusions (Stanley et al. 2022). Further, overconfidence from increased media literacy positively affects engagement with fakery (Miller et al. 2024). Third, people likely accept information that aligns with prior views, the so-called confirmation bias. Pre-existing opinions and beliefs determine whether a person believes new information as shown from data (Kim and Dennis 2019, Moravec et al. 2022) and model-based evidence (Rabin and Schrag 1999). While the information that challenges one's opinions receives little cognitive attention and is less likely believed, individuals believe and engage with aligning information (Kim et al. 2019, Moravec et al. 2019). Fourth, continued influence shapes individuals' beliefs even when incorrect information is retracted. Building on relationships integrated into one's beliefs and the closing of a causal gap in a mental model make individuals more likely to believe in fakery even after it is retracted (Chaxel 2022, Cowley and Janus 2004, Hamby et al. 2020). However, for negative stories in which accuracy is more important, this effect is attenuated (Hamby et al. 2020).

Ease of Content Dissemination: Technology does not only foster fakery but also tolerates it. Even though technology allows for more access to information, Shi et al. (2022) show that online articles, review articles, or discussion forums do not reduce but instead amplify fakery. While velocity and processability attenuate fakery, anonymity, rehearsability, parallelism, and engagement with machines instead of humans enhance it (Cohn et al. 2022, Harrison 2018). Further, sender ambiguity strongly predicts rumormongering during crises—personal involvement and anxiety have lower predictive power (Oh et al. 2013). In addition, technologies reduce the control of legal or societal forces to reduce fakery (Nikitkov and Bay 2008).

Algorithmic Bias: Algorithms or recommendation agents enhance fakery. Recommendation agents are generally used to reduce information overload and support consumers' decision-making processes. However, recommendations can be biased and primarily benefit sellers instead of consumers sensitive to those systems (Adomavicius et al. 2013, Xiao and Benbasat 2015). These systems can be easily manipulated, providing a biased perception of the most popular results and potentially leading to filter bubbles that result in a lack of exposure to opposing content (Prawesh and Padmanabhan 2014).

Sender Credibility: Individuals tend to proxy with the information sender for the accuracy of a piece of content—whose credibility becomes increasingly challenging to assess in the digital age (e.g., Cheung et al. 2012, Stanley et al. 2022). Jensen et al. (2013) confirm that reviewer credibility is strongly associated with perceptions of product quality and that this reviewer credibility is affected positively by the two-sidedness and negatively by affect intensity. Algarni et al. (2017) further outline that in social engineering, the perception of a sender in terms of sincerity, competence, attraction, and worthiness makes receivers more susceptible to falling for fakery.

Message Characteristics: Several message characteristics strengthen the belief in fakery. First, the ease of processing, also known as fluency, affects the tendency to fall for fakery by visual signals, natural sequencing, or repetition (King and Auschaitrakul 2020, Stanley et al. 2022). In reviews, this is enhanced by argument quality, consistency, and two-sidedness, i.e., presentation of positive and negative aspects (Cheung et al. 2012). Also, messages can be intentionally interpreted wrongly according to fluency by rhetorical means (Beisecker et al. 2024) or unintentionally where qualifying language—e.g., "unlikely", "improbable"—is used (Stanley et al. 2022). Roggeveen and Johar (2002) find evidence that fluency is driven by subjective familiarity, but they also show that implausible claims become more credible when derived from several senders. The strength of the fluency effect differs by age (Law et al. 1998, Skurnik et al. 2005, Stanley et al. 2022). Second, besides the belief in fakery predicting the intent to share a piece of information (Chen et al. 2021), further message characteristics drive sharing.

London Jr et al. (2022) outline that users tend to share unverifiable messages if they perceive them as helpful or novel, for which they proxy with content and non-content characteristics, namely plausibility, vividness, and sender credibility. Also, soft and hard news is more likely retweeted than general news and messages that aim to manipulate are more likely retweeted than messages that cover the latest news (Akar et al. 2021). Some further message features such as mentions, emojis, punctuation, or the number of tweets are associated with a retweeting intention on the social media platform Twitter (Akar et al. 2021).

Receiver Perception: Several receiver perceptions indicate some moderating role for the tendency to fall for fakery and spread it further. First, research investigates the state of individuals to understand factors that attenuate or strengthen the belief in fakery. Model-based evidence shows a lower ability to detect fake news from overconfidence (Kartal and Tyran 2022). Disturbance, mind-wandering, memory impairment, emotions, cognitive load, and effort during the information processing reduce the ability to discern true from fake (Appan and Browne 2012, Craig et al. 2012, Deng and Chau 2021, Stanley et al. 2022). For example, social techniques, such as interviews, are more likely to induce misinformation in the communication, which is later on believed (Appan and Browne 2012). However, in contrast to Craig et al. (2012) that provide evidence for adverse effects of cognitive load, Twyman et al. (2020) show that multitasking as a form of increased cognitive load can reduce communication performance and, therefore, the tendency to fall for fakery. Second, the medium affects the extent to which fakery is believed. George et al. (2018) find that media with fewer cues to deception make individuals less likely to detect fakery. Contrasting perceived deception in online and offline retailing, Riquelme and Román (2014) outline that cognitive factors are more relevant in online retailing. Third, external factors determine the spread of fakery. Information overload, trust in online information, mobile connectivity, and political freedom positively relate to sharing unverified information (Laato et al. 2020, Shirish et al. 2021). Further, individuals share fakery even if they are incentivized not to, proving their low ability to discern truth from fakery, but also shedding light on the negative consequences if receivers know about the incentive to only share accurate information and trust the information more (Serra-Garcia and Gneezy 2021).

Table 2.2: Summary of Results for the Receiver of Fakery
Note: The table shows the main findings for the receiver of fakery on digital platforms.

	Definition	Relevant Publications
Age	• Higher susceptibility to false claims and misinformation for older adults • Higher susceptibility to social engineering for younger adults	Algarni et al. (2017), Gaeth and Heath (1987), Skurnik et al. (2005)
Literacy	• Positive effects of literacy on detection of fakery in most, but not all studies	George and Robb (2008), Johar (2022), Miller et al. (2024)
Partisanship	• Influence of partisanship on belief in politically aligned fakery	Horner et al. (2021), Turel and Osatuyi (2021)
Cultural Differences	• Lower susceptibility for prevention-focused people • Higher likelihood of spread of fakery in sparse and less connected networks • Detection ability irrelevant from the cultural background	George et al. (2008), Kirmani and Zhu (2007), Mostagir et al. (2022)
Cognitive Biases	• Initial acceptance of all information as true (truth bias) • Inability to discern fakery from the truth • Acceptance of information in alignment with prior views (confirmation bias) • Continued influence of false information after retraction	Chaxel (2022), Hamby et al. (2020), Kim and Dennis (2019), Moravec et al. (2022), Rabin and Schrag (1999), Serra-Garcia and Gneezy (2021), Stanley et al. (2022)
Ease of Content Dissemination	• Toleration and enhancement of fakery via new media driven by novelty, anonymity, and low control	Harrison (2018), Nikitkov and Bay (2008), Shi et al. (2022)
Algorithmic Bias	• Enhancement via algorithmic recommendations and filter bubbles	Adomavicius et al. (2013), Prawesh and Padmanabhan (2014)

Table 2.2: Summary of Results for the Receiver of Fakery (cont.)

	Definition	Relevant Publications
Sender Credibility	• Proxy for the accuracy of information with the sender	Cheung et al. (2012), Stanley et al. (2022)
Message Characteristics	• Belief in fakery driven by message fluency and familiarity • Sharing further enhanced by novelty and liveliness	Akar et al. (2021), King and Auschaitrakul (2020), London Jr et al. (2022), Roggeveen and Johar (2002), Stanley et al. (2022)
Receiver Perception	• Enhancement of belief in fakery because of cognitive load and disturbance, negative emotions, and social influence • Differences in the medium driven by a variety of cues to deception	Appan and Browne (2012), Craig et al. (2012), Deng and Chau (2021), George et al. (2018), Riquelme and Román (2014), Stanley et al. (2022), Twyman et al. (2020)

Open Questions

Open questions remain about firms as receivers of fakery, message cues that affect falling for fakery, and peer influence.

First, non-individual receivers have not been studied. There is substantial research on individuals as receivers of fakery (e.g., Horner et al. 2021, Kim and Dennis 2019, London Jr et al. 2022, Miller et al. 2024, Moravec et al. 2022, Mostagir et al. 2022, Oh et al. 2013, Serra-Garcia and Gneezy 2021, Turel and Osatuyi 2021, Xiao and Benbasat 2015). In contrast, non-individuals, e.g., firms, as receivers of fakery are poorly understood, e.g., by investigating how firms are affected by fakery by gatekeepers (Wilbur and Zhu 2009) or competitors (Song et al. 2019), as well as spillover effects (Darke et al. 2010). In this context, it is important to carry out research that understands (1) whether non-individuals are receivers of fakery, (2) which factors moderate the targeting, and (3) whether consequences for these receivers differ from individuals.

Second, there remains an open question in the context of message cues. Research has started to investigate the context and content of messages to assess differences in believability, e.g., communication medium (George et al. 2018), message sender (Cheung et al. 2012, Jensen et al. 2013, Stanley et al. 2022), or photographic evidence (Stanley et al. 2022), as well as perceived quality and consistency of reviews (Cheung et al. 2012). Further, sharing is determined by novelty (London Jr et al. 2022) or message features (Akar et al. 2021, Serra-Garcia and Gneezy 2021). Future research should investigate how certain cues of fake objects determine whether they are believed and shared, e.g., the format of objects (videos, images, text), tone, language professionalism, engagement metrics, timing, or platform reputation.

Third, the extent to which peer influence manifests in the belief of fakery remains an open question. It is shown that individuals tend to recall misinformation from an interview during the information requirements determination process (Appan and Browne 2012). Also, the literature provides evidence for peer influence in other settings (e.g., Muchnik et al. 2013, Wang et al. 2018). It becomes crucial to understand to what extent peer influence drives the belief and sharing of fakery on digital platforms.

2.4.4. What are the Consequences of Fakery?

State-of-the-Art

Fakery has various consequences, affecting online communication, individuals, society, senders, and bystanders. Table 2.3 summarizes the main results.

Consequences for Online Communication: The literature finds a change in communication behavior when exposed to fakery. Depending on network density and connectedness, the participation of social bots even below 5% can often change opinions (Ross et al. 2019). This is closely related to the theory of the spiral of silencing, which indicates that individuals are unlikely to express their own opinion publicly if they perceive it to differ from that of the majority (Elisabeth 1974). If bots create the illusion that the majority reflects their opinion, it becomes the prevalent opinion.

Consequences for Individuals: New information—true or false—requires individuals to update their beliefs. This proves particularly challenging in the presence of contradicting statements and fakery (Sadler 2021). Cognitive and affective mechanisms are activated once fakery is in place—advertisers try to create arousal and pleasure, which affect consumer perception and behavior (Xiao and Benbasat 2011). While sophisticated individuals learn well in an environment with sufficient accurate information, naive societies can outperform sophisticated societies in an environment with increasing fakery where sophisticated individuals cannot agree on a true state and become paranoid (Mostagir and Siderius 2022). In particular, when confronted with fakery, individuals with low levels of competence are likely to overestimate their competence (Kruger and Dunning 1999). This has real-life implications for individuals as false information and assumptions drive beliefs and decision-making (e.g., Barone and

Miniard 1999, Dellarocas 2006a, Johar 1995, Shi et al. 2022). In particular, highly involved consumers tend to develop purchasing intentions based on incorrect inferences while processing fakery through misleading advertisements and enhanced by other online sources (Johar 1995, Shi et al. 2022). This leads to misleading decisions if fakery does not align with true quality and often results in a social loss (Chen et al. 2022, Dellarocas 2006a). A more diversified angle indicates that the text of incentivized reviews can still provide valuable insights despite potential bias in the numerical rating (Qiao and Rui 2023). To complement, the literature investigates how individuals are differently affected by fakery. For example, Van Bommel (2003) outlines that honestly appearing fakery can influence others' decision-making—in this sense, informed investors can spread rumors through imprecise trading advice to manipulate the market prices and benefit from the overshooting. Inequality is also shown depending on different forms of societies—if agents induce fakery in a network and impede proper learning, communities with no access to knowledgeable agents are weakened (Mostagir and Siderius 2023a).

Consequences for Society: Negative effects emerge on collective decision-making. As individuals are likely to overestimate their competence in identifying fakery according to the Dunning-Kruger effect (Kruger and Dunning 1999), deficiencies in collective decisions emerge (Kartal and Tyran 2022). George et al. (2021) identify persuasion, conviction, polarization, and aversion as leading implications of exposure to fakery. Negative real-life effects are driven by incorrect beliefs from fakery in diverse areas of life. Fake news increases voting for populist parties despite not explaining the whole effect and influences attitudes toward societal challenges such as the causes and the importance of global warming (Cantarella et al. 2023, Cho et al. 2011). Further, exposure to fake reviews negatively affects consumer behavior, particularly for experience goods, with declining marginal returns, and reduces the overall credibility of reviews (Zhao et al. 2013).

Consequences for Senders: Consequences for the sender are positive but become negative once fakery is revealed. First, the research investigates the initial positive reactions of consumers and investors. In terms of consumers, fakery can improve the performance of firms, e.g., sponsored listings without explicit disclosure on e-commerce platforms can lead to misinterpretation and improve performance, or fake reviews can push visibility in search results (Deng et al. 2021, He et al. 2022b, Lappas et al. 2016). Fakery can also attract attention, e.g., Rao (2022) show that fabricated overly-positive information about products in the sense of fake ads attracts significant site visits. Also, increasing a price but introducing a list price simultaneously improves sales ranks and profit margins for sellers (Park et al. 2023). However, Darke et al. (2010) outline that deviations between implementation and expectation lead to a decline in trust and adverse spillover effects for unrelated products and firms. Regarding investors, fakery gets substantial attention on social media, e.g., Clarke et al. (2021) show that fake articles on Seeking Alpha generate 83.4% more page views on average in comparison to legitimate articles, and Jia et al. (2020) show that rumors attract tweeting volume, which can potentially distort price discovery concerning highly speculative merger rumors. Increased attention manifests in stock market reactions—Clarke et al. (2021) find evidence that there is increased trading volume on the release date of fake news and Jia et al. (2020) show that higher tweeting volume is significantly associated with market reactions even though not a signal of accuracy. Also Ullah et al. (2014) find evidence for abnormal returns and trading volume in response to fakery—with the effect even holding despite false information being denied.

Second, research investigates the adverse effects on firms. Pu et al. (2022) find that in equilibrium, fakery in the form of quality misrepresentation always hurts low-quality sellers, while effects for high-quality sellers depend on market conditions. Further, anticipated fakery can force firms to engage in fakery: Fakery in a marketing setting can be detrimental to firms and leads to a rat race if consumers anticipate fakery and accordingly discount the value of online ratings (Dellarocas 2006a). Also, there are forms of fakery in which firms are victims. For example, platform firms or competitors can use click fraud to increase search advertising costs for firms without increasing sales (Wilbur and Zhu 2009).

Third, the literature investigates the negative reactions of consumers and investors when fakery is revealed. Consumers penalize fakery when revealed, e.g., there is a significant decline in website visits and product demand in response to consent orders of the Federal Trade Commission (Rao and Wang 2017), a decline in brand attitudes and purchase intentions (Xie et al. 2015), and reduced satisfaction and loyalty (Román 2010). If they have a negative prior evaluation of the advertiser, consumers hold them more accountable for fakery (Johar 1996). Also, consumers taking the perspective of a salesperson engaging in fakery makes them less tolerant of this unethical behavior if they have high moral self-awareness (Xie et al. 2022). To assess practices directed against competitors, Song et al. (2019) show that engaging in pseudo-harm crises to harm competitors is detrimental to both the offending and the victim firm regarding consumer sentiment. Aligning with the latter, Tergiman and Villeval (2023) show that in the interaction of project managers and investors, reputation mechanisms do not increase honesty but make project managers shift from detectable to deniable lies. Not only consumers but also investors react to the revelation of fakery (e.g., Kogan et al. 2023, Tipton et al. 2009, Wiles et al. 2010). Empirical evidence shows a decline in abnormal stock returns in response to revealed corporate misconduct, corporate illegalities, and deceptive marketing (Davidson III and Worrel 1988, Murphy et al. 2009, Tipton et al. 2009, Wiles et al. 2010). For example, Wiles et al. (2010) find adverse effects of revealed deceptive advertising practices, i.e., regulatory reports of misleading ads, which are associated with abnormal returns of -0.91% mitigated by omission bias, i.e., the reactions are more negative when information is misrepresented instead of omitted, and reputation. Kogan et al. (2023) provide evidence for a drop in trading volume and price volatility following an SEC investigation's revelation of fraudulent news.

Consequences for Bystanders: Research also investigates spillover effects when deceptive practices are revealed. When quality deviates from expectation, emerging distrust, self-protection, and skepticism lead to adverse reactions to unrelated advertisements, products, firms, and news (Darke and Ritchie 2007, Darke et al. 2010, Kogan et al. 2023). For example, Kogan et al. (2023) show that the revelation of fraudulent news indicated by an SEC investigation reduces the overall effect of the news on trading behavior, also for legitimate ones.

Open Questions

Overall, studies seem to align in the direction of findings. Fakery shapes the formation of attitudes (Cho et al. 2011, Ross et al. 2019) and affects consumers (He et al. 2022b, Rao 2022), investors (Clarke et al. 2021, Jia et al. 2020), and society (Cantarella et al. 2023). They all associate exposure to fakery that remains undetected with positive effects for the sender, leading to opinion distortion in favor of the sender by changing opinions (Cho et al. 2011, Ross et al. 2019), resulting in higher visibility (Deng et al. 2021, He et al. 2022b, Jia et al. 2020), and ultimately leading to improved performance (Clarke et al. 2021, Jia et al. 2020). However, the awareness or perception of fakery reduces these effects (Szabo and Webster 2021, Zhao et al. 2013) and overall, fakery is not as effective as organic signaling (Deng et al. 2021). At the same time, researchers also align in the effects after fakery is revealed, which are consistently negative in terms of consumer (Rao and Wang 2017, Rao 2022, Xie et al. 2015) and investor reactions (Murphy et al. 2009, Tipton et al. 2009, Wiles et al. 2010).

However, there is a need to further detail the consequences of unrevealed fakery, mitigation of consequences, and spillover effects.

First, there is little insight into the second-order consequences of fakery regarding anticipation effects (Dellarocas 2006a, Mostagir and Siderius 2023b, Zhao et al. 2013). Future research should more thoroughly investigate whether the adverse effects of fakery appear even if not revealed—this can be the case implicitly from experience and unfulfilled expectations (Darke et al. 2010) but also explicitly from contradicting statements. Tergiman and Villeval (2023) highlight that the relevance of reputation makes managers shift from detectable to deniable lies or Foerderer and Schuetz (2022) show how firms strategically time data breach announcements, and it becomes crucial to understand whether similar behavior is observed for other senders of fakery. This

Table 2.3: Summary of Results for the Consequences of Fakery
Note: The table shows the main findings for the consequences of fakery on digital platforms.

Consequences for ...	Definition	Relevant Publications
Online Communication	• Influential power of social bots in shaping discussions and silencing human users	Elisabeth (1974), Ross et al. (2019)
Individuals	• Distortion in the updating of beliefs due to noisy information • Decision-making derived from false assumptions • Exploitation of false information under access to true information • Negative implications for firms from fakery of search engines	Barone and Miniard (1999), Johar (1995), Mostagir and Siderius (2022, 2023a), Sadler (2021), Shi et al. (2022), Van Bommel (2003), Wilbur and Zhu (2009)
Society	• Negative implications for collective decision-making driven by overconfidence • Reduction in helpfulness and trust toward information • Real-life implications in the formation of opinions	Cantarella et al. (2023), Cho et al. (2011), Kartal and Tyran (2022), Kruger and Dunning (1999), Zhao et al. (2013)
Senders	• Positive effects on firm performance • Negative effects for low-quality firms, “rat races” among firms • Negative consumer and investor reactions to the revelation of fakery	Clarke et al. (2021), Dellarocas (2006a), He et al. (2022b), Jia et al. (2020), Kogan et al. (2023), Lappas et al. (2016), Pu et al. (2022), Rao (2022), Rao and Wang (2017), Román (2010), Tipton et al. (2009), Ullah et al. (2014), Wiles et al. (2010), Xie et al. (2015)
Bystanders	• Negative spillover effects after fakery revelation	Darke and Ritchie (2007), Darke et al. (2010), Kogan et al. (2023)

sheds light on whether fakery proves as the lucrative business it is often assumed to be.

Second, little is known about potential mitigation strategies. In other settings, it is shown that firm engagement is positively related to customer satisfaction and firm performance (e.g., Chung et al. 2020). However, it is unclear how this is also valid in a setting of fakery performed by firms or individuals. Future research should (1) understand whether the engagement of senders with related stakeholders can mitigate the adverse effects of the revelation of fakery and (2) investigate how they should communicate with consumers and investors.

Third, it is not fully clear how spillover effects of fakery manifest (e.g., Darke et al. 2010, Kogan et al. 2023). For example, Song et al. (2019) outline that pseudo-harm crises directed at competitors harm both the sender and the receiver. However, what happens to the demand remains an open question. Is there only a short-term effect? Do individuals switch completely to other industries, or is there an overall behavior change, i.e., a decline in trading or consumption? To understand the actual behavior and underlying thoughts, qualitative approaches such as interviews could provide helpful answers.

2.4.5. What are Countermeasures Against Fakery?

State-of-the-Art

This review differentiates between preventive and corrective measures to answer the question of what different forms of countermeasures exist. For preventive measures, this review covers awareness creation, warnings, and platform design decisions. Table 2.4 summarizes the main results.

Awareness Creation: One way to reduce susceptibility to fakery is to enhance awareness of fakery. Policies that regulate fakery should consider the autonomy of consumers and advertisers by improving self-criticism and a sense of responsibility (Attas 1999, Sher 2011). One strategy is to reduce cognitive load (Craig et al. 2012) or develop literacy via training to create skepticism (Gaeth and Heath 1987, George et al. 2021, Johar 2022). Training can anticipate and forestall fakery to create skepticism (Johar 2022) or induce a feeling of responsibility when spreading information (Lamy 2023). For example, Moravec et al. (2022) show that asking users about their knowledge when rating news stories makes them more skeptical and less likely to believe and share fakery. Also, access to information that makes individuals aware of fakery reduces the incentive for senders to engage in fakery (Heese et al. 2022). As literacy is shown to be of high relevance in distinguishing fakery from the truth, it becomes crucial to understand how literacy training can be most effective. Various scholars highlight the relevance of literacy training and education to combat fakery (Gaeth and Heath 1987, George et al. 2021). However, it is unclear how exactly individuals should be trained. Wilson et al. (2022) highlight that specific training likely makes individuals more susceptible to other forms of fakery. To overcome this, measures that enhance general skepticism should be developed instead of relying on specific training programs. Future research should elaborate on how general training for literacy development should be designed to be most effective in various settings.

Warnings: Warnings before exposure to fakery are discussed in the literature (e.g., Moravec et al. 2020, Stanley et al. 2022, Xiao and Benbasat 2015). Warnings are most effective if they encourage critical questioning of information to which individuals will be exposed and should trigger both automatic and deliberate (e.g., training) cognition (Moravec et al. 2020, Stanley et al. 2022). Trendel et al. (2018) show that image-based warnings on deceptive advertising or product recalls are superior to text-based warnings.

Platform Design: Platforms need to define an optimal signaling mechanism that recommends engaging with content if it is below a certain threshold of inaccuracy (Candogan and Drakopoulos 2020)—there are certain strategies to provide such helpful information. First, platforms use verification and ratings to reduce fakery. Verification can help to reduce anonymity and increase accountability and consequently reduce the likelihood of engaging in fakery, but a verification badge negatively moderates the effect (Wang et al. 2021a). For example, Mayzlin et al. (2014) provide support for allowing for verified and non-verified reviews in different formats, and Kokkodis et al. (2022) highlight that an optional disclosed verification strategy can increase review quality and helpfulness. Also, labeling incentivized reviews as such and only providing qualitative and not quantitative reviews can help (Qiao and Rui 2023). However, Mostagir and Siderius (2023b) show that policy interventions that try to reduce the number of bribing firms for fake reviews potentially have unintended consequences. Second, assessing the quality of users before allowing them to engage in content creation can reduce fakery. Evaluating the ability of users to edit an article in terms of quality, reputation, and editing patterns in managed wikis can help to overcome the spam problem (Wöhner et al. 2015). Closely related are ratings of credibility to increase helpfulness of ratings (Jabr 2022). In contrast, mutual reviews reduce the incentive to provide negative reviews, ultimately reducing helpfulness (Donaker et al. 2019). However, surprisingly, Wu and Geylani (2020) find that stricter regulation potentially reduces consumer surplus as consumers adapt their expectations toward less fakery. Third, technical design decisions can combat fakery. For example, ranking designs affect how consumers perceive items and should give equal weight to older reviews (Lappas et al. 2016) and recommendation systems should be designed to be robust against fakery (Prawesh and Padmanabhan 2014, Van Roy and Yan 2010). Filtering technologies that reduce the ability for fakery adapt consumers' expectations and reduce the pressure on senders to engage in fakery (Dellarocas 2006a). However, increased costs for fakery, higher detection rates, and reduced search costs for consumers are not always beneficial to reducing fakery (Chen and Papanastasiou 2021, Jin et al. 2023). Besides the presentation of recommendations, the presentation of the content itself matters. For example, Kim and Dennis (2019) show that highlighting the sender of an article and including negative sender ratings enhance skepticism and reduce believability. At the same time, downstream measures such as a high return leniency or delayed payment can reduce the incentive to engage in fakery (Edelman 2009, Pu et al. 2022). Also, to avoid fakery, the intervention of third parties could be beneficial, e.g., a third party could audit the click fraud independently (Chen et al. 2015b, Wilbur and Zhu 2009).

Table 2.4: Summary of Results for the Preventive Countermeasures of Fakery

Note: The table shows the main findings for the preventive countermeasures of fakery on digital platforms.

	Definition	Relevant Publications
Awareness Creation	• Reduction of cognitive load • Enhancement of literacy development via general training • Awareness creation during the processing of information	Craig et al. (2012), Gaeth and Heath (1987), Johar (2022), Moravec et al. (2022), Wilson et al. (2022)
Warnings	• Warnings that encourage critical questioning and trigger cognition • Superiority of image-based in comparison to text-based warnings	Moravec et al. (2020), Stanley et al. (2022), Trendel et al. (2018), Xiao and Benbasat (2015)
Platform Design	• Verification of reviewers and content creators • Provision of information quality signals via ratings and labeling • Enhancement of skepticism via sender highlighting • Smart design of recommendation systems • Use of filtering technology • Lower levels of fakery with higher return leniency • Third-party audits for detection of fakery • Effects of lower search costs, higher brushing costs, and higher detection rate not always positive	Chen et al. (2015b), Dellarocas (2006a), Donaker et al. (2019), Jabr (2022), Jin et al. (2023), Kim and Dennis (2019), Kokkodis et al. (2022), Lappas et al. (2016), Mayzlin et al. (2014), Prawesh and Padmanabhan (2014), Pu et al. (2022), Qiao and Rui (2023), Van Roy and Yan (2010), Wang et al. (2021a), Wilbur and Zhu (2009), Wöhner et al. (2015)

For the corrective approaches, this review investigates corrective messages, fact-checking, crowd-based approaches, automated approaches, and response strategies. Table 2.5 summarizes the main results.

Corrective Messages: Several studies confirm that corrections can successfully refute fakery (Johar and Roggeveen 2007). Corrective messages should be kept general and align with prior beliefs and identities of individuals (Johar 2022, Stanley et al. 2022). The sender of the correction in terms of authority (e.g., company vs. FTC) does not seem to matter (Armstrong et al. 1979). However, several studies question the effectiveness of corrective messages. Fluency and repetition of fakery via corrections enhance the continued influence (Hamby et al. 2020), and corrective messages do not reduce fakery but attract more in following periods (King and Auschaitrakul 2020).

Fact-Checking: Fact-checks rely on checks by experts to provide labels about the truth of statements as a promising measure to combat fakery (e.g., Khan et al. 2022a, Moravec et al. 2020, Schuetz et al. 2021). For example, Schuetz et al. (2021) find more desirable protection behavior during the COVID-19 pandemic in response to fact-checks. However, fact-checking is heavily criticized, e.g., Moravec et al. (2019) show that labels increase time spent on an article but do not shape beliefs. To make fact-checking more effective, correct facts should be added to refutations (Schuetz et al. 2021), and verifications and refutations should both be provided to counter ambiguity (Pennycook et al. 2020). This shows that whether rebuttals are effective and how they should be designed is unclear. The differing insights require future research to understand what drives successful rebuttals and fact-checks compared to alternatives.

Crowd-Based Approaches: To reduce their costs, platforms can complement their measures by involving users in the content inspection process if content validity is unknown ex-ante (Chua et al. 2007, Papanastasiou 2020). Community-based crowd intelligence can be exploited to detect social bots (Benjamin and Raghu 2023) or fraudulent reviews (Donaker et al. 2019), but also for content rating (Kim et al. 2019) and reporting (Gimpel et al. 2021, Wang et al. 2022). For example, Kim et al. (2019) show that sender ratings affect the believability of fake news. At the same time, the inclusion of social norms in the social media design, as well as video content and more intense sentiments expressed in a text, enhance the use of reporting options (Gimpel et al. 2021, Wang et al. 2022).

Automated Approaches: As fakery is constantly increasing on digital platforms and manual monitoring is impossible, platform firms increasingly invest in automated measures for detection (He et al. 2022b, Khan et al. 2022a). While content monitoring is often done as a corrective measure, predicting one’s likelihood to engage in fakery can also be preventive. First, literature investigates content-based detection approaches using linguistic and content-based cues of messages (Clarke et al. 2021, Siering et al. 2016, Zhou et al. 2004). For example, the language in fake reviews differs from authentic reviews because different types of memory are used (Kronrod et al. 2023). To obtain training data, Ng et al. (2023) suggest using human intelligence. Second, besides text-based approaches, the sender of a message can be assessed. The literature investigates sender-based assessment to detect fakery in the form of fake reviews relying on the behavior of users in their reviewing behavior, e.g., review gap, count, scores, and length, rating entropy and deviation, time of review, or tenure (Kumar et al. 2018, 2019). Zhou et al. (2004) further show that aggregating message cues on a subject level can reduce data points without losing classifier accuracy. Third, the increasing sophistication of fakery leads to new challenges for automated detection. Reviewers disguise their fake reviews as organic over time in terms of verbal features, and non-verbal features become increasingly important to assess fakery (Abdulqader et al. 2022, Ho et al. 2016, Ludwig et al. 2016, Zhang et al. 2016). For example, Ho et al. (2016) find evidence for cues of fakery related to affection, immediacy, cognitive load, or wordiness. Scholars suggest combining user judgment with machine intelligence (Wei et al. 2022), using external data (Zhang et al. 2022), or analyzing trace data (Weinmann et al. 2022).

Response Strategies: The research investigates how to proceed after fakery is suspected or detected. For example, firms can compensate for fake reviews by confronting fake reviews or investing in marketing (Lappas et al. 2016). At the same time, platforms can remove fraudulent reviews even though it is often helpful to leave them for awareness (Ananthakrishnan et al. 2020). For example, Ng et al. (2021) show that fake news flags reduce the forwarding of fake news while a forwarding restriction for accounts reduces the survival time of information. However, Piccolo et al. (2018) show that in some cases, it can make sense for firms to tolerate fakery in the form of deceptive advertising by low-quality competitors to increase profits.

Open Questions

Despite these insights, there are some gaps in the literature about the effectiveness and design of countermeasures, unintended fakery or unintended negative consequences of countermeasures, and interactions between manual and automated detection approaches.

First, despite the existence of various measures to combat fakery, it becomes increasingly important to understand how to assess the effectiveness of such measures. Carson et al. (1985) criticize the attempt to determine the need for intervention based on a cost-benefit ratio, as high costs are often not outweighed by perceived benefits that ignore societal costs. Following this statement and several high-level techniques to assess fakery as suggested by Gardner (1975) or Sher (2011), future research should consider how (1) the need for intervention is determined, (2) the success is measured, and (3) whether external pressure is required if it is not profitable internally to engage in the countering of fakery.

In that context, it remains unclear to what extent measures that amplify credible content can complement measures that focus on the pruning of fakery. In particular, such measures align with the free speech fostered on digital platforms. Prior studies show that corrective messages and fact-checking that refute fakery can effectively combat fakery (e.g., Johar and Roggeveen 2007, Pennycook et al. 2020, Schuetz et al. 2021),

Table 2.5: Summary of Results for the Corrective Countermeasures of Fakery

Note: The table shows the main findings for the corrective countermeasures of fakery on digital platforms.

	Definition	Relevant Publications
Corrective Messages	• General form of corrective messages • No immediate sender effects for the sender of corrective messages • Unintended consequences of corrective messages by continued influence and attracting fakery	Armstrong et al. (1979), Hamby et al. (2020), Johar and Roggeveen (2007), Johar (2022), King et al. (2021)
Fact-Checking	• Positive effects of fact-checks that include corrections • Need for fact-checks in terms of rebuttals and verifications • Criticism in terms of the effectiveness of fact-checks	Moravec et al. (2019), Pennycook et al. (2020), Schuetz et al. (2021)
Crowd-Based Approaches	• Use of crowd intelligence for identification of social bots, fake reviews, or untrustworthy senders • Higher likelihood for the use of reporting options under the presence of social norms, for videos, and higher sentiment intensity	Benjamin and Raghu (2023), Donaker et al. (2019), Gimpel et al. (2021), Kim et al. (2019), Wang et al. (2022)
Automated Approaches	• Content-based approaches • Sender-based approaches • Performance improvement via non-verbal features, message features, and external cues • Initial evidence for using trace data	Ho et al. (2016), Kumar et al. (2018, 2019), Siering et al. (2016), Zhang et al. (2016), Zhou et al. (2004)
Response Strategies	• Enhancement strategies for firms affected by the fakery of competitors • Superiority of displaying fraudulent reviews in comparison to censoring • Attenuation of dissemination of fakery via flags • Reduction of the survival time of fakery for account-level forwarding restrictions	Ananthakrishnan et al. (2020), Lappas et al. (2016), Ng et al. (2021)

but at the same time show that continued influence can still shape beliefs (Hamby et al. 2020). To overcome this, a potential strategy should be investigated to amplify credible content instead of focusing on detecting and correcting fakery to make users better informed. Some studies started doing so by introducing ratings or verifications for contents and sources (e.g., Jabr 2022, Kim and Dennis 2019, Mayzlin et al. 2014, Qiao and Rui 2023, Wang et al. 2021a), however, these studies provide mixed results and do not evaluate the effect on content consumption for credible content. Future studies should build upon that to better understand the implications of such strategies. One measure, namely promoting content from credible sources, will be covered in the empirical study in Chapter 4 of this dissertation.

Second, it is unclear how to combat unintentional fakery and overcome unintended negative consequences of countermeasures to combat fakery. In terms of unintentional fakery, the literature indicates that fakery is potentially transmitted unintentionally (Ferrara et al. 2016, Stanley et al. 2022). This effect is then reinforced by technology, e.g., echo chambers or recommendation agents that only provide selective exposure to information and filter bubbles emerging from it (e.g., Adomavicius et al. 2013, Allcott et al. 2019, Benbasat and Wang 2005, Pariser 2011, Prawesh and Padmanabhan 2014). These findings open up new research areas to investigate platform decisions. Future research should understand how platforms can be designed to (1) nudge users toward accuracy checking and (2) reduce echo chambers and filter bubbles. Additionally, countermeasures are generally assessed for their effectiveness in reducing fakery; however, the side effects of these interventions are often unclear. For example, optional verification as an entry barrier is discussed in various papers as a measure to reduce fakery (e.g., Donaker et al. 2019, Kokkodis et al. 2022, Mayzlin et al. 2014, Wang et al. 2021a). However, Mayzlin et al. (2014) highlight that verification also reduces contributions overall as it limits the set of users able to contribute and suggest an optional verification approach where Wang et al. (2021a) show that such strategies are ineffective. Overall, future research should investigate the unintended side effects of countermeasures, particularly those consequences that harm a platform firm. When incentives for platform firms are not in line with the outcomes of the countermeasures, either (1) regulation becomes necessary or (2) countermeasures need to be adapted and aligned with the platforms' business interests.

Third, the extent to which automated measures can be used effectively in combination with manual detection approaches remains an open question. Further, automated detection approaches are often used as a corrective measure; however, future research could also investigate its potential as a preventive measure, e.g., by monitoring users. Facing the rising complexity and developments in fakery, future research should investigate how to design and adapt algorithms, e.g., for multimodal approaches, to keep up with the pace of development.

2.5. Conclusion

This scoping review of the literature on fakery on digital platforms gives an overview of the current state of knowledge of the different forms of fakery on digital platforms. It develops an adapted sender-receiver framework that investigates how fakery manifests on digital platforms and what research questions emerge around this phenomenon. It provides an overview of the state-of-the-art literature on platform fakery and identifies open questions after synthesizing relevant papers.

This literature review and current public debates motivate the two follow-up empirical studies in the context of fakery on digital platforms.

First, departing from the various forms of fakery, the review indicates little research on fake accounts, particularly fake followers. While research does not focus on this form of fakery, the public considers it a strong challenge for digital platforms (Confessore et al. 2018, Federal Trade Commission 2023). Considering the public interest and the gap in the literature, the first empirical study in this dissertation thoroughly investigates the extent to which fake accounts are used and evaluates their risks for investors.

Second, the COVID-19 pandemic and the associated infodemic have recently shed light on the dimension of fakery in critical situations, with the role of social media being major (Pian et al. 2021). To make digital platforms more responsible, the public pressures them and asks for a stronger commitment to measures against misinformation (Colomina et al. 2021, Duffy 2022). The literature review hints at the complementary nature of countermeasures that amplify credible information. Departing from that gap in the literature, the second empirical study in this dissertation investigates a countermeasure in the health context that promotes credible content by featuring content from authoritative sources in the search results and providing a label to their content. The empirical analysis investigates whether such an approach enhances the consumption and production of credible content.

2.6. Appendix for Chapter

Table A2.1: Overview of Literature Selection
Note: The table shows how many papers were found and selected for the main search, the backward search, and the forward search. For the main search, there is a split by journals.

Outlet	Hits	Selected
Main Search	4,215	129
Academy of Management Journal	53	
Academy of Management Review	16	
Accounting, Organizations and Society	79	
Administrative Science Quarterly	19	
American Economic Review	75	3
Business & Information Systems Engineering	27	1
Contemporary Accounting Research	131	
Econometrica	68	1
Entrepreneurship Theory and Practice	12	
European Journal of Information Systems	41	7
Harvard Business Review	107	2
Human Relations	44	
Human Resource Management	15	
Information Systems Journal	22	1
Information Systems Research	82	14
Journal of Accounting & Economics	78	1
Journal of Accounting Research	73	
Journal of Applied Psychology	247	
Journal of Business Ethics	531	13
Journal of Business Venturing	17	
Journal of Consumer Psychology	94	7
Journal of Consumer Research	135	4
Journal of Finance	58	1
Journal of Financial and Quantitative Analysis	41	1
Journal of Financial Economics	89	1
Journal of Information Technology	27	1
Journal of International Business Studies	23	1
Journal of Management	54	
Journal of Management Information Systems	78	19
Journal of Management Studies	40	
Journal of Marketing	54	1
Journal of Marketing Research	108	10
Journal of Operations Management	19	
Journal of Political Economy	33	
Journal of Strategic Information Systems	13	
Journal of the Academy of Marketing Science	53	1

Table A2.1: Overview of Literature Selection (cont.)

Outlet	Hits	Selected
Main Search	4,215	129
Journal of the Association for Information Systems	32	1
Management Information Systems Quarterly	81	10
Management Science	256	12
Manufacturing and Service Operations Management	30	
Marketing Science	84	6
MIT Sloan Management Review	21	
Operations Research	91	2
Organization Science	78	
Organization Studies	49	
Organizational Behavior and Human Decision Processes	239	2
Production and Operations Management	70	3
Quarterly Journal of Economics	37	1
Research Policy	58	1
Review of Accounting Studies	52	
Review of Economic Studies	36	
Review of Finance	18	1
Review of Financial Studies	65	
Strategic Entrepreneurship Journal	2	
Strategic Management Journal	64	
The Accounting Review	196	
Backward Search		9
Forward Search		4
TOTAL		142

Table A2.2: Concept Matrix of Literature on Fakery
Note: The table describes the concept matrix of the final set of papers. It shows whether a paper investigates fakery from a general or a platform-specific perspective, what method it applies, and with which framework category it is associated.

Reference	Focus	Method	Category
Adomavicius et al. (2013)	Platform	Empirical (Experiment)	Receiver
Algarni et al. (2017)	Platform	Empirical (Observational Data)	Receiver
Ananthakrishnan et al. (2020)	Platform	Empirical (Experiment)	Countermeasures
Anderson and Simester (2014)	Platform	Empirical (Observational Data)	Sender

Table A2.2: Concept Matrix of Literature on Fakery (cont.)

Reference	Focus	Method	Category
Appan and Browne (2012)	General	Empirical (Experiment)	Receiver
Armstrong et al. (1979)	General	Empirical (Experiment)	Countermeasures
Attas (1999)	General	Conceptual	Countermeasures
Barone and Miniard (1999)	General	Empirical (Experiment)	Consequences
Beisecker et al. (2024)	Platform	Empirical (Experiment)	Receiver
Benjamin and Raghu (2023)	Platform	Design	Countermeasures
Candogan and Drakopoulos (2020)	Platform	Theoretical	Countermeasures
Cantarella et al. (2023)	Platform	Empirical (Observational Data)	Consequences
Carson et al. (1985)	General	Conceptual	Countermeasures
Chaxel (2022)	General	Conceptual	Receiver
Chen et al. (2022)	Platform	Theoretical	Sender / Consequences
Chen and Papanastasiou (2021)	Platform	Theoretical	Sender / Countermeasures
Chen et al. (2015b)	Platform	Theoretical	Countermeasures
Cheung et al. (2012)	Platform	Empirical (Survey)	Receiver
Cho et al. (2011)	Platform	Empirical (Experiment)	Consequences
Chua et al. (2007)	Platform	Empirical (Case Study)	Countermeasures
Clarke et al. (2021)	Platform	Empirical (Observational Data)	Consequences / Countermeasures
Cohn et al. (2022)	General	Empirical (Experiment)	Receiver
Cowley and Janus (2004)	General	Empirical (Experiment)	Receiver
Craig et al. (2012)	General	Empirical (Experiment)	Receiver / Countermeasures
Darke et al. (2010)	General	Empirical (Experiment)	Receiver / Consequences
Darke and Ritchie (2007)	General	Empirical (Experiment)	Consequences
Davidson III and Worrel (1988)	General	Empirical (Observational Data)	Consequences
Dellarocas (2006a)	Platform	Theoretical	Consequences / Countermeasures
Deng and Chau (2021)	Platform	Empirical (Experiment)	Receiver
Deng et al. (2021)	Platform	Empirical (Observational Data)	Consequences
Donaker et al. (2019)	Platform	Conceptual	Countermeasures

Table A2.2: Concept Matrix of Literature on Fakery (cont.)

Reference	Focus	Method	Category
Edelman (2009)	Platform	Conceptual	Countermeasures
Gaeth and Heath (1987)	General	Empirical (Experiment)	Receiver / Countermeasures
Gardner (1975)	General	Conceptual	Countermeasures
George et al. (2018)	General	Empirical (Experiment)	Receiver
George et al. (2021)	General	Conceptual	Sender / Receiver / Consequences / Countermeasures
Gimpel et al. (2021)	Platform	Empirical (Experiment)	Countermeasures
Hamby et al. (2020)	General	Empirical (Experiment)	Receiver / Countermeasures
Harrison (2018)	Platform	Empirical (Observational Data / Survey)	Receiver
He et al. (2022b)	Platform	Empirical (Observational Data)	Sender / Consequences / Countermeasures
Heese et al. (2022)	General	Empirical (Observational Data)	Countermeasures
Ho et al. (2016)	Platform	Empirical (Experiment)	Countermeasures
Horner et al. (2021)	Platform	Empirical (Experiment)	Receiver
Jabr (2022)	Platform	Empirical (Observational Data)	Countermeasures
Jensen et al. (2013)	Platform	Empirical (Experiment)	Receiver
Jia et al. (2020)	Platform	Empirical (Observational Data)	Consequences
Jin et al. (2023)	Platform	Theoretical	Sender / Countermeasures
Johar (2022)	General	Conceptual	Receiver / Countermeasures
Johar (1996)	General	Empirical (Experiment)	Consequences
Johar and Roggeveen (2007)	General	Empirical (Experiment)	Countermeasures
Johar (1995)	General	Empirical (Experiment)	Consequences
Kartal and Tyran (2022)	General	Theoretical / Empirical (Experiment)	Receiver / Consequences
Keppo et al. (2022)	General	Theoretical	Sender
Khan et al. (2022a)	Platform	Conceptual	Sender / Countermeasures
Kim and Dennis (2019)	Platform	Empirical (Experiment)	Receiver / Countermeasures
Kim et al. (2019)	Platform	Empirical (Experiment)	Receiver / Countermeasures
King and Auschaitrakul (2020)	General	Empirical (Experiment)	Receiver

Table A2.2: Concept Matrix of Literature on Fakery (cont.)

Reference	Focus	Method	Category
King et al. (2021)	Platform	Theoretical / Empirical (Observational Data)	Countermeasures
Kirmani and Zhu (2007)	General	Empirical (Experiment)	Receiver
Kogan et al. (2023)	Platform	Empirical (Observational Data)	Consequences
Kokkodis et al. (2022)	Platform	Empirical (Observational Data)	Countermeasures
Kronrod et al. (2023)	Platform	Empirical (Observational Data / Experiment) / Design	Countermeasures
Kumar et al. (2019)	Platform	Design	Countermeasures
Kumar et al. (2018)	Platform	Design	Countermeasures
Laato et al. (2020)	Platform	Empirical (Survey)	Receiver
Lamy (2023)	General	Conceptual	Countermeasures
Lappas et al. (2016)	Platform	Empirical (Observational Data)	Consequences / Countermeasures
Law et al. (1998)	General	Empirical (Experiment)	Receiver
Lee et al. (2018b)	Platform	Empirical (Observational Data)	Sender
London Jr et al. (2022)	Platform	Empirical (Observational Data / Experiment)	Receiver
Luca and Zervas (2016)	Platform	Empirical (Observational Data)	Sender
Ludwig et al. (2016)	Platform	Empirical (Observational Data)	Countermeasures
Mayzlin (2006)	Platform	Theoretical	Sender
Mayzlin et al. (2014)	Platform	Theoretical / Empirical (Observational Data)	Sender / Countermeasures
Miller et al. (2024)	Platform	Empirical (Experiment)	Receiver
Moravec et al. (2020)	Platform	Empirical (Experiment)	Countermeasures
Moravec et al. (2022)	Platform	Empirical (Experiment)	Receiver / Countermeasures
Moravec et al. (2019)	Platform	Empirical (Experiment)	Causes (Receiver) / Countermeasures
Mostagir and Siderius (2023a)	General	Theoretical	Consequences
Mostagir et al. (2022)	General	Theoretical	Receiver
Mostagir and Siderius (2023b)	Platform	Theoretical	Sender / Countermeasures
Mostagir and Siderius (2022)	General	Theoretical	Consequences
Mullainathan and Shleifer (2005)	General	Theoretical	Sender

Table A2.2: Concept Matrix of Literature on Fakery (cont.)

Reference	Focus	Method	Category
Murphy et al. (2009)	General	Empirical (Observational Data)	Consequences
Ng et al. (2023)	Platform	Design	Countermeasures
Ng et al. (2021)	Platform	Empirical (Observational Data)	Countermeasures
Nie et al. (2022)	Platform	Empirical (Observational Data)	Sender
Nikitkov and Bay (2008)	Platform	Empirical (Observational Data)	Receiver
Oh et al. (2013)	Platform	Empirical (Observational Data)	Receiver
Papanastasiou (2020)	Platform	Theoretical	Countermeasures
Park et al. (2023)	Platform	Empirical (Observational Data)	Sender / Consequences
Pennycook et al. (2020)	Platform	Theoretical / Empirical (Experiment)	Countermeasures
Piccolo et al. (2018)	General	Theoretical	Countermeasures
Prawesh and Padmanabhan (2014)	Platform	Theoretical / Design	Receiver / Countermeasures
Pu et al. (2022)	Platform	Theoretical	Sender / Consequences / Countermeasures
Qiao and Rui (2023)	Platform	Empirical (Observational Data / Experiment)	Sender / Consequences / Countermeasures
Rabin and Schrag (1999)	General	Theoretical	Receiver
Rao (2022)	Platform	Empirical (Observational Data)	Consequences
Rao and Wang (2017)	General	Empirical (Observational Data)	Consequences
Riquelme and Román (2014)	Platform	Empirical (Survey)	Receiver
Rockmann and Northcraft (2008)	Platform	Empirical (Experiment)	Sender
Roggeveen and Johar (2002)	General	Empirical (Experiment)	Receiver
Román (2010)	Platform	Empirical (Survey)	Consequences
Ross et al. (2019)	Platform	Theoretical	Consequences
Sadler (2021)	General	Theoretical	Consequences
Schuetz et al. (2021)	Platform	Empirical (Survey)	Countermeasures
Serra-Garcia and Gneezy (2021)	General	Empirical (Experiment)	Receiver

Table A2.2: Concept Matrix of Literature on Fakery (cont.)

Reference	Focus	Method	Category
Sher (2011)	General	Conceptual	Sender / Countermeasures
Shi et al. (2022)	Platform	Empirical (Observational Data)	Receiver / Consequences
Shirish et al. (2021)	General	Empirical (Observational Data)	Receiver
Siering and Janze (2019)	Platform	Empirical (Observational Data)	Sender
Siering et al. (2016)	Platform	Design	Countermeasures
Skurnik et al. (2005)	General	Empirical (Experiment)	Receiver
Song et al. (2019)	Platform	Empirical (Observational Data)	Consequences
Stanley et al. (2022)	General	Conceptual	Receiver / Countermeasures
Tergiman and Villeval (2023)	General	Empirical (Experiment)	Consequences
Tipton et al. (2009)	General	Empirical (Observational Data)	Consequences
Trendel et al. (2018)	General	Empirical (Experiment)	Countermeasures
Turel and Osatuyi (2021)	Platform	Empirical (Experiment)	Receiver
Twyman et al. (2020)	General	Empirical (Experiment)	Receiver
Ullah et al. (2014)	General	Empirical (Observational Data)	Consequences
Van Bommel (2003)	General	Theoretical	Consequences
Van Roy and Yan (2010)	Platform	Theoretical	Countermeasures
Wang et al. (2021a)	Platform	Empirical (Observational Data)	Countermeasures
Wang et al. (2022)	Platform	Empirical (Observational Data / Experiment)	Countermeasures
Wei et al. (2022)	Platform	Design	Countermeasures
Weinmann et al. (2022)	Platform	Empirical (Experiment)	Countermeasures
Wilbur and Zhu (2009)	Platform	Theoretical	Consequences / Countermeasures
Wiles et al. (2010)	General	Empirical (Observational Data)	Consequences
Wilson et al. (2022)	Platform	Empirical (Experiment)	Countermeasures
Wöhner et al. (2015)	Platform	Design	Countermeasures
Wu and Geylani (2020)	General	Theoretical	Countermeasures
Xiao and Benbasat (2015)	Platform	Empirical (Experiment)	Receiver / Countermeasures

Table A2.2: Concept Matrix of Literature on Fakery (cont.)

Reference	Focus	Method	Category
Xiao and Benbasat (2011)	Platform	Conceptual	Consequences
Xie et al. (2022)	General	Empirical (Experiment)	Consequences
Xie et al. (2015)	General	Empirical (Experiment)	Consequences
Xu et al. (2012)	General	Empirical (Experiment)	Sender
Yip and Schweitzer (2016)	General	Empirical (Experiment)	Sender
Zhang et al. (2016)	Platform	Design	Countermeasures
Zhang et al. (2022)	Platform	Design	Countermeasures
Zhao et al. (2013)	Platform	Theoretical / Empirical (Observational Data)	Consequences
Zhou et al. (2004)	General	Design	Countermeasures

3. Extent and Implications of Fake Followers

3.1. Introduction

Despite the negative public attention to the practice of boosting social media popularity with *fake followers*, there is little research on this topic, as identified in the literature review. Fake followers are accounts created to artificially increase follower counts, purchased by companies to appear more popular than they actually are. If one is to believe the media, fake follower use is a widespread issue that deceives investors (e.g., Browne 2018, Jacobs 2018). The New York Times reported about a “follower factory,” which was accused of having sold more than 200 million fake followers to boost their clients’ follower counts (Confessore et al. 2018). The FTC has gone so far as to propose new rules that would prohibit selling or buying fake followers (Federal Trade Commission 2023).

These speculations are in line with the information systems literature that establishes a strong link between firms’ activities on social media platforms such as Twitter, Instagram, or TikTok and their valuation (Chung et al. 2020, Deng et al. 2018, Lis and Neßler 2014, Nofer and Hinz 2015). Recent years have shown that social media can have important positive effects and that online reputation plays a major role, for instance, driving a considerable portion of sales, but also has adverse effects, such as outrage or backlash (Gutt et al. 2019, Luo et al. 2013, Song et al. 2019, Teubner et al. 2016, 2020, Wang et al. 2021b). It is not surprising, therefore, that investors pay close attention to firms’ social media activities (Deng et al. 2018, Kim and Youm 2017, Luo et al. 2013, Nofer and Hinz 2015). In the context of fake followers, it becomes crucial to understand whether they are used to strategically influence popularity and investor behavior.

To connect the public debate on fake followers with the existing academic understanding of social media influence, this study aims to understand the true magnitude of fake follower use – especially from the eye of the investor. Two main research questions emerge. First, are large parts of firms’ follower counts inflated by fake followers, or is it only a minor fraction? Second, from a risk perspective, how much of a loss in shareholder value should investors expect if it becomes evident that a firm purchased fake followers? These questions are essential as they help to think more clearly about potential countermeasures.

Based on the assessment of existing research, no study has yet attempted to address these questions. Prior work documented various forms of corporate misconduct on social media, including sentiment manipulation, spreading of fake news, forged reviews, spam, and ad fraud, but not fake follower use (e.g., Hawlitschek et al. 2018, Lee et al.

2018b, Nie et al. 2022, Wang et al. 2022, Wei et al. 2022, Xiao and Benbasat 2011). It is also challenging to study fake follower use because distinguishing between fake and real followers is usually impossible without insider knowledge. Some research has proposed algorithms to identify fake followers, but these show high error rates (Gallwitz and Kreil 2021, Rauchfleisch and Kaiser 2020).

To address these questions, the use and risks of fake followers are grounded in theory. Following the theory of attention economics, firms compete for attention and, therefore, are incentivized to buy fake followers to stand out (Davenport and Beck 2001, Goldhaber 1997, Simon 1971). Especially those firms that have a greater need to compete for attention—those in competitive industries or small firms—should be buying more fake followers. Along the same lines, investors will pay close attention to a firm's social media practices. If they learn that a firm's followership is (partially) fake, this lowers their expectations of its expected profits.

The empirical design follows a panel data study with a fixed-effects panel regression and a stock market event study. To overcome the hurdles in distinguishing fake and genuine followers, the study exploits a unique event on the Twitter platform. Over two days in July 2018, Twitter removed tens of millions of fake accounts from the platform responsible for approximately 6% of the overall follower count (Confessore and Dance 2018). Using detailed daily data on stock market-listed firms, the use of fake followers is estimated from the follower counts immediately before and after the purge. Then, a stock market event study assesses how shareholders reacted to the revelation of firms' use of fake followers. This setup is particularly valuable because it allows to overcome the challenges of detecting fake followers (Rauchfleisch and Kaiser 2020). In addition, Twitter's follower purge was immediate, exogenous to firms, and came as a surprise. There is some reason to believe that Twitter acted honest in removing fake followers due to a pending FTC investigation (Confessore and Dance 2018).

The results paint a picture that differs from the debate around fake followers in the media and prior research. Although there is statistically significant and robust evidence that fake followers inflate the average follower count, only a small fraction of about 1.17% of firms' followers are fake. In addition, while firms in competitive industries and small firms – as suggested by the theory of attention economics – have a larger share of fake followers than the average firm, the effect remains overall still modest (between 1.4%-1.6%).

In line with the theory of attention economics, evidence supports the hypothesis that stock markets react negatively when they learn that firms use fake followers. A 1% higher share of fake followers is associated with negative cumulative abnormal returns of approximately -0.078%. An empirical extension of the analysis explores why investors reacted negatively. Considering the tweet sentiment around the purge, there is no support that investors' reaction is linked to concerns over reputational damage arising for the firm. Instead, there is evidence that investors interpret fake follower use as new information about firms' social media reach. In this regard, a lower follower count is associated with less influence. The findings remain consistent across various robustness checks, including quasi-experimental evidence based on a difference-in-differences design. These numbers show that fake follower use presents a relatively minor challenge on digital platforms with limited shareholder risk.

The chapter is structured as follows. Section 3.2 provides the theoretical foundation from established theories and prior research. Section 3.3 describes the Great Purge on Twitter as well as the details of the data collection and empirical analysis. Section 3.4 presents the results. Section 3.5 discusses the implications of the study.

3.2. Theoretical Background

3.2.1. Related Research

Two primary streams of research inform the study. A first stream of research seeks to understand the drivers and extent of deceptive behaviors by firms on social media (e.g., Bello Rinaudo et al. 2022, Hawlitschek et al. 2018, Luca and Zervas 2016, Mayzlin et al. 2014). Various practices have been investigated, including fake news and claims (e.g., Hawlitschek et al. 2018, Kim and Dennis 2019, Wang et al. 2022, Wei et al. 2022), forged reviews (e.g., Kumar et al. 2018, Luca and Zervas 2016, Nie et al. 2022), fake accounts (e.g., Ferrara et al. 2016, Huang and Liu 2023), or sentiment manipulation (e.g., Dellarocas 2006b, Lee et al. 2018b). Some research studied so-called "bots" (e.g., Bessi and Ferrara 2016, Hagen et al. 2022). In understanding the behavior of these bots on Twitter, Salge et al. (2022) found that bots amplify information existing on the platform by further dissemination, but also that they obtain information from other sources and spread it on the platform. Other studies focus on the negative aspects of bots and mostly investigate the manipulation of opinions in public discussions (Bessi and Ferrara 2016, Hagen et al. 2022). Amongst these, only a few studies investigate fake followers. Other studies are in the domain of computer science and seek to detect and describe fake accounts using metadata, social network structure, friends, tweeting behavior and content, or sentiment (e.g., Benjamin and Raghu 2023, Cresci et al. 2015, Rauchfleisch and Kaiser 2020, Sayyadiharikandeh et al. 2020). One exception and perhaps closest to this study are Huang and Liu (2023), who develop a theoretical model of influencers' incentives to purchase fake followers. By contrast, this study is empirical and about firms. Also, Silva and Proksch (2021) study fake followers on Twitter but focus on politicians and do not study investor reactions.

A second related stream of work studied the link between firms' social media engagement and various outcomes of firm performance. Firms use social media to reach out to consumers and promote products (e.g., Chen et al. 2015a, Chevalier and Mayzlin 2006, Lis and Neßler 2014, Luca 2011), to build brands, and to foster sales through interaction (e.g., Chung et al. 2020, Hennig-Thurau et al. 2015, Wang et al. 2021b). In this context, studies observe associations between firms' social media engagement and sales, firm equity, and stock prices, but also show the relevance of consumer engagement for firm performance (Gutt et al. 2019, Jabr and Zheng 2014, Kim and Youm 2017, Kumar et al. 2016, Luo et al. 2013, Nofer and Hinz 2015, Teubner et al. 2016). In this context, research highlights the role of follower count (Hinz et al. 2011, Gelper et al. 2021, Wies et al. 2023). However, existing literature does not yet understand to what degree firms' use of fake followers influences this link.

3.2.2. Fake Follower Use and Investor Reactions

The theory of attention economics suggests that individuals' attention is scarce. Accordingly, firms that stand out from their competitors will be more successful (Davenport and Beck 2001, Simon 1971). Social media is an important broker of attention (e.g., Chen et al. 2015a, Hennig-Thurau et al. 2015, Lis and Neßler 2014, Nofer and Hinz 2015). For example, social media broadcasting positively influences sales (Chen et al. 2015a) and positive online reviews affect consumers' purchasing behavior (e.g., Chevalier and Mayzlin 2006, Luca and Zervas 2016, Teubner et al. 2016). In light of these arguments, firms likely have the incentive to purchase fake followers. More followers can signal greater quality (Huang and Liu 2023) and are important for seeding strategies (Gelper et al. 2021, Hinz et al. 2011). Therefore, firms have the incentive to build large online followerships, and fake followers can be a means to increase them (Cha et al. 2010, Cresci et al. 2015).

Based on these considerations, some firms should use fake followers to a larger extent, especially firms facing greater industry competition and smaller firms. Regarding industry competition, individuals' connections on social media are driven by homophily—they are more likely to connect to others that are similar (McPherson et al. 2001). In the business context, this translates to heterogeneous susceptibility to firms' attempts to enhance attention, with similar susceptibility across similar firms, i.e., firms of the same industry—this strengthens the need to allocate attention (Ackoff 1989, Simon 1971). Some evidence supporting this claim comes from prior research, finding that restaurants, hotels, and moviemakers engage more in fraudulent behavior with increasing competition (Lee et al. 2018b, Luca and Zervas 2016, Mayzlin et al. 2014).

The theory of attention economics also suggests that smaller firms have a greater incentive to use fake followers for two reasons. First, smaller firms face resource constraints regarding marketing practices (Carson 1985). Second, smaller firms are at a visibility and trust disadvantage resulting from a lower sales volume and fewer employees and customers (Carson 1985). Thus, fake followers provide an easy and affordable means to enhance attention. In contrast, large firms are more known, usually have more organic followers, and face higher risks of fraudulent behavior (Carson 1985, Mayzlin et al. 2014). This is indirectly supported by empirical evidence. For example, studies on fraud indicate higher manipulation for small business owners and low-budget movies, i.e., when resource constraints are present (Lee et al. 2018b, Luca and Zervas 2016, Mayzlin et al. 2014).

Investors decide about their capital allocation based on the expectations about the future value of a firm (Graham and Zweig 2003, Loibl and Hira 2009). To proxy for future value, expectations incorporate future profits and risks associated with a firm (Fama et al. 1969, Fama 1970). In decision-making, investors consider fundamental and technical data, but also information from the media, e.g., news coverage, search engine data, or social media (Hirshleifer and Teoh 2009). This aligns with studies finding an association between social media and stock markets (Deng et al. 2018, Kim and Youm 2017, Luo et al. 2013, Nofer and Hinz 2015).

Following the theory of attention economics, investors will also incorporate information about firms' use of fake followers. In particular, investors will react negatively when it becomes evident that firms rely on fake followers. This is for two mechanisms, labeled as revealed influence and reputational damage. First, investors react negatively because the information changes their perception of that firm's (quantitative) social media reach. When firms' use of fake followers is exposed, it becomes apparent that the perceived reach has been manipulated (Cha et al. 2010, Cresci et al. 2015). After the revelation, investors have information about the actual follower count and the extent of manipulation. Given the importance of social media for firm performance (Chen et al. 2015a, Kim and Youm 2017, Kumar et al. 2016, Lis and Neßler 2014), investors reduce their valuation. In other words, when fake follower use is exposed, it reveals a more accurate picture of firms' social media followership, thereby influencing investors' valuation. Second, according to a reputational damage mechanism, investors react negatively because the information that a firm has used fake followers negatively affects the firm's (qualitative) standing on social media. Social media users can express their dissatisfaction from perceived betrayal on social media, which can be observed by investors (Greve et al. 2010, Matook et al. 2022). As a result, the revelation of fake followers should manifest in investors' valuation. There is some evidence backing this mechanism. Reputational damages have been driving, for instance, stock market reactions to celebrity endorsement scandals (Knittel and Stango 2014), data breaches

(Martin et al. 2017), or regulatory investigations (Jain et al. 2010).

3.3. Method and Data

3.3.1. Empirical Setting: Twitter and the *Great Purge*

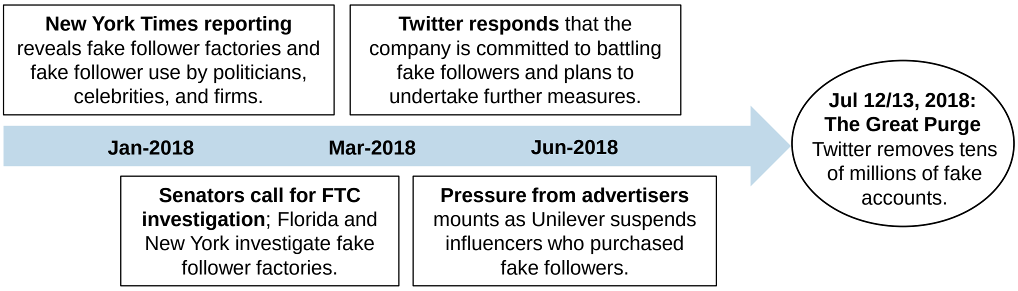
The empirical study is set on the Twitter platform, where users share short texts called tweets. Other users can interact with these tweets but can also become *followers* to receive updates on another user's activities. Twitter is well-suited for investigating fake followers for three reasons. First, research has documented the considerable influence of Twitter (Hennig-Thurau et al. 2015, Kumar et al. 2016). Second, investors pay close attention to information on Twitter (Twitter Data 2016, Morse 2016), and popular trading platforms, including Bloomberg, and news media report information from Twitter (Bloomberg 2015). Third, follower counts are crucial for firms on Twitter (Cha et al. 2010, Wies et al. 2023).

The event considered for this study is the *Great Purge*: On July 11, 2018, Jack Dorsey, the Twitter CEO, announced the removal of user accounts that showed "suspicious activity" making it by far the most extensive intervention so far carried out over two days (Confessore and Dance 2018). On July 12 and 13, Twitter closed tens of millions of user accounts. This so-called Great Purge received considerable coverage in international media.

Technically, Twitter removed so-called "locked accounts". The purged accounts belonged to different large providers and were "created only to simulate a static audience" (Social Puncher 2018, p. 23). They often had blank profiles, zero tweets, and followed a large number of accounts. In other words, these accounts were not *bots* that interacted with other users but simply served the purpose of being fake followers.

Twitter's Great Purge represents a unique empirical opportunity for three reasons. First, the purge is a unique chance to reduce measurement errors that plague the study of fake accounts. Although prior research has developed advanced algorithms to distinguish fake followers from real users, these algorithms have been shown to have a considerable rate of type II errors (Gallwitz and Kreil 2021, Rauchfleisch and Kaiser 2020). Second, the event serves as a valid empirical shock because Twitter did not announce its course of action beforehand. Thus, this design can refute concerns over investors or firms anticipating the purge and, therefore, changing their behavior. Finally, Twitter had a great interest in acting honestly. As shown in Figure 3.1, some weeks before the purge, U.S. Senators had formally asked the FTC for an investigation of fake followers on Twitter after increasing public scrutiny. Therefore, Twitter had a great interest in acting honestly and removing fake followers to the greatest degree possible.

Figure 3.1: Timeline of Twitter's Great Purge
Note: The figure describes the events preceding the Great Purge in 2018. In January, The New York Times reported about a company that sold fake Twitter followers to several politicians, celebrities, and firms. In February, U.S. Senators formally asked the FTC for an investigation. In March, Twitter highlighted its commitment to reducing fake followers on the platform. On July 12 and 13, Twitter carried out the Great Purge.


New York Times reporting reveals fake follower factories and fake follower use by politicians, celebrities, and firms.
Twitter responds that the company is committed to battling fake followers and plans to undertake further measures.
Jan-2018
Mar-2018
Jun-2018
Jul 12/13, 2018: The Great Purge Twitter removes tens of millions of fake accounts.
Senators call for FTC investigation; Florida and New York investigate fake follower factories.
Pressure from advertisers mounts as Unilever suspends influencers who purchased fake followers.

3.3.2. Data

To construct the dataset, the sample departs from the S&P 1500 Composite Index, which comprises the top 1,500 public U.S. firms by market capitalization. This index is suitable as (1) it covers a broad range of firms across sizes and industries, (2) securities show variation in prices and trading, (3) it covers around 90% of the U.S. market capitalization, and (4) the S&P has been used in other research (e.g., Kim and Youm 2017, Lee et al. 2017). Identifying the corresponding Twitter account for each firm in the index underlies a manual inspection. The Twitter account is primarily determined from a link to a firm's Twitter account on its global website. If no global website is available, the U.S. website serves as the source for the link. If no link is available, the Twitter account is identified from a manual search on the platform. The number of accounts is restricted to one account per firm: Twitter accounts in the dataset present the global or U.S. account representing the whole company, not subsidiaries or departments. Twitter accounts are further dismissed if they did not exist at the beginning of the observation period or have no logos or pictures. As not all firms operate a Twitter account, Twitter handles are identified for 1,226 firms.

The data collected for each firm in the index consists of the following. First, following existing research, Social Blade is used to obtain follower data (Lin et al. 2022, Sjöblom et al. 2019). For approximately 7.3% of the firm-day observations, Social Blade data is incomplete. The primary strategy to deal with missing values is to follow the standard approach and record them as "missing", thereby ignoring rows with missing data in the analysis (Peng et al. 2023). The findings are robust to the alternative approach of *multiple imputation* following Peng et al. (2023). Next, via the research access to the Twitter Full Archive Search API, firms' tweets are obtained. The API provides a historical archive of all tweets ever posted unless deleted by the user or due to Twitter policies, therefore not subject to criticism of the tweet sampling with the Streaming API (Gerlitz and Rieder 2013). The set of relevant tweets excludes retweets to avoid oversampling and tweets with images and videos for analysis reasons. This results in more than 9.9 million tweets for the observation period. Last, firms' abnormal stock market returns are extracted from CRSP via WRDS. This data source also allows access to fundamental data on firms, i.e., the volume of assets, firm sector, EBIT, and liquidity. By only considering firms in the sample for whom data is consistently available during the observation period, the final dataset contains 839 firms.

The resulting dataset is a firm-day panel that comprises 252,539 observations based on 839 firms over 150 days before and after the purge.

3.3.3. Variables

Table 3.1 defines the variables and reports their descriptive statistics. *Followers* is the total number of followers of firm i on day t. The variable is log-transformed to account for skewness. *Share fake* is the proportion of followers of firm i deleted during the Great Purge relative to that firm's total number of followers before the purge, and 0 in case no followers were removed.

Competitive intensity and firm size define the moderators in the study. *Competitive intensity* is assessed based on the inverted Herfindahl-Hirschman index (HHI) (Kwieciński 2017, Rhoades 1993). The HHI is widely used and available for all industries, which makes it advantageous over other measures that require surveying firms directly. The HHI is calculated by squaring the market shares of firms in an industry and then summing them up (Rhoades 1993). The HHI is obtained from the U.S. Census Bureau's Economic Census, where it is reported every five years for the year preceding the purge, i.e., 2017. *Firm size* is assessed by following existing research, namely by inferring it from the volume of assets (Campbell and Shang 2022, Shalit and Sankar 1977, Yang et al. 2012). The variable is log-transformed to account for skewness and inverted, i.e., multiplied with -1, for consistency with the hypothesis.

Several further variables serve as independent variables or controls. *After purge* is a binary that is 0 on days before the purge (i.e., before July 12) and 1 afterward. *Sector* is the (industry) sector of firm i as based on the two-digit sector description of the NAICS code. *EBIT* are the Earnings before Interest and Taxes of firm i. The variable is log-transformed to account for skewness. The variable *Cash*, which is the total cash of

Table 3.1: Variables and Descriptive Statistics
Note: The table describes the data set along the variables over the observation period. It relies on 252,539 firm-day observations based on 839 firms over 150 days before and after the Great Purge.

Variable	Description	Mean	SD	Min	Median	Max
Followers [in Thousands]	Number of followers of firm i on day t	373.846	2,866.327	0.032	11.673	62,950.238
Share fake	Share of followers of firm i removed during the Great Purge	0.012	0.025	0.000	0.008	0.370
CAR	Cumulative abnormal stock returns (CAR) of firm i for [0,1] of the Great Purge	-0.003	0.025	-0.196	-0.003	0.105
Competitive intensity	Normalized and inverted Herfindahl-Hirschman Index (2017) for firm i based on the NAICS code	0.519	0.268	0.000	0.507	0.995
Firm size	Total volume of assets of firm i in US$ Bn (2018)	40.362	171.019	0.024	6.184	2,622.532
After purge	1 if t is after the purge, 0 otherwise	0.502	0.500	0.000	1.000	1.000
EBIT	Earnings before Interest and Taxes of firm i in US$ Bn (2018)	2.151	5.511	-0.253	0.503	70.662
Cash	Cash of firm i in US$ Bn (2018)	1.705	6.154	0.000	0.314	130.547

firm i, accounts for liquidity. The variable is log-transformed to account for skewness. Twitter sentiment about firm i is measured based on the tweets that mention a firm's name using a hashtag (e.g., #Amgen, #Boeing), following extant work (Bessi and Ferrara 2016, Lee et al. 2018b). VADER, a robust approach for tweets relying on lexical, grammatical, and syntactical structures (Deng et al. 2022, Hutto and Gilbert 2014, Zhang and Luo 2023), is used to assess the sentiment of each tweet. It assigns each tweet a continuous value between -1 and 1, with 0 representing a neutral sentiment. Continuous scaling has the advantage of capturing even tiny sentiment changes. Tests show that the obtained measure is highly similar to if obtained using alternative sentiment algorithms, making the results not dependent on the sentiment measure (see Appendix Table A3.1).

3.3.4. Empirical Framework and Model

This study uses two separate estimation models. To investigate fake follower use by firms, the changes in followers caused by the Great Purge are estimated with a fixed-effects panel regression model of the form:

$$Log(Followers)_{i,t} = \beta_0 + \beta_1 \times After\ purge_t + \theta_i + \tau_t + \kappa_{i,t} + \epsilon_{i,t} \tag{3.1}$$

whereby i iterates over firms and t over days, with β_1 being the coefficient of interest.

To address concerns over omitted variables or pre-existing heterogeneity, the model includes firm fixed effects θ_i (Wooldridge 2019). Firm-level fixed effects cancel out any difference across firms that is static within the observation period, for instance, their industry or business model. Moreover, the model adds the variable list τ_t, which contains time-level fixed effects in terms of dummies for day-of-week, week-of-year, and month-of-year to control for time-invariant heterogeneity constant within these time units (Wooldridge 2019). Moreover, one could be concerned that there could be a change in the dependent variable due to fluctuations over time. To adjust for a trend in followers, the model includes the control κ in terms of a linear time trend. The primary analysis considers changes over an observation period of 30 days before and after the purge, but the results remain robust to this choice, as documented in the robustness checks.

To understand the shareholder value effects of fake follower use, a stock market event study is conducted (Brown and Warner 1985, MacKinlay 1997, Sorescu et al. 2017). The idea is to estimate value effects caused by an event by estimating so-called abnormal returns, namely comparing the change in the stock price around the event date with the expected stock price had the event not taken place. The abnormal return $AR_{i,t}$ is the difference between the actual return $r_{i,t}$ and the expected return $\hat{r}_{i,t}$, i.e., the returns if the event would not have occurred. The abnormal returns are then cumulated over an *event window* to account for the fact that information about an event might leak to the stock markets slightly before an event or that some investors might react later. These so-called cumulative abnormal returns (CAR) are then $CAR_i = \sum_{t=t_1}^{t_2} AR_{i,t}$, where t_1 and t_2 represent the boundaries of the event window. For this study, the main effect relies on an event window of [0,1] around the purge to capture the immediate effects of the event, but the results for windows of various lengths are reported. The event window must be long enough to capture reactions to an event and short enough not to incorporate confounding events. The study follows the recommendation to choose relatively narrow event windows to account for a fast market reaction, but also for the fact that the purge was carried out over two days (July 12 to July 13) (Konchitchki and O'Leary 2011, Sorescu et al. 2017).

The expected return $\hat{r}_{i,t}$ is calculated using the Fama-French Three Factor model (FF3FM) because of its widespread use (Carhart 1997, Jegadeesh and Titman 1993, Sorescu et al. 2017). The FF3FM estimates the expected stock return based on regressing stock returns on the overall market returns $r_{m,t}$ and controls over the *estimation window* before the event (Brown and Warner 1985). More specifically, the Fama-French Three Factor model is $\hat{r}_{i,t} = r_{f,t} + \hat{\beta}_{1,i}(r_{m,t} - r_{f,t}) + \hat{\beta}_{2,i}(SMB_t) + \hat{\beta}_{3,i}(HML_t) + \hat{\epsilon}_{i,t}$ with $r_{f,t}$ as the risk-free rate of return, $r_{m,t}$ as the value-weighted return on all stocks in the NYSE, AMEX, and NASDAQ, SMB$_t$ as the difference in returns between small-cap vs. large-cap stocks, and HML$_t$ as the difference in returns between high book-to-market and low book-to-market stocks. The Fama-French Plus Momentum model, used to validate the findings, additionally includes a momentum effect MOM$_t$ to adjust for falling or rising tendencies. The consistency of the results with alternative models is reported in the robustness section. The estimation window accounts for 30 days with a gap between the estimation and event window of two days to maximize data availability; however, Sorescu et al. (2017, p. 203) outline that the "length of the estimation window is not likely to have a big impact on the final result."

The model then takes the form:

$$CAR_i = \beta_0 + \beta_1 \times Share\ fake_i + \kappa_i + \epsilon_i \tag{3.2}$$

where the coefficient of interest is β_1. The variable list κ_i contains the following controls obtained from Compustat to adjust for heterogeneity. First, *Sector* accounts for heterogeneous effects on the industry level (e.g., Bharadwaj et al. 1999, Foerderer and Schuetz 2022). Second, *Firm size* accounts for differences between larger and smaller firms (e.g., Dehning et al. 2003, Im et al. 2001). Third, *EBIT* and *Cash* adjust for performance differences between firms (e.g., Kohli et al. 2012).

3.4. Results

3.4.1. Fake Follower Use

Results and Interpretation

Prior to discussing the regression results, descriptive evidence is presented. Figure 3.2 plots the daily change in followers before and after the purge. The follower count is marginally changing before the purge. On the days of the purge (*t* and *t+1*), there is a sharp and marked decline in the growth of followers. The decline in followers is specific to the purge – no other marked changes are visible over the remainder of the observation period. Thus, the decline in followers is particular to this event and unlikely an artifact of any other event that occurred beforehand.

Figure 3.2: Descriptive Evidence of Fake Follower Use – Firms' Daily Change in Followers [in %]

Note: The figure plots the daily growth in *Followers* before and after the Great Purge. The dashed vertical denotes the day of the purge. The blue area denotes the estimation period of 30 days used for equation 3.1.

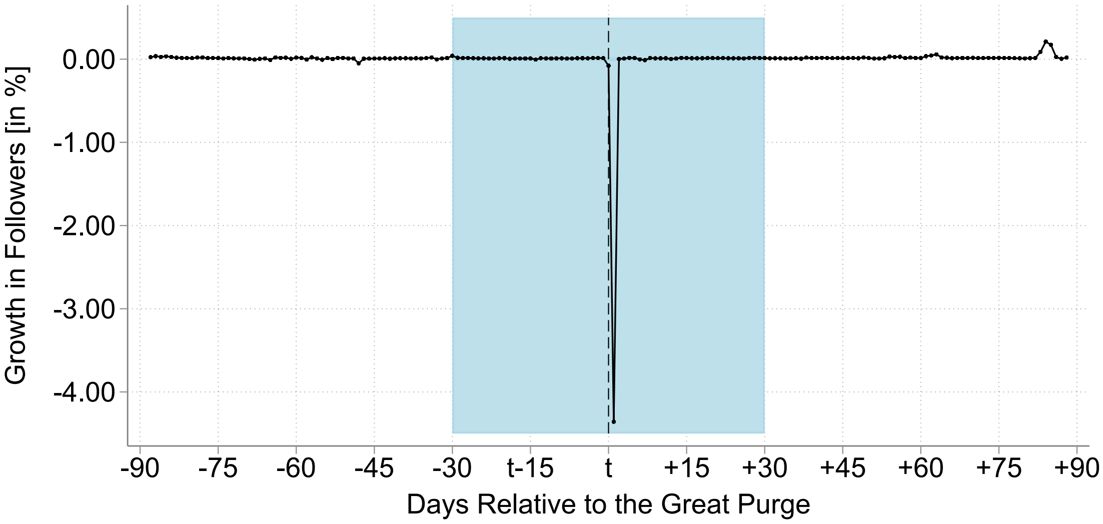


Table 3.2 shows the regression results. Column (1) shows the estimates for equation 3.1. The coefficient on *After purge* is negative and statistically significant. However, regarding the effect magnitude, only a small fraction of the average firm's followers are fake: 1.17% or approximately 12,000 followers on average. This is much less than estimated by the few existing studies of fake followers among all Twitter users and using fake follower detection algorithms, namely 15-23% (Bessi and Ferrara 2016, Hagen et al. 2022). Columns (2) and (3) explore whether firms in competitive industries and those that are smaller show greater fake follower use. In both columns, the coefficient

on the interaction term is negative and significant, which means that firms in competitive industries and those that are smaller use fake followers to a larger extent.

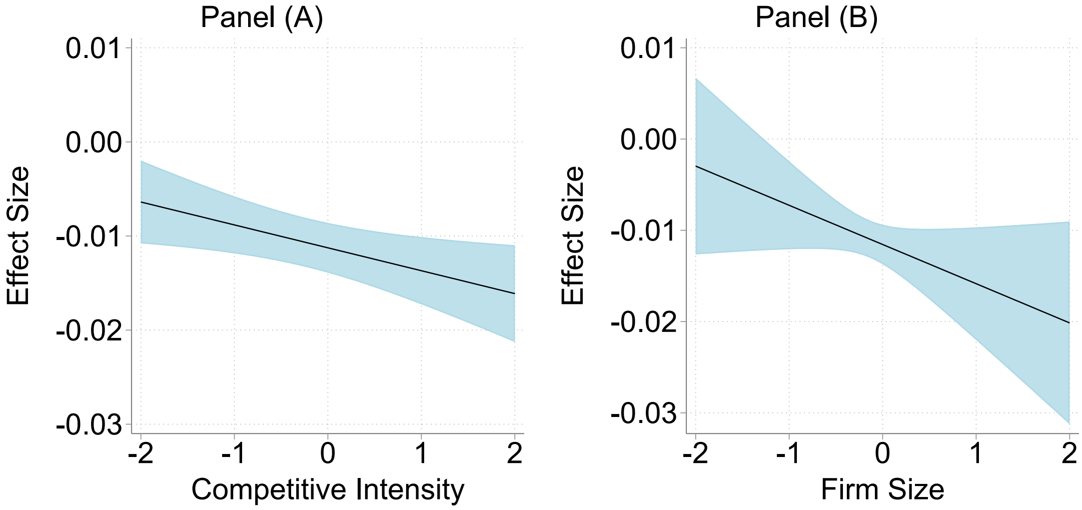
Table 3.2: Estimation of Fake Follower Use
Note: The table reports the test for fake follower use. Column (1) shows the estimates of equation 3.1. Column (2) adds *Competitive intensity* as a moderator. Column (3) adds *Firm size*. For comparability, both moderators are standardized to have a mean of 0 and a standard deviation of 1, and Columns (2) and (3) rely on observations within the 10% and 90% percentiles of *Log(Followers)*. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1) Log (Followers)	(2) Log (Followers)	(3) Log (Followers)
After purge	-0.012***	-0.011***	-0.012***
	(0.001)	(0.001)	(0.001)
After purge × Competitive intensity		-0.002*	
		(0.001)	
After purge × Firm size			-0.004^{+}
			(0.003)
Observations	51,011	28,305	40,688
Controls	X	X	X
Firm Fixed Effects	X	X	X
Time Fixed Effects	X	X	X

Figure 3.3 plots the interaction effects to ease the interpretation of the effect magnitudes. The horizontal axis shows the moderator variable, and the vertical axis shows the magnitude of fake follower use. The effects are reported together with 95% confidence intervals. Even when exploring these conditions for higher-than-expected effects, fake follower use is shown to be of modest magnitude. As of Panel (A), for firms facing a one-standard-deviation higher competitive intensity, the proportion of fake followers is 1.4% (i.e., the effect is only 0.2 percentage points higher compared to when competitive intensity is on its mean). Similarly, a one-standard-deviation smaller firm has 1.6% fake followers (i.e., the effect is only 0.4 percentage points higher than when on its mean).

The conclusion that fake follower use is modest among firms is also confirmed when considering individual firms in the sample. Only a few firms showed considerable fake follower use, including The Joint Corp (37.01%), Hologic Inc (36.02%), Cadence Design Systems Inc (19.59%), MSCI Inc (19.44%), and ICU Medical Inc (19.02%).

Figure 3.3: Interaction Plots for Competitive Intensity and Firm Size
Note: The figure shows the interaction plots for *Competitive intensity* and *Firm size* in Panel (A) and Panel (B), respectively. The horizontal axis gives the moderator variable centered on the mean (0), with the values indicating standard deviations from the mean. The vertical axis gives the share of fake followers, i.e., the estimate of the loss in *Followers* in the purge. The solid line thus shows the effect size conditional on the moderator variable, together with 95% confidence intervals.

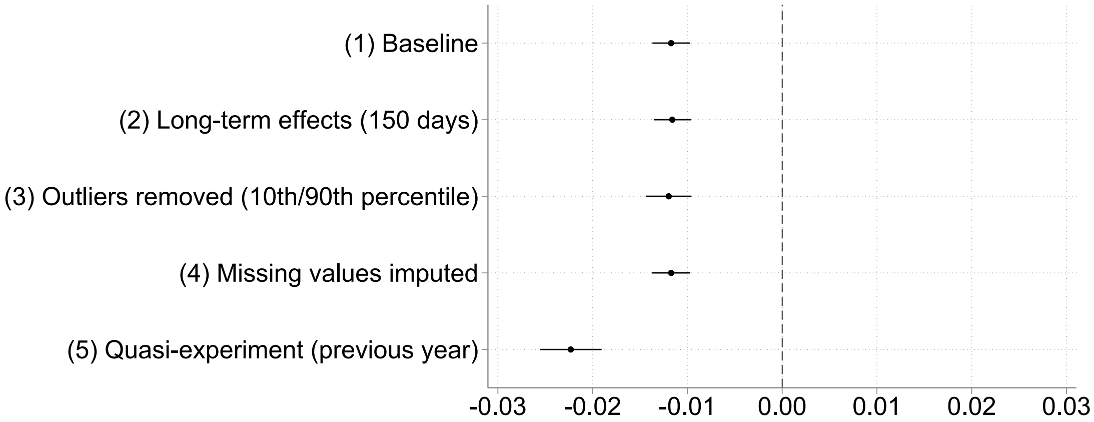


Robustness Checks

Several robustness checks support these conclusions. Figure 3.4 plots the estimate obtained from these checks together with 95% confidence intervals. Row (1) shows the baseline from Table 3.2 for comparison. Row (2) shows that results are not an artifact of the choice of the observation period. Using an extended observation period of 150 days (i.e., five months) around the purge, there is an almost identical effect (1.15%). Row (3) shows that a few firms with non-representative follower counts do not drive the results. Observations not between the 10% and 90% percentiles of *Log(Followers)* are removed from the dataset. The resulting estimate is significant and almost identical (1.19%). Row (4) documents that the results are robust to imputing missing observations for followers (1.16%) (Peng et al. 2023). Row (5) confirms the estimates from an alternative difference-in-differences quasi-experimental setup. To rule out omitted variable bias, the study would ideally rely on a quasi-experimental research design. However, there is no natural control group, as all Twitter users are equally affected. Similar platforms outside the U.S. differ in adoption and follower numbers. An option is to use an artificial control group from historical observations similar to others (e.g., Chen et al. 2020, Tafti et al. 2016). The effective number of firms that can be used in this design is smaller (455 firms) because Social Blade does not have data going back one year before the Great Purge for all firms. This analysis relies on the same set of firms with observations for the preceding year, assuming that follower numbers would have followed a similar trend, and conducts a difference-in-differences analysis that results in a similar estimate (2.21%) (Angrist and Pischke 2008).

Figure 3.4: Estimates of Fake Follower Use (β_1) for Various Robustness Checks

Note: The figure shows the estimates of fake follower use (i.e., β_1 of equation 3.1) as obtained from several robustness checks. Line (1) shows the baseline estimate. Line (2) relies on an extended observation window of 150 days before and after the purge. Line (3) excludes outliers of *Log(Followers)*. Line (4) imputes missing values of the dependent variable. Line (5) uses a difference-in-differences estimation from a historical control group. The markers denote the estimate, and the whiskers give the 95% confidence intervals. The dashed vertical denotes a coefficient of 0, i.e., a null effect. The estimates are documented in Appendix Table A3.2.



3.4.2. Stock Market Reactions to Fake Follower Use

Results and Interpretation

Table 3.3 summarizes the results. Column (1) reports the estimates for the main event window of [0,1]. The coefficient on *Share fake* is negative and significant. Column (2) adds controls, and the observed coefficient remains negative and significant. Column (3) uses a longer window of [0,3], which leads to a consistent coefficient. Column (4) uses the largest event window, and the coefficient is again negative and significant.

Based on these estimates, the shareholder value effects are minor. If 1% of a firm's followers are exposed to be fake, a firm's stock price declines by about 0.078%. In terms of total shareholder value loss, considering a median market capitalization of US$ 7.85 Bn before the purge, a firm with average fake follower use experiences a loss in market capitalization by US$ 7.11 Mn due to the purge (i.e., median market capitalization $\times$ average follower loss $\times$ β_1).

Table 3.3: Estimation of Stock Market Reactions To Fake Follower Use
Note: The table reports the test for the stock market reactions. Column (1) shows the estimates for equation 3.2 for the event window [0,1] without control variables. Column (2) adds control variables. Column (3) shows the estimates for the event window [0,3] with control variables. Column (4) shows the estimates for the event window [0,5] with control variables. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}\ p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

	(1) CAR [0,1]	(2) CAR [0,1]	(3) CAR [0,3]	(4) CAR [0,5]
Share fake	-0.078*	-0.078**	-0.091**	-0.135**
	(0.032)	(0.030)	(0.035)	(0.042)
Constant	-0.002*	0.001	0.026	0.074***
	(0.001)	(0.014)	(0.019)	(0.022)
Controls		X	X	X
Observations	807	691	691	691

Second, the effect magnitude can be compared to other events that harm firms' valuation. The reaction is comparable to losses firms encounter when announcing incidents damaging their internal operations. For example, breaches of customer data can be associated with cumulative abnormal returns between -0.15% to -0.29%, which would correspond to the reaction if approximately 2% to 3% of firms' followers are fake (Martin et al. 2017). Another example would be product recalls, which have been observed to be linked to cumulative abnormal returns of -0.18%, which would correspond to a fake follower share of about 2.3% (Liu et al. 2017).

Empirical Extension: Why Did Investors React Negatively?

The study suggested two explanations in line with the theory of attention economics that could have led investors to react negatively, labeled for brevity as revealed influence and reputational damage. Revealed influence means that investors reacted negatively because learning that a firm relied on fake followers provides the information that a firm's social media reach is less than expected. Reputational damage means that investors do not necessarily react negatively to the information about fake follower use but potentially ensuing backlash from social media users who learned about the misconduct. To explore both mechanisms, follow-up analyses test predictions that should hold if one of the mechanisms is present following Pierce et al. (2015).

To test for the revealed influence mechanism, it can be utilized that some accounts deleted during Twitter's Great Purge were restored several weeks after the purge (Dave 2018). Twitter granted purged account owners the right to restore their accounts by filing an appeal and providing proof (i.e., personal identity verification). Any such verified account was restored between October 3 and 5, 2018. If revealed influence explains the results, investors should react more positively the more of a firm's followers are restored. Method-wise, the event study is reimplemented following equation 3.2 but with estimates of CAR for the October event and the share of followers restored. Because the restoration took place over three days, the event window is [0,3].

To test for the reputational damage mechanism, user reactions on Twitter are inferred from changes in sentiment toward firms using tweets that mention firms in the form of hashtags (e.g., #Amgen, #Boeing) in line with prior research (e.g., Rust et al. 2021). Following Liu (2012), the analysis relies on the event study setup in equation 3.2 but estimates cumulative abnormal *sentiment* instead of returns. In line with Liu (2012), the counterfactual is estimated for each firm to calculate abnormal sentiment considering an estimation window of 30 days, a gap of two days, and an event window of [0,1] around the purge. Similar to the abnormal returns, $AS_{i,t}$ is the difference between the actual sentiment $s_{i,t}$ and the expected sentiment $\hat{s}_{i,t}$. The abnormal sentiments are added over the event window to capture the cumulative abnormal sentiment (CAS).

Table 3.4 shows the results. Column (1) reports the test for the revealed influence mechanism regarding the stock market reactions when some of the purged Twitter accounts were restored. There is evidence supporting this test. For a 1% higher share of followers restored, firms experienced a 0.572% increase in stock returns. Investors react positively when they learn that some fake followers were falsely purged, corroborating the revealed influence mechanism. The findings remain robust across several checks, including various time window variations, outliers, and the expected returns model (see Appendix Table A3.3). Overall, this estimate provides evidence for the revealed influence mechanism. Column (2) reports the test for the reputational damage mechanism. The coefficient on the term shows an insignificant estimate. Follow-up checks validate that the non-significant results are not an artifact of the event windows, outliers, or the collection of tweets as findings are replicated using the "@" syntax (see Appendix Table A3.4). Overall, there is no support for reputational damage.

Robustness Checks

Figure 3.5 plots the estimate for β_1 of equation 3.2 as obtained from various robustness checks together with 95% confidence intervals. Line (2) shows that results remain similar for an event window of [0,20] with increasing magnitude and confidence intervals (0.28%). Line (3) confirms the findings when controlling for outliers. Observations with the *Log(Followers)* not within the 10% and 90% percentiles are dropped from the dataset. The estimate is slightly smaller (0.063%), but confidence intervals remain small. Line (4) shows consistency for the Fama-French Plus Momentum model (FFPM) that controls for differences in stock returns associated with prior returns (Carhart 1997, Jegadeesh and Titman 1993). The estimate is similar to the baseline estimate (0.064%).

Table 3.4: Empirical Extension – Why did Investors React Negatively?
Note: The table explores why investors reacted negatively. Column (1) is a test of the revealed influence explanation. Column (2) is a test of the reputational damage explanation. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1) Cumulative Abnormal Returns when Followers Restored	(2) Cumulative Abnormal Sentiment around the Purge
Share restored	0.572^{+}	
	(0.331)	
Share fake		-0.227
		(0.955)
Constant	0.077**	-0.012
	(0.025)	(0.280)
Controls	X	X
Observations	693	270

Additional quasi-experimental evidence confirms the findings, considering that (1) not all firms lose followers and (2) not all firms are on Twitter. The estimates are documented in lines (5) and (6). The estimates indicate that firms that rely on fake followers experience a significant decline in abnormal returns by 0.4% compared to those without fake followers or those that do not operate a Twitter account, significant at a 10% level.

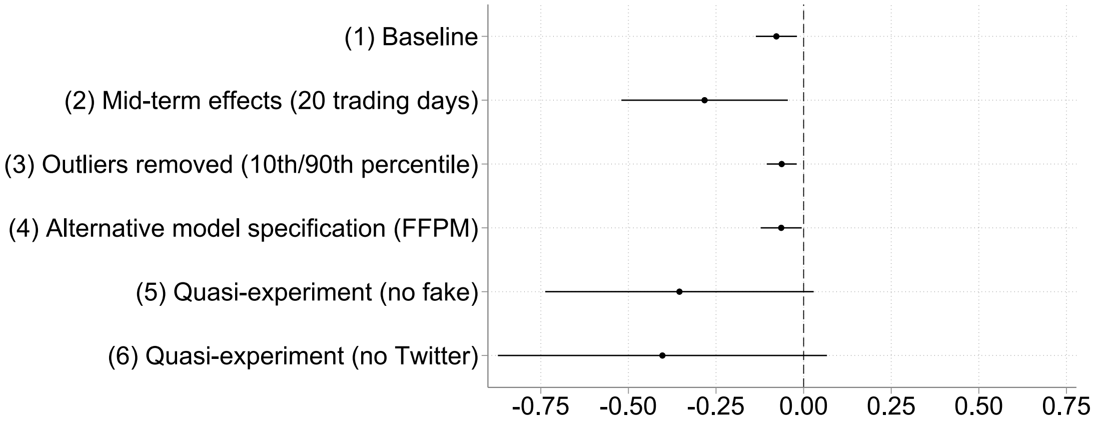
3.5. Discussion

Two main findings emerge from the investigations. First, relying on the purge of tens of millions of fake accounts, only 1.2% or 12,000 of firms' followers are identified as fake on average. A moderator analysis indicates that firms facing greater competitive pressure and smaller firms use fake followers to a larger degree, with the overall magnitude still modest. Second, investors react negatively when it is revealed that firms have engaged in fake follower use: A 1% higher share of fake followers is associated with cumulative abnormal returns of around -0.08%. This effect is not due to reputational damage inferred from social media sentiment but to the new information that firms' social media reach is lower than expected.

The findings make two main contributions to existing research. First, this study complements research on deceptive behavior by firms on social media (e.g., Hawlitschek et al. 2018, Huang and Liu 2023, Lee et al. 2018a, Nie et al. 2022, Wang et al. 2022). Prior research looked into distortion that can be attributed to misinformation, fake reviews, deceptive ads, sentiment manipulation, or social bots.

Figure 3.5: Estimates of Investor Reactions (β_1) for Various Robustness Checks

Note: The figure shows the estimates of investor reactions to fake follower use (i.e., β_1 of equation 3.2) as obtained from various robustness checks. Line (1) shows the baseline estimate. Line (2) relies on an extended event window of 20 trading days. Line (3) excludes outliers of *Log(Followers)*. Line (4) uses the Fama-French Plus Momentum model to calculate the cumulative abnormal returns. Line (5) uses a difference-in-differences estimation from a control group of accounts without fake followers. Line (5) uses a difference-in-differences estimation from a control group of accounts without a Twitter account. Lines (1) to (4) refer to the abnormal returns in percent for a 1% higher share of fake followers, Lines (5) and (6) refer to the abnormal returns in percent in comparison to the control group. The markers denote the estimate, and the whiskers give the 95% confidence intervals. The dashed vertical denotes a coefficient of 0, i.e., a null effect. The estimates are documented in Appendix Table A3.5.



While there have been attempts to investigate fake accounts and social bots (Cresci et al. 2015, Huang et al. 2018, Salge et al. 2022), there is no understanding of firms' fake follower use and the consequences for investors as identified in the literature review. To fill this gap, this study empirically assessed the extent of fake follower use and the risks from the practice as inferred from stock market reactions.

Second, this study contributes to research on social media engagement and outcomes for firm performance (e.g., Chen et al. 2015a, De Vries et al. 2017, Hennig-Thurau et al. 2015, Lis and Neßler 2014, Nofer and Hinz 2015, Teubner et al. 2016). Prior research has provided great insights into the effects that social media engagement can have on product promotion, brand building, and sales, showing the relevance of social media for firm outcomes. However, there were no insights about the role that fake followers play when inflating follower counts. This study provides first insights into how firms strategically use fake followers and how this fraudulent practice influences investors' decision-making.

This study provides practical implications for several stakeholders. First, the findings affect social media platforms. The study finds that, overall, fake followers present a minor problem. Firms are penalized for using fake followers, making it a non-profitable business to engage in this fraudulent practice when detection measures are in place. This makes it crucial for platforms to understand how to detect fake followers and to enhance the risk from the publicity of fake follower use. Second, the findings affect policymakers. Platform initiatives are driven by external pressure, and fake followers on Twitter are removed in response to an upcoming Federal Trade Commission (FTC) investigation. In an effort to protect, the Federal Trade Commission (FTC) has recently proposed new rules that would bar the selling of fake followers (Federal Trade Commission 2023). Despite the challenges of fake accounts, platform firms potentially have the incentive to tolerate fake accounts for growth purposes. This sheds light on the need for policy intervention. Third, the findings of the study affect individuals. Investigating the extent of fake follower counts, the findings sensitize laypeople to critically question social media metrics, particularly considering the large variation in fake follower use.

3.6. Conclusion

In light of the increasing public scrutiny over firms' social media practices, the first empirical study of this dissertation investigated whether firms use fake followers and to what degree follower counts are inflated by fake followers. In addition, it studied how stock markets react when investors learn that fake followers inflate a firm's follower count. The Great Purge in July 2018, initiated by Twitter, presents the empirical setting during which tens of millions of suspicious user accounts were removed from the platform. The findings show that firms lost around 1.2% of their followers due to the purge, indicating moderate use of fake accounts for attention-seeking and artificial popularity. However, there is heterogeneity concerning competitive intensity and firm size. Firms with higher competitive intensity and smaller firms are more likely to rely on fake followers; however, fake follower use is still modest. Distinguishing firms by follower loss, there is evidence for a decline in stock returns in response to the revelation of fake followers driven by the revealed influence of a firm. Still, the risk associated with fake followers is limited. These findings shed light on concerns about social media platforms concerning realness, credibility, and transparency but also show that the problem of fake followers is not as pronounced as assumed in the media.

3.7. Appendix for Chapter

Table A3.1: Correlation of Various Sentiment Measures
Note: The matrix shows the pairwise correlation between different sentiment measures. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	AFINN	SentiStrength	VADER
AFINN (Nielsen 2011)	1.000		
SentiStrength (Thelwall et al. 2010)	0.612***	1.000	
VADER (Hutto and Gilbert 2014)	0.783***	0.620***	1.000

Table A3.2: Robustness Checks: Fake Follower Use
Note: The table tests for the extent of fake follower use. Column (1) tests for an extended observation window of 150 days before and after the purge. Column (2) excludes outliers of *Log(Followers)*. Column (3) imputes missing values of the dependent variable. Column (4) applies a difference-in-differences estimation with a historical control group. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)
	Long-term	Outliers removed	Missing values imputed	Quasi-experiment
After purge	-0.012***	-0.012***	-0.012***	0.003***
	(0.001)	(0.001)	(0.001)	(0.000)
Treat × After purge				-0.022***
				(0.002)
Observations	234,104	40,810	51,179	54,728
Controls	X	X	X	X
Firm Fixed Effects	X	X	X	X
Time Fixed Effects	X	X	X	X

Table A3.3: Robustness Checks: Revealed Influence Mechanism
Note: The table tests for the revealed influence mechanism. Column (1) tests for an extended event window of [0,5]. Column (2) excludes outliers of *Log(Followers)* (p=0.220). Column (3) uses the Fama-French Plus Momentum model (p=0.189). OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)
	CAR [0,5]	Outliers removed	FFPM
Share restored	1.025*	0.353	0.358
	(0.491)	(0.287)	(0.273)
Constant	0.060	0.096***	0.042^{+}
	(0.038)	(0.028)	(0.025)
Observations	693	549	693
Controls	X	X	X

Table A3.4: Robustness Checks: Reputational Damage Mechanism
Note: The table tests for the reputational damage mechanism. Column (1) tests for an extended event window of [0,5]. Column (2) tests for an extended event window of [0,7]. Column (3) excludes outliers of *Log(Followers)*. Column (4) uses the sentiment of tweets that include mentions as an alternative dependent variable. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)
	CAS [0,5]	CAS [0,7]	Outliers removed	Mentions
Share fake	0.009	1.235	-1.269	-1.262
	(2.279)	(2.842)	(1.110)	(1.920)
Constant	-0.530	-0.900	0.280	0.347
	(0.679)	(0.970)	(0.381)	(0.427)
Observations	270	270	206	253
Controls	X	X	X	X

Table A3.5: Robustness Checks: Stock Market Reactions to Fake Follower Use

Note: The table tests for the investor reactions. Column (1) tests for an extended event window of [0,20]. Column (2) excludes outliers of *Log(Followers)*. Column (3) uses the Fama-French Plus Momentum model. Column (4) applies a difference-in-differences estimation with channels without fake followers as the control group. Column (5) applies a difference-in-differences estimation with channels without a Twitter account as the control group. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)	(5)
	Mid-term	Outliers	FFPM	Quasi-experiment	Quasi-experiment
		removed		no fake	no Twitter
Share fake	-0.283*	-0.063**	-0.064*		
	(0.121)	(0.022)	(0.030)		
Treat				-0.004^{+}	-0.004^{+}
				(0.002)	(0.002)
Constant	0.130^{+}	-0.005	-0.005	0.018**	0.013*
	(0.068)	(0.016)	(0.014)	(0.006)	(0.006)
Observations	691	547	691	673	627
Controls	X	X	X	X	X

4. Tackling Misinformation by Promoting Credible Content

4.1. Introduction

Following the aforementioned challenges of fakery, among others misinformation, one of the most pressing questions surrounding online platforms is which interventions are effective in combating it (e.g., Allcott and Gentzkow 2017, Bhargava 2022, Hwang and Lee 2024, Pennycook et al. 2020). Online platforms like YouTube, TikTok, or Instagram enable anyone to create and distribute informational content, making them susceptible to the spread of inaccurate information. As evident in recent years, platform misinformation can have disastrous consequences, especially in domains such as healthcare, elections, or climate policy (Lazer et al. 2018, Vosoughi et al. 2018).

Research into interventions against misinformation has so far focused on measures that reduce the spread of fakery but much less on the complementary approach of amplifying content from credible sources, as identified in the literature review. In particular, considerable research has been devoted to approaches such as fact-checking, warnings, and user literacy training (e.g., Berger et al. 2023, Borwankar et al. 2022, Moravec et al. 2020, Roozenbeek et al. 2022). While these approaches are important, they focus on pruning inaccurate information or helping users distinguish between correct and incorrect. However, several platforms have recently started to implement an alternative strategy, namely, to certify the credibility of creators and amplify their content. So far, the effectiveness of this strategy remains unassessed, creating a gap in the understanding of how best to combat misinformation.

This study addresses this gap by examining YouTube's *Authoritative Health Information* program. The program seeks to improve the spread of accurate health information by promoting content from what the platform refers to as *authoritative* channels – namely, channels operated by hospitals, medical schools, and certified doctors. Those channels are promoted in the search results for health keywords, and their videos receive a label that explains that the channel's content is credible. This study uses large-scale empirical data to understand how YouTube's intervention affected the views of authoritative channels' content over that of other channels, the mechanism through which the program affected channel viewership, and the downstream consequences for content production on YouTube.

The intervention presents a quasi-experiment whose effects are estimated using a difference-in-difference-in-differences design. The study exploits the fact that YouTube introduced the program in different countries at different times. Germany was one of the first countries to experience the program in February 2023; French YouTube channels did not see the rollout until much later that year. Thus, the study uses German YouTube channels as the treatment group and French YouTube channels as a control group. This design is appropriate because YouTube users' statistics are comparable among Germany and France, and the fit between the countries has been shown in prior studies (Aguiar et al. 2018, Calzada and Gil 2020, DataReportal et al. 2022). Furthermore, despite being geographically located next to each other, significant language barriers prohibit channels from producing content in the other country's language. The study relies on a 21-week daily panel of YouTube channels with their views, subscribers, and videos collected via the YouTube API.

The analysis shows three main findings. First, YouTube's intervention yielded positive, albeit modest, results. Viewership of channels deemed authoritative by YouTube increased by approximately 4.3%. This uptick suggests that YouTube's strategy successfully nudged a portion of its audience toward credible content. Conversely, non-authoritative channels saw a decline in viewership by around 4.6%. Moreover, the findings indicate that these effects are not fleeting. Over time, the increased viewership of authoritative channels and the corresponding decline in non-authoritative viewership have persisted. This is an important observation given that previous research highlighted the transient nature of some countermeasures against misinformation (Barrera et al. 2020, Berger et al. 2023). However, statistical significance and persistence aside, the effects of YouTube's Authoritative Health Information program are modest when compared to prior studies that investigate other kinds of content that were promoted (Bockstedt and Goh 2011, Dewan et al. 2023, Huang et al. 2022a).

Second, the study unveils the mechanics of YouTube's intervention, which involved two changes: the feature and the label. Follow-up analyses sought to determine which of these changes was responsible for the observed shift in viewership. This can be investigated by taking advantage of the fact that not all channels were effectively featured in the search results, as the feature was only displayed for some search terms but not others. It becomes apparent that the label alone had no discernible impact on viewership. The entire positive effect observed can be attributed to the feature. In essence, the outcome of the intervention hinges entirely on the adjustments to the visibility of the authoritative content, not on the labels assigned to it. Therefore, labeling content as authoritative is perhaps insufficient, despite the positive effects usually observed for other kinds of quality certifications and badges on platforms (e.g., Dewan et al. 2023, Hui et al. 2007, Oezpolat et al. 2013), but confirms the hypothesis that labels might not be effective in reducing engagement with misinformation (Borwankar et al. 2022, Kim et al. 2019, Moravec et al. 2023). In addition, however, the findings show that one opportunity to combat misinformation lies in making authoritative content discovered more easily and entering the consideration set of platform users.

Finally, the findings reveal that YouTube's intervention had little effect on the content production of both authoritative and non-authoritative channels. Specifically, authoritative channels increased their output by only 0.8%, while non-authoritative channels did not significantly adapt production at all. These negligible changes indicate that the intervention did not significantly motivate channels to alter their production strategies. One possible explanation for these minimal effects is that the shifts in viewership were too small to influence content creation decisions. However, the current data does not fully explain these results. Future research could provide more insight by surveying channels about their decision-making processes or conducting in-depth interviews.

The chapter is structured as follows. Section 4.2 reviews related work. Section 4.3 describes the YouTube Authoritative Health Information program as well as the theoretical expectations. Section 4.4 explains the difference-in-difference-in-differences design and the data. Section 4.5 presents the results. Section 4.6 discusses the implications for theory and practice.

4.2. Related Work

False information refers to factually incorrect information, and it is commonly categorized into misinformation and disinformation based on the intention of the sender (George et al. 2021, Hernon 1995, Tandoc Jr et al. 2018, Wardle and Derakhshan 2017). Misinformation results from "an honest mistake" (Hernon 1995, p. 134) and includes factually wrong information that is not disseminated deliberately, such as false connections, misleading quotes and images, or satire (Wardle and Derakhshan 2017). Disinformation emerges from "a deliberate attempt to deceive or mislead" (Hernon 1995, p. 134), i.e., it involves fabricated content with the goal to manipulate others by exploiting human biases (Allcott and Gentzkow 2017, French et al. 2023, Miller et al. 2024, Wardle and Derakhshan 2017).

A growing number of papers in information systems and adjacent fields has been studying countermeasures against the spread of false information on online platforms (e.g., see reviews from Chen et al. 2023, George et al. 2021, Li et al. 2022). So far, considerable work has been devoted to understanding interventions that intend to solve the issue by reducing the spread or impact of inaccurate information, such as fact-checking (Berger et al. 2023, Moravec et al. 2020, Nyhan et al. 2020, Porter and Wood 2021, Schuetz et al. 2021), crowd-based content inspection (Borwankar et al. 2022, Pennycook and Rand 2019), inoculation and literacy training (Badrinathan 2021, Lewandowsky and Van der Linden 2021, Pennycook et al. 2020, Roozenbeek et al. 2022), identity verification (Wang et al. 2021a), as well as draining of advertisement money through greater transparency (Ahmad et al. 2024).[1] However, these contributions notwithstanding, they have not studied the complementary approach of promoting the spread of accurate information.

Some research has examined the efficacy of labels alerting users to potentially inaccurate information related to a particular content or source (Borwankar et al. 2022, Bradshaw et al. 2021, Kim and Dennis 2019, Kim et al. 2019, Moravec et al. 2019, 2023, Pennycook and Rand 2019). Many of these studies are lab experiments in which users are presented with labels of different purposes and designs. For example, Kim et al. (2019) manipulated social media posts linking to news outlets such that they show a star rating which informs about the reliability of that outlet, and Moravec et al. (2023) manipulated Facebook posts to show a label that tells users if the post originates from government-controlled pages. One exception is Borwankar et al. (2022), who studied the effects of Twitter's crowdsourced Birdwatch program using that platform's data. The findings cannot confirm consistent changes in users' engagement with labeled content, but they suggest that users must notice the label, understand its meaning, and trust its reliability.

Despite a lack of research investigating the consequences of amplifying credible content, one noteworthy study is Hwang and Lee (2024). They study Twitter's 2019 intervention to display a link to a government or non-profit website considered reliable when users searched for health topics, such as *vaccines.gov* or *WHO.int*. They find that misinformation spreads less after the intervention's rollout. The intervention presented in the second empirical study of this dissertation is different (i.e., it seeks to certify reliable information sources on the platform and propel them in the search results) and investigates whether credible information is viewed more often and how it impacts the source of credible as opposed to the spread of inaccurate content.

The study also connects to the more general research on labels (or certificates, seals, etc.) and promotions on multi-sided platforms (e.g., Dewan et al. 2023, Hui et al. 2007, Oezpolat et al. 2013, Rietveld et al. 2019). The typical motivation for these interventions is to resolve information asymmetries between the demand and supply side on platforms (Boudreau and Hagiu 2009, Parker and Van Alstyne 2005). For example, AirBnB highlights some hosts with a "Superhost" badge, and Google promotes some apps as "Editors' Choice". Many of these studies find positive effects of these interventions for

[1] Some findings are mixed, indicating that the effectiveness of interventions also depends on the setting in which they are rolled out. For example, Schuetz et al. (2021) observe that fact-checking positively influenced user behavior around the COVID-19 pandemic, whereas Berger et al. (2023) find that fact-checking is only effective in the short term for disputed health and nutrition topics.

demand (e.g., Dewan et al. 2023, Oezpolat et al. 2013, Terlaak and King 2006). However, the key question is whether such an intervention works in the case of (mis)information. One counterargument could be that users find labels trustworthy when it comes to evaluating products or sellers but ignore or discredit labels when it comes to domains where they have strong directional motivations (see Nyhan 2020). For example, Bradshaw et al. (2021) find that YouTube's state media label had no meaningful effect on user engagement with channel content, perhaps because users have pre-existing ideologies and political leanings that may prompt them to ignore the labels.

4.3. The YouTube Authoritative Health Information Program

4.3.1. Description

YouTube serves as a bustling hub for health information channels, offering a diverse array of content for education and inspiration (e.g., Liu et al. 2020, Mitkina et al. 2023). Such content includes, for example, healthy lifestyles (e.g., Andrew Huberman), prevention or treatment of medical conditions (e.g., Doctor Mike), mental self-care (e.g., The School of Life), dietary trends (e.g., Bryan Johnson), and fitness routines (e.g., Yoga with Adriene). Especially in Europe, more than half of the EU citizens from the age 16 to 74 regularly search for health information online; in Germany, these numbers are even higher than this, at approximately 70 % (Eurostat 2021). Amongst these, YouTube is the most accessed source (Kodura 2023). At the same time, the platform has been plagued by the spread of inaccurate information, for example, during health crises such as the COVID-19 pandemic (Duffy 2022, Li et al. 2020).

On October 18, 2022, YouTube announced the Authoritative Health Information program (Graham 2022c, see full text in Appendix B4.1). The program distinguishes so-called authoritative health information channels and seeks to promote their content on the platform. In particular, authoritative channels receive two treatments, as shown in Figure 4.1. First, authoritative channels become featured at the top of the search results for health-related keywords (*search feature*). When users now search for health information, the results display videos from authoritative channels at the very top, titled "From health sources." Second, videos from authoritative channels receive a label that states that the content is from an authoritative source and what renders that channel authoritative (*label*).

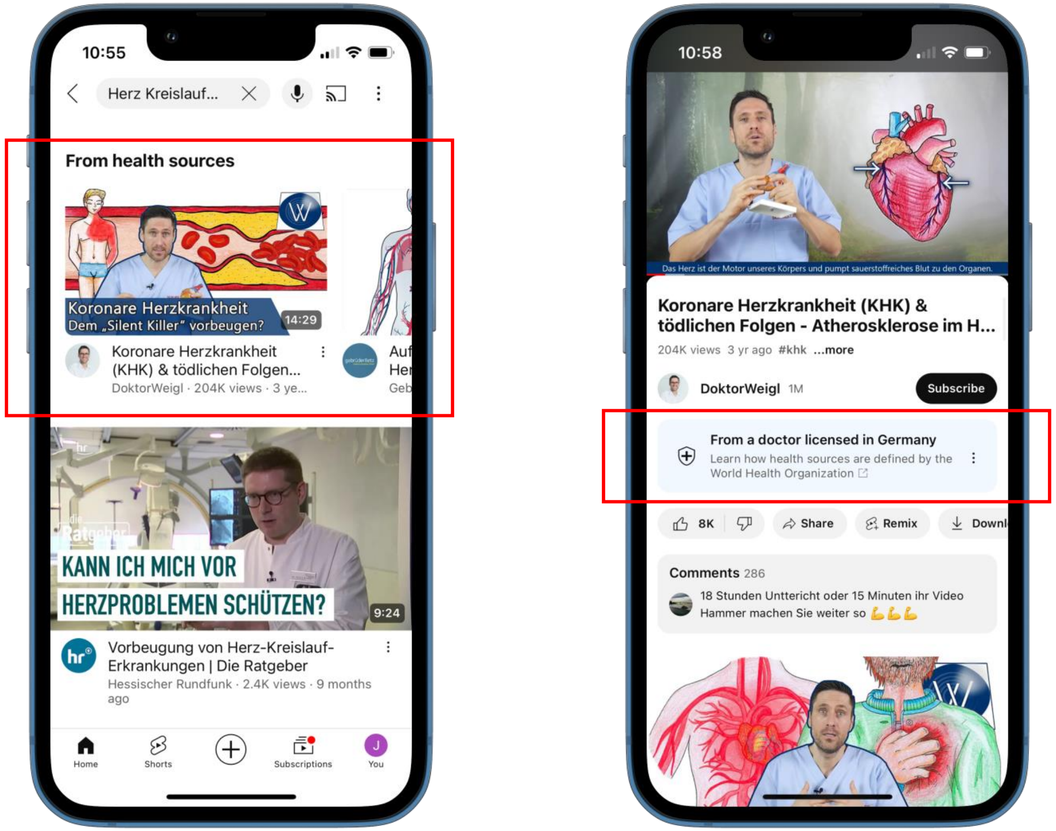
To define authoritative channels, YouTube followed the guidance of the National Academy of Medicine for social media platforms, namely that any medical information a

Figure 4.1: YouTube’s Authoritative Health Information Program

Channels that YouTube declared as authoritative experienced two changes ..

(i) Feature

(ii) Label



10:55

Herz Kreislauf...

From health sources

Koronare Herzkrankheit (KHK) & tödlichen Folgen...
DoktorWeigl · 204K views · 3 ye...

Vorbeugung von Herz-Kreislauf-Erkrankungen | Die Ratgeber
Hessischer Rundfunk · 2.4K views · 9 months ago

10:58

Das Herz ist der Motor unseres Körpers und pumpt sauerstoffreiches Blut zu den Organen.

Koronare Herzkrankheit (KHK) & tödlichen Folgen - Atherosklerose im H...
204K views 3 yr ago #khk ...more

DoktorWeigl 1M

From a doctor licensed in Germany
Learn how health sources are defined by the World Health Organization

Comments 286
18 Stunden Untterricht oder 15 Minuten ihr Video Hammer machen Sie weiter so 💪💪💪

..their videos are featured at the top of the search results

..their videos receive a label that certifies authority status channel gives comes from “medically trained and qualified professionals” (Kington et al. 2021, p. 28). YouTube primarily considered organizations with pre-existing, standardized vetting mechanisms (i.e., certification, government accountability), such as healthcare organizations, university clinics, educational institutions, public health departments, health insurers, and government organizations. YouTube automatically defined channels that fell into this circle as authoritative. All other channels that wanted to be described as authoritative had to formally apply and undergo an external review by the third-party auditor LegitScript (YouTube Help 2024a, 2023). Eligibility criteria were restrictive and required the channels to fulfill the following criteria: proof of being a certified doctor or psychologist; no violation of the YouTube policies; playback time of more than 1,500 hours in the last 12 months, or shorts with more than 1.5 million views in the previous 90 days; content that is science-based, objective, and transparent as defined by the National Academy of Medicine (Burstin et al. 2023). While there is no information about how many channels were denied from the program, it becomes evident that the review was strict in the sense that some channels were mandated to provide scientific references for their claims and to remove or refine some of their videos for borderline statements. For example, the physiotherapy channel Liebscher & Pracht reported to have reviewed more than 1,400 of their videos to remove or refine statements that could not entirely be supported by scientific evidence; an effort that involved 17 employees over three months (Kreienbrink 2023).

4.3.2. Theoretical Expectations

Effect on the Views of Authoritative Content

The feature and the label, as introduced by the YouTube Authoritative Health Information program, should cause authoritative content to be watched more. The driving forces behind this effect are two distinct theoretical mechanisms through which the feature and the label operate.

Regarding the feature, *sequential search theory* suggests that the intervention will attract more viewers to authoritative channels because these channels are now positioned at the top of the search results where they receive relatively more attention (McCall 1970, Weitzman 1979). In particular, the theoretical mechanism is that when individuals (i.e., viewers) are confronted with a ranked list (i.e., the search results), they evaluate the different list items (i.e., videos) one by one and by proceeding stepwise from top to bottom. Whatever is ranked higher in the list will be considered earlier, and moving on will be considered in light of the search cost and uncertainty of obtaining a better option. Especially the positive effects of the first rank have been documented in a wide variety of digital settings, including product rankings or search engines (e.g., Ghose et al. 2013, Oestreicher-Singer and Sundararajan 2012, Ursu 2018). Departing from these arguments, authoritative channels will receive more views since they are positioned at the top.

Regarding the label, *signaling theory* suggests that authoritative channels will attract greater demand because they will be deemed more credible (Ross 1973, Spence 1973). Signaling theory addresses the issue of information asymmetry in market transactions, where one party often lacks sufficient information about the other party's ability to meet their needs (e.g., Jensen and Meckling 1976, Stiglitz 2000). This uncertainty can hinder or even prevent transactions from occurring. Signaling involves one party credibly conveying information to reduce this uncertainty and assure the other party of their capabilities or intentions (for a review, see Connelly et al. 2011). The labels displayed along authoritative channels' videos act as a signal. Videos are experience goods whose information quality is difficult to assess before usage. As a result of this uncertainty, users' decisions will likely be influenced by the judgment of others, as manifested in the label (e.g., Dimoka et al. 2012, Ter Huurne et al. 2017). Especially on YouTube, a setting with an abundance of prices that could give a cue about quality, the label is expected to attract viewership to authoritative channels.

Downstream Effects on Channel Popularity and Content Production

As the intervention promotes content quantity, authoritative channels are expected to prioritize a higher content volume using more resources (Poch and Martin 2015, Törhönen et al. 2020). Evidence from prior research shows that external incentives enhance content contributions to exploit them. For example, audience size and perceived popularity strongly drive contributions as they affect recognition and enhance motivation (Cao et al. 2023, Chen et al. 2018, Goes et al. 2014, Qiu and Kumar 2017). Further, platforms effectively enhance contribution via monetary or social-norm based incentives, badges, awards, or improved functionalities (Burtch et al. 2018, Cavusoglu et al. 2021, Chen et al. 2019b, Gallus 2017). In line with the latter, the literature shows that content contributors align their contents with the most promising areas, i.e., by producing content in line with the awarded ones (Burtch et al. 2022, Foerderer et al. 2021) or by prioritizing quality or quantity depending on which one is incentivized (Claussen et al. 2013, Rubin et al. 2018).

In contrast, theory suggests that the perceived quality of the content declines. In line with the resource-based view of firms, the reallocation of resources toward quantity from their strategic perspective to exploit the intervention makes fewer resources available for quality (Barney 1991). Such a quantity-quality trade-off is not a new phenomenon but has also been shown in other settings, e.g., the interaction between quantity and quality of children under a constant household income (Becker and Lewis 1973). Similarly, the more current literature on content contributions confirms this suggestion. For example, Khern-am nuai et al. (2018) show that while the quantity of reviews increases with financial incentives, the effort spent on single reviews decreases, resulting in shorter reviews and reduced helpfulness.[2]

At the same time, in line with the unintended effects of awards, it is also possible that channels relax their efforts overall as they can reach the same number of views despite lower effort. For example, Malmendier and Tate (2009) show that company CEOs that get a superstar status, i.e., that benefit from prestigious awards, underperform and use their time for more enjoyable activities. Further, He et al. (2023) show that threshold-based incentives encourage users to contribute only until reaching the threshold.

4.4. Method and Data

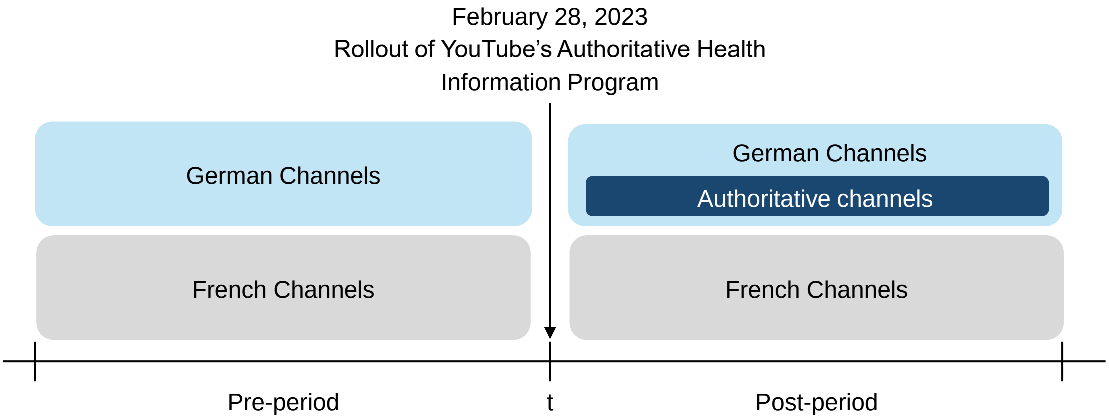
4.4.1. Research Design

Figure 4.2 illustrates the difference-in-difference-in-differences (DDD) setup that is used to understand the impact of YouTube's program. In general, such a design is well-suited for isolating causal effects because it compares outcome changes over time between a treatment group and a control group, allowing researchers to control for time-invariant confounders (Angrist and Pischke 2008). By examining how the treatment group's outcomes diverge from the control group's outcomes after an intervention, DDD effectively captures the causal impact of the treatment while mitigating the influence of other factors that remain constant over time. The setup follows a standard two-way fixed effects difference-in-difference-in-differences design, with the rollout of the program being the intervention and the analysis being conducted at the level of the channel. The data, which will be described in the next section, is a channel-day panel.

For assignment into the treated and control group, the study exploits the fact that YouTube introduced the program only in Germany but not in other countries. Therefore, German health channels are considered treated channels. French channels serve as a comparable control group. They were not part of YouTube's Authoritative Health Information program and, as such, did not experience any change at that time. French channels are well-suited compared to other countries. They share similar health trends, especially an aging population, alcohol, and drug consumption above average, lack of

[2] In contrast to a decline in quality, the study from Kovács and Sharkey (2014) indicates that a decline in popularity can be explained by (1) a diversification in the audience that potentially dislikes content and (2) the fact that people tend to dislike popular content. However, as the analyses do not only consider average effects that are prone to bias from newly acquired users, this cannot be seen as the only potential explanation.

Figure 4.2: Research Design: Difference-in-Difference-in-Differences
Note: The figure illustrates the research design. It relies on a daily panel of YouTube health channels before and after the rollout of the intervention. The study compares German (treated) to French (control) channels before and after the rollout and between channels defined as authoritative and non-authoritative after the intervention.



physical activity, high health expenditure, and good healthcare quality and coverage (OECD/European Union 2020). Also, Germany and France are geographically and culturally close. According to DataReportal et al. (2022), Germany is the European country with the most YouTube users as of 2024, accounting for 65.7 million users, followed by the UK and France.

This group design is well-suited in light of the Stable Unit Treatment Value Assumption (SUTVA). The SUTVA would be violated if the channels in Germany and France are considerably interconnected, i.e., if treated channels cater to the control channels' audience or vice versa. In general, it is not impossible for French viewers to watch German channels or vice versa. However, this overlap is likely to be minimal. French YouTube channels predominantly target French-speaking audiences, whereas German YouTube channels primarily cater to German-speaking viewers, creating a language barrier. This language barrier is a significant obstacle since French is a Romance language (i.e., similar to Spanish, Italian, and Portuguese), whereas German is a Germanic language (i.e., closely related to English, Dutch, and the Scandinavian languages). While French and German may have a tiny overlap in vocabulary due to historical interactions and borrowings, their grammar, pronunciation, and overall structure are fundamentally different. Also, note that English channels (i.e., UK or U.S.) are not suited as a control country because the English language is widely used in many countries, also where it is not the primary language. In other studies, France has been used as a control group for Germany, showing its fit as a control (Aguiar et al. 2018, Calzada and Gil 2020). It is unlikely that confounding events would have affected French channels at the time of the study after inspecting YouTube's platform announcements. However, the robustness section also provides empirical evidence that the results are consistent when using a control group other than French channels.

The intervention is also well-suited because channels could not anticipate it, and it was rolled out swiftly. YouTube did not disclose the details of the program or provide a timeline after the announcement. It is also likely that the change was not a completely deliberate decision of YouTube. In 2022, an open letter, signed by more than 80 fact-checking organizations around the globe, asked YouTube's CEO Susan Wojcicki for more commitment to reducing misinformation (Duffy 2022). Moreover, motivated by their experience during the COVID-19 pandemic, legislators around the globe were drafting or introducing bills to oblige platforms to take stricter measures against misinformation (Colomina et al. 2021). In 2021, for instance, Facebook's Mark Zuckerberg, Twitter's Jack Dorsey, and Google's Sundar Pichai had to testify before Congress on misinformation on their platforms (McCabe 2021). This shows that the event is not particularly driven by the platform's intrinsic motivation but rather by external pressure. The rollout then took place overnight and was implemented for all devices (mobile, desktop, TV).[3]

We use the program's rollout date to define the pre- and post-periods. YouTube announced the program on October 18, 2022, opened applications on October 27, 2022, and rolled out the changes (i.e., the search feature and the label) on February 28, 2023 (Graham 2022c, Weiß 2023). For the estimation, the study relies on a period of 5 weeks before and 16 weeks after the intervention. The data collection was prepared after the program's announcement, determining the beginning of the observation period. The end of the observation period is motivated by having enough variation available to observe changes but to avoid confounding events from other changes to the platform.[4]

[3] The only exception is that no label is displayed on TV (Appendix Figure A4.1). However, as only a tiny fraction of users in Germany access YouTube from the TV, this does not meaningfully affect the results (ARD and ZDF 2019).

[4] Before the observation period, YouTube had rolled out the program in the U.S., Brazil, India, Japan, or the UK (Graham 2022a,b). YouTube gradually expanded the Authoritative Health Information program to other countries, also to France in September 2023 – which is not concerning for the study because it is long after the end of the observation period (Phelippeaux 2023, Caruso 2023).

4.4.2. Data Collection

A channel-day panel is built from data retrieved from the official YouTube API via its research access.[5] Since no ready-to-use index of health channels is available, such an index was compiled manually. The index was constructed by searching for health-related keywords using the YouTube API v3 and then recording the obtained channels. The keywords used are the health topics published by the World Health Organization (2022). These account for 189 significant healthcare topics (e.g., Ageing, Alcohol, Cancer, Diabetes, Healthy Diet), thereby presenting a comprehensive search radius. These search terms were translated to German and French to query the YouTube API. To ensure ongoing relevance and avoid capturing only a snapshot in time, the search was conducted on a daily basis, starting from December 18, 2022. The process was concluded on December 27, once the discovery of new channels fell below 5% consistently for five consecutive days. To provide a clean index of channels, false positives in terms of channels for which the language or location attribute indicated that they are neither German nor French, as well as non-health channels (i.e., as determined by inspecting the channel description), were removed from the set of channels.

The channel-day panel relies on the daily data collection on the indexed YouTube channels. The relevant variables contain various channel characteristics i.e., the number of videos published, views, and subscribers. To understand whether YouTube defined a channel as authoritative, and since the information was not recorded in the API, a manual check of the YouTube channels was done to check whether a channel's videos had received the label and whether the search feature was available for respective search terms. The resulting data includes 2,159 German (treated) and 1,560 French (control) channels.

4.4.3. Variables

Dependent Variables: The primary dependent variable is the total number of views (*Views*) of channel i's videos on day t, following prior research (e.g., Garg et al. 2023). YouTube considers a video as "viewed" if a viewer watches it for at least 30 seconds. Additionally, this count includes multiple views from the same viewer (McLachlan and Cooper 2022). YouTube conducts several accuracy checks, e.g., by accounting for spam views (YouTube Help 2024b). *Subscribers* is the number of subscribers of channel i on day t. For the regressions, they are normalized on the number of views on that day

[5] Using the YouTube API has the advantage that there is no need to rely on data from third-party providers or web scraping, which can be prone to errors, but instead, the data is available directly from the origin. A few recent studies have also begun to use the YouTube API (El-Komboz et al. 2023, Kerkhof 2024).

to account for the relative popularity. *Videos* is the total number of videos uploaded by channel i on day t. All variables are measured on a daily basis and log-transformed. For views and videos, 1 is added to the variable for log-transformation to account for values of zero. As negative values result from the deletion of videos, these observations are missing once the log-transformation is done. This ensures that the results are not biased by the reduction of views in response to the deletion of videos but that the variables reliably measure the number of views and videos in response to the intervention.

Independent variables: To map the DDD setup, three variables are created. *After* is 1 on days after the introduction of the YouTube Authoritative Health Information program, and 0 otherwise. *Treat* is 1 for treated (i.e., German) and 0 for control (i.e., French) channels. *Authoritative* is 1 for all channels that YouTube defines as authoritative channels after the introduction of the program, and 0 otherwise. Several further variables are created. *Pre-vetted* is 1 for channels with existing standardized vetting mechanisms, and 0 otherwise. Following the definition of YouTube (YouTube Help 2023), channels are coded as pre-vetted if they are health organizations with certification mechanisms in place or government accountability, including hospitals, pharmacies, governmental authorities, health insurance, and medical schools of universities. Not included are individuals, such as private doctors or medical staff. *Age* is the number of days since a channel joined YouTube, measured until the day of the introduction of the YouTube Authoritative Health Information program.

Table 4.1 defines the variables and reports their descriptive statistics.

4.4.4. Matching

To balance the treated and control group in size and to reduce heterogeneity, coarsened exact matching (CEM) is employed on the data. The channels are matched on the dependent variable *Views*. Further, they are matched on *Videos* to control for the production output of a channel. They are also matched on *Age* to reduce imbalance along the experience of a channel. The matching procedure uses CEM's automatic binning algorithm, "sturges rule" in Stata, and requires k2k to ensure identical group size (Blackwell et al. 2009, Iacus et al. 2012). Table 4.2 shows that the differences among the groups are close to zero and insignificant.

After matching, the resulting panel is on the channel-day level. It contains 445,860 channel-day observations, based on 3,050 channels over 35 days before and 112 days after the intervention. Of the 1,525 German channels, 1,364 are non-authoritative, while 161 are authoritative.

Table 4.1: Variables and Descriptive Statistics
Note: The table describes the data set along the variables over the observation period after matching. It relies on 445,860 channel-day observations based on 3,050 channels over 35 days before and 112 days after the introduction of the YouTube Authoritative Health Information program.

Variable	**Description**	**Mean**	**SD**	**Min**	**Median**	**Max**
Views [in Thousands]	Number of views of channel i received on day t	1.201	39.728	-13,089.978	0.060	12,293.341
Subscribers [in Thousands]	Number of subscribers of channel i on day t	9.446	37.538	0.000	0.716	1,100.000
Videos	Number of videos of channel i uploaded on day t	0.076	1.364	-323.000	0.000	161.000
After	1 if day t is after the intervention, 0 otherwise	0.764	0.425	0.000	1.000	1.000
Treat	1 if channel i is German, 0 otherwise	0.502	0.500	0.000	1.000	1.000
Authoritative	1 if channel i is certified as an authoritative source, 0 otherwise	0.053	0.224	0.000	0.000	1.000
Search feature	1 if search feature is available for channel i for at least one video, 0 otherwise	0.452	0.498	0.000	0.000	1.000
Pre-vetted	1 if channel i is pre-vetted by standardized vetting mechanisms, 0 otherwise	0.048	0.213	0.000	0.000	1.000
Age	Number of days since channel i joined YouTube until the introduction of the YouTube Authoritative Health Information program	2,185.835	1,313.326	64.000	2,017.000	6,191.000

Table 4.2: Test for Group Differences
Note: The table tests for differences in means between treated and control channels before and after the matching. Column “Difference in Means” reports the t-test. Column “Difference in Trends” reports a regression estimate for the difference in time trend before the treatment. $^{+}\ p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

	Before Matching				**After Matching**			
	Control	Treatment	Difference in Means	Difference in Trends	Control	Treatment	Difference in Means	Difference in Trends
Subscribers	17,048.077	12,517.017	4,531.061^{+}	-0.582	9,983.975	8,198.865	1,785.111	-0.361
Views	2,010,968.110	2,449,000.932	-438,032.822	360.032	1,081,900.162	1,286,300.901	-204,400.738	283.024
Videos	131.883	135.064	-3.181	0.004	106.351	114.746	-8.395	0.009
Age	2,192.058	2,311.629	-119.571**	-	2,186.081	2,184.061	2.020	-
Obs.	1,560	2,159			1,525	1,525		

4.4.5. Estimation Model

The study relies on a fixed-effects difference-in-difference-in-differences estimation. The regression model takes the form:

$$\begin{aligned} Y_{i,t} &= \beta_0 + \beta_1 \times After_t + \beta_2 \times Treat_i \times After_t + \\ &\beta_3 \times Authoritative_i \times Treat_i \times After_t + \theta_i + \tau_t + \epsilon_{i,t} \end{aligned} \tag{4.1}$$

where β_2 captures the effect of the intervention on $Y_{i,t}$ of treated channels, and β_3 captures the additional effect on $Y_{i,t}$ for authoritative channels.

Several controls are included to account for channel and time heterogeneity. First, to account for the fact that individual channel characteristics could drive the effects, the estimation includes channel fixed effects θ_i. Channel fixed effects allow to consider channel heterogeneity that remains constant over time, i.e., the background and education of a channel owner, the equipment for video production, topic expertise, or existing quality signals available for channels such as their name (Wooldridge 2019). Second, to account for heterogeneity among different points in time, for instance, during weekends or flu season, the estimation includes time fixed effects τ_t (Wooldridge 2019).

4.5. Results

4.5.1. Effect of the Intervention on Channel Views

Figure 4.3 plots the coefficients obtained from estimating equation 4.1 using *Views* as the dependent variable, together with 95% confidence intervals. The dashed line denotes a null effect. The Authoritative Health Information program had a statistically significant impact on viewership patterns. It is observed that the intervention reduced views for non-authoritative channels by 4.6%, while it increased the views of authoritative channels by 4.3%. Both effects are statistically significant. In other words, the intervention attracted a statistically significantly larger number of views to authoritative channels while reducing the number of views for non-authoritative channels. In terms of economic significance – and given a median number of daily views in the control group of 55 views per day – an average decline in daily viewership can be inferred. For non-authoritative channels, views decline by 2.53 views per day, on average. For the authoritative channels, the views increase by 2.38 views per day, on average. Because the pre-intervention median daily views for authoritative channels are

101 compared to 63 for non-authoritative channels, the findings indicate that the intervention has a small effect on increasing the gap between both groups.

Figure 4.3: Effects of YouTube's Authoritative Health Information Program on Channel Views

Note: The figure shows the estimates for the coefficients of β_2 (light blue) and β_3 (dark blue) of equation 4.1 as obtained from various robustness checks. Line (1) shows the baseline estimates. Line (2) excludes outliers of the dependent variable. Line (3) winsorizes the dependent variable. The markers denote the estimate, and the whiskers give the 95% confidence intervals. The dashed vertical denotes an effect size of 0, i.e., a null effect. The estimates are documented in Appendix Table A4.1.

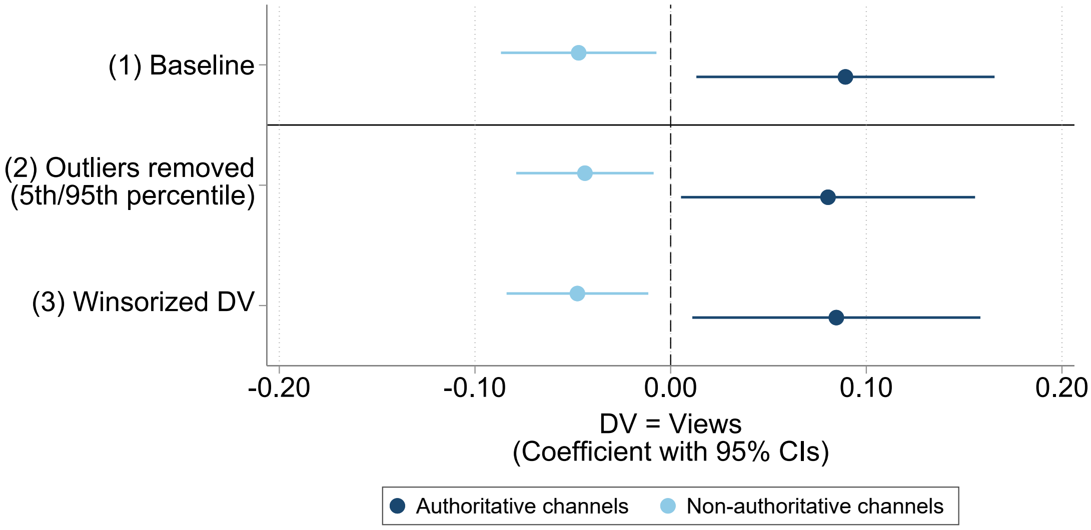
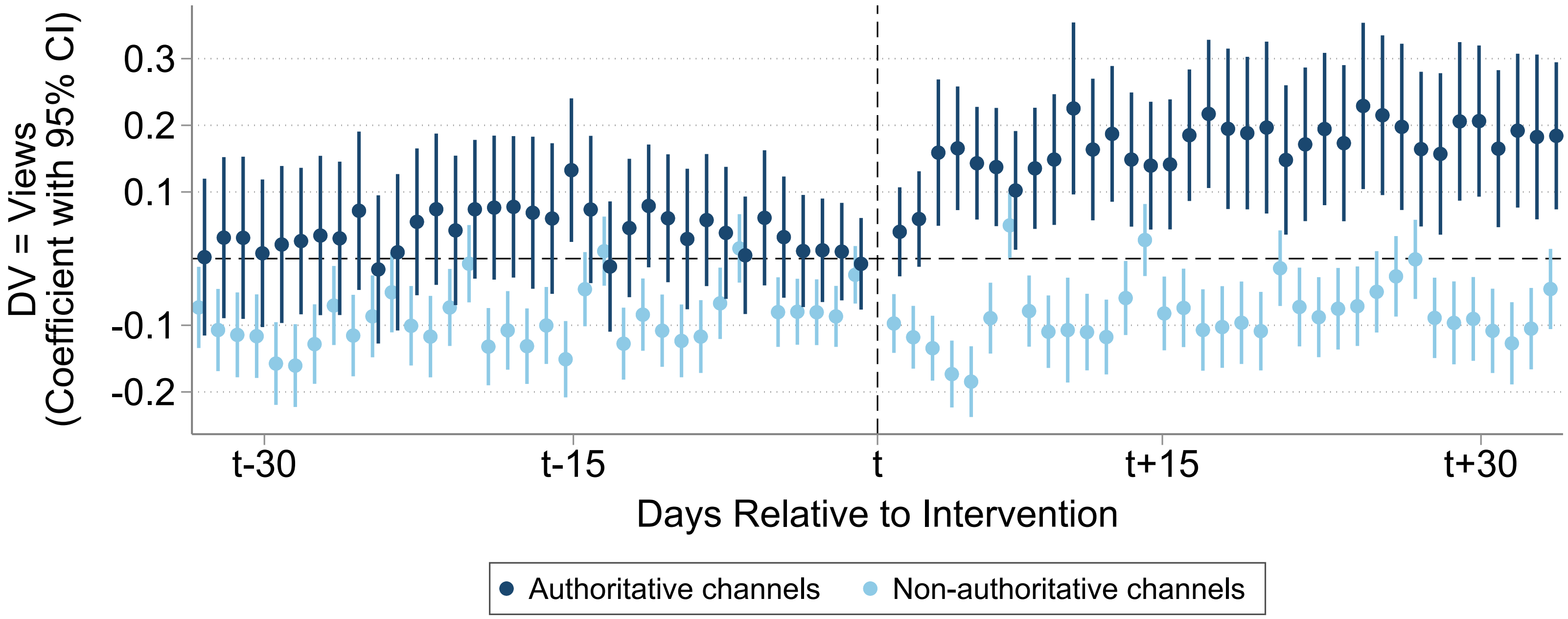


Figure 4.4 shows the coefficients obtained from an alternative leads and lags formulation of the DDD framework over a 35-day time window before and after the intervention, in which β_2 and β_3 are allowed to differ by time. As before, the markers denote the point estimate, and the whiskers give the 95% confidence intervals. It can be observed that the coefficients are close to zero and insignificant before the intervention, which further corroborates the assumption over parallel trends. After the rollout of the intervention, channel views of non-authoritative channels decline gradually over time, especially around 12 weeks after the rollout. The views of authoritative channels increase immediately after the rollout and are significantly more positive even weeks afterward. The temporal pattern here aligns with research investigating how users access YouTube videos (Goodrow 2021, Zhou et al. 2016). Zhou et al. (2016) show that YouTube search and recommendations of videos drive views. As the recommendations are determined based on others' preferences, the effect of the content manipulation on the search level only manifests over time as right after the introduction the recommendations are potentially still driven by non-authoritative content. However, the non-authoritative channels lose visibility over time, and their videos get fewer views.

Figure 4.4: Effects of YouTube's Authoritative Health Information Program on Channel Views (Daily)

Note: The figure shows the daily estimates for the coefficients of β_2 (light blue) and β_3 (dark blue) of equation 4.1 for 35 days before and after the rollout of the intervention. The day of the rollout (t) is the baseline, denoted by a dashed vertical. The markers denote the estimate, and the whiskers give the 95% confidence intervals. The dashed horizontal denotes a coefficient of 0, i.e., a null effect.



4.5.2. Mechanism: Feature vs. Label

To understand how the intervention affects channel views, the next test assesses whether the effects are driven by (1) the feature or (2) the label. To do so, it utilizes the fact that not all channels were subjected to the feature intervention. Not all channels are featured in the search because YouTube implemented it only for some search terms, and thus, only channels whose videos catered to those search terms are featured. YouTube offers the search feature only for a subset of all searches for health-related topics, such that approximately 37% of the authoritative channels in the sample are never receiving a feature and only a label.

This allows to disentangle the effects of the label and the search feature. *Search feature* is 1 for all channels subject to the search feature and 0 otherwise. It is obtained from a manual search of the search terms after the introduction of the program. Although all channels that YouTube defined as authoritative received the label, not all were featured in search. Econometrically, this allows to modify equation 4.1 and introduce an interaction with the binary variable *Search feature*.

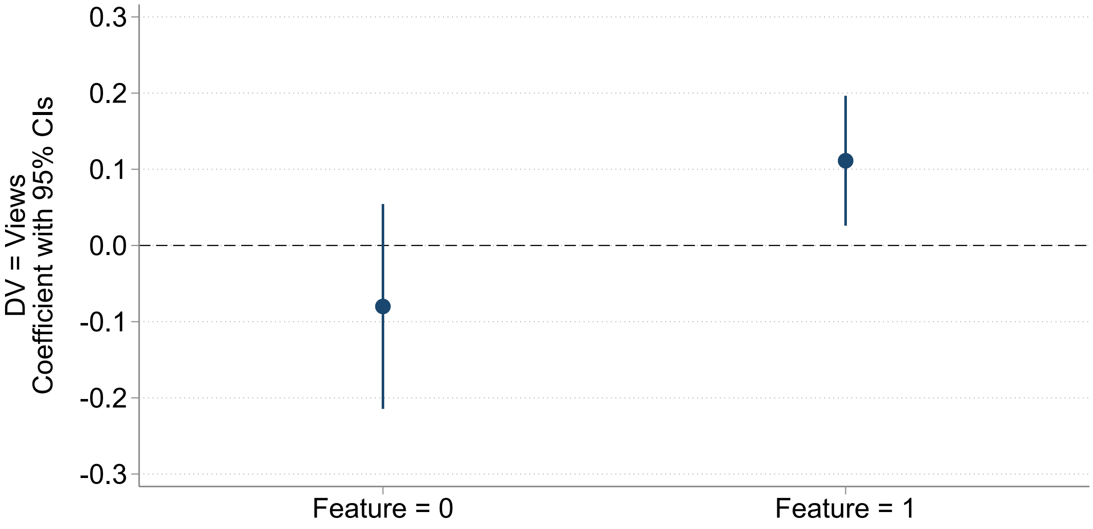
Figure 4.5 shows the resulting coefficient plot, again with the whiskers denoting 95% confidence intervals. The primary conclusion from the data is that the intervention's effects are primarily driven by the search feature and not the label. Authoritative channels that merely received the label – but not the feature – did not experience a statistically significant change in viewership. The confidence intervals are broad and include zero. By contrast, authoritative channels that received the feature and the label showed a considerable increase in views by about 11.8%. This result indicates that the changes in viewership result from the feature and that the label alone has no material effect.

4.5.3. Downstream Effects on Channel Subscribers and Video Production

Figure 4.6 shows the regression estimates for equation 4.1 with the dependent variables being *Log(Subscribers)* and *Log(Videos)*.

Line (1) shows the effects on the subscriber count. While the subscriber count normalized by views increases for non-authoritative channels, it declines for authoritative channels. Overall, the number of subscribers per view increases by 4.1% for non-authoritative channels, while for authoritative channels, it declines by 8%. Appendix Table A4.4 provides evidence that this is not an artifact of the measure but that the findings can be replicated using the absolute number of subscribers or the number of likes per view. Considering the mean number of subscribers in the control group of 819

Figure 4.5: Mechanism of the Effect on Channel Views: Feature vs. Label
Note: The figure shows the effect of YouTube's Authoritative Health Information program on channel views for channels admitted to the program, distinguishing between those that received no search feature (Feature = 0) and those that received a search feature (Feature = 1). The markers denote the estimate, and the whiskers give the 95% confidence intervals. The dashed horizontal denotes a coefficient of 0, i.e., a null effect. The estimates are documented in Appendix Table A4.2.



subscribers, subscribers declined by 8.75 for non-authoritative channels and 27.9 for authoritative channels compared to the control group. For authoritative channels, this can be traced back to a decline in the popularity of the contents; for the non-authoritative ones, this is instead an artifact of the decline in views as a result of the intervention.

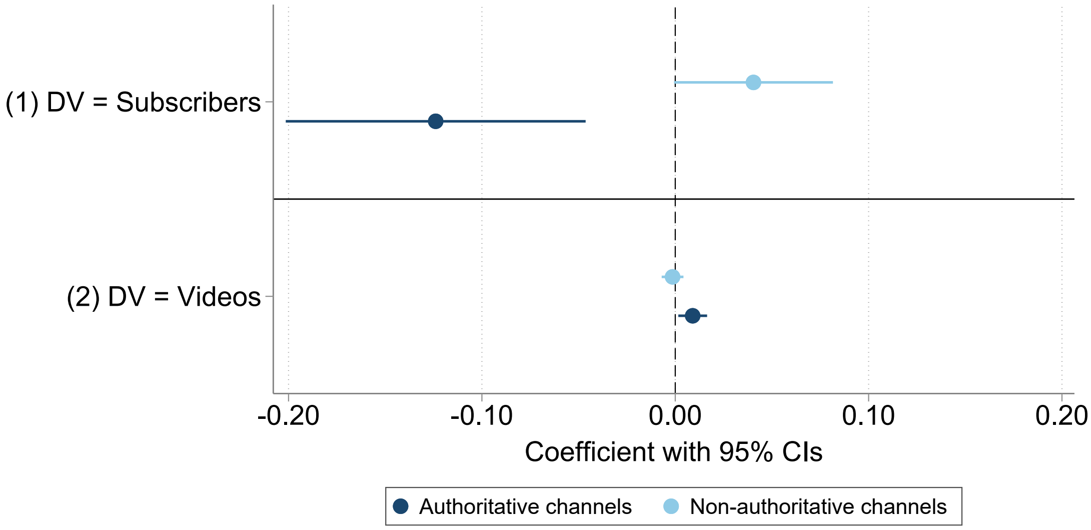
Line (2) shows the effects of the number of videos produced by authoritative vs. non-authoritative channels. There is no effect on non-authoritative channels and only a negligible increase of 0.8% in the number of videos for authoritative channels. In terms of economic significance, this leads to a rise in the number of videos by 0.0006 per channel per day (%increase x average video count in the control group). This effect seems relatively small and will not considerably increase the availability of authoritative content on the platform.

4.5.4. Robustness Checks

Selection Bias: Further analyses seek to rule out selection bias. In this analysis, selection bias could occur if systematic differences exist between the channels that apply for YouTube's Authoritative Health Information program and those that do not.

Figure 4.6: Downstream Consequences of YouTube's Authoritative Health Information Program on Channel Subscribers and Videos

Note: The figure shows the estimates for the coefficients of β_2 (light blue) and β_3 (dark blue) of equation 4.1 with subscribers and videos as the dependent variables. Line (1) shows the estimates for the number of subscribers per view. Line (2) shows the estimates for the number of videos. The markers denote the estimate, and the whiskers give the 95% confidence intervals. The dashed vertical denotes an effect size of 0, i.e., a null effect. The estimates are documented in Appendix Table A4.3.



This would then be carried over into the estimation.[6] To rule it out empirically, a robustness check is conducted in which channels are removed from the sample if they entered into the program after applying for it, thus keeping only those channels in the sample that were automatically – without application – defined by YouTube as authoritative. Appendix Table A4.5 shows that within this subsample, the estimates are highly similar to the baseline estimates. This shows that the findings of the study are not driven by channel characteristics that are correlated with the incentive to apply for the program but are associated with the intervention.

Within-Group Matching: To make the groups comparable, matching is performed on treated and control channels (German vs. French). However, since the analysis

[6] In general, there are a couple of ex-ante reasons why these differences should not influence the results. First, all channels were likely aware of the program and had the opportunity to apply, given that they fulfilled the criteria. The program had been announced by YouTube and was covered in high-circulation German newspapers, such as Die Welt or ZEIT (Welt 2022, ZEIT Online 2022). The program does not impose significant barriers to entry, such as high costs or complex application procedures. It relies on an online format and can be completed in a reasonable time. Second, YouTube provides clear information about the eligibility criteria and benefits of the program, such that channels are more likely to make informed decisions about whether to apply based on their alignment with the program's objectives rather than on factors related to viewership. Third, the channel fixed effects included in the regression model will likely correlate with omitted variables influencing both the decision to apply for the program and viewership outcomes.

considers within-group differences (authoritative vs. non-authoritative), it would bolster the findings if they account for potentially pre-existing within-group differences. Therefore, for an additional test, matching is performed separately for authoritative and non-authoritative channels on French control channels. As before, comparable groups are obtained (see Appendix Tables A4.6 and A4.7). The estimates from a standard difference-in-differences regression are displayed in Appendix Tables A4.8 and A4.9, confirming the results. The effect magnitude is slightly larger than in the baseline (i.e., +9.3% for authoritative channels and -3.8% for non-authoritative channels). Thus, the results are not artifacts of within-group differences between authoritative and non-authoritative channels.

Alternative Control Group: There may be concerns over the suitability of French channels as the control group. Since the control group serves as the counterfactual, it is crucial to ensure that the choice of the control group does not influence the results but that the results are robust to alternative choices. As outlined in the method section, French channels constitute a suitable control group, and spillover effects or confounding events are unlikely to induce bias in the estimation. However, as it cannot be ruled out that some unobserved factors reduce the comparability between the treatment and the control group, the analysis was also conducted using another control group. This alternative control group consists of Spanish and Italian health channels. Italy and Spain are well-suited control groups because, after France, they have the largest user base on YouTube in Europe (DataReportal et al. 2022) and share similar healthcare factors (OECD/European Union 2020). In line with the main approach, coarsened exact matching is applied to make the groups comparable (see Appendix Table A4.10). To obtain a larger set of channels to match from, Spanish and Italian channels are pooled. The results, displayed in Appendix Table A4.11, once again confirm the findings. The effects are consistent with the baseline (+4.3% for authoritative and -3.6% for non-authoritative channels). Taken together, the results are confirmed when using an alternative control group.[7]

4.6. Discussion

To advance the understanding of countermeasures against false information on online platforms, this study investigated a novel intervention by YouTube. The intervention

[7] One question concerns the influence of the feature's design, namely presenting the user with a horizontal carousel. However, users are likely familiar with navigating carousels because they have been adopted for other formats on YouTube (e.g., for *Shorts*), as well as on Google, Amazon, and Twitch, to name a few. Experimental evidence indicates that horizontal displays are easier to process for humans because they match the human binocular vision field (which is horizontal in direction) and, therefore, simplify eye movement and information processing (Deng et al. 2016).

promoted content from sources declared authoritative, by positioning their content prominently in the search results and adding a label certifying their authoritativeness. Previous research has mostly investigated the causes and spread of misinformation (e.g., Allcott and Gentzkow 2017, Pennycook and Rand 2021, Vosoughi et al. 2018). This study complements a nascent set of studies on countermeasures against false information on online platforms, especially those that investigated fact-checking, credibility nudges, crowd-based content inspection, source ratings, or inoculation, by understanding a novel intervention, namely promoting authoritative content (e.g., Berger et al. 2023, Borwankar et al. 2022, Hwang and Lee 2024, Kim et al. 2019, Roozenbeek et al. 2022).

The analysis shows that the intervention had significantly positive but minor effects on consuming authoritative content. Authoritative channels saw 4.3% more viewers. This suggests that the intervention was effective in nudging viewers toward sources that YouTube had declared as authoritative. Concurrently, the intervention appeared to have a mirror effect on non-authoritative channels, which experienced a decline in viewership by approximately 4.6%. This drop indicates a shift in audience preference away from less reliable sources, likely as a direct consequence of the intervention. Moreover, the analysis shows that the effects of the intervention are persistent over time, indicating that the observed changes in viewership are not merely transient but reflect a sustained shift of attention toward authoritative channels. From a theoretical perspective, this durability is crucial, as prior research has uncovered differences in the persistence of countermeasures against false information (Barrera et al. 2020, Berger et al. 2023). For example, in an online experiment involving Facebook posts about COVID-19, Berger et al. (2023) found that fact-checking only has short-term effects. The sustained effect also matters because, over time, this might lead to a more discerning audience less susceptible to misinformation in the long run.

However, statistical significance and persistence aside, compared to other interventions that promoted content on online platforms, the effects of YouTube's Authoritative Health Information program are minor. For example, Dewan et al. (2023) studied AirBnB's introduction of the AirBnB Plus seal and observed that it increased booking rates by 6.8%. Huang et al. (2022a) ran a field experiment in an image-sharing online community and found that featuring an image in the feed of users increased views of that image by about 34.0%. Bockstedt and Goh (2011) studied different attributes of eBay listings and observed that being featured at the top of the search results increases listing views by about 120.8%. Thus, the small magnitude of the effect underscores the conclusion that this intervention is not a panacea but actually has limited influence over directing information consumption toward authoritative content. Perhaps it is best used to complement other, perhaps regulative, interventions.

To further inform the understanding of the intervention, the study disentangles whether the feature or the label caused the change in viewership. The analysis reveals that the effect is driven solely by the feature, and that the label has no effect. Therefore, the study suggests that simply labeling content as authoritative is insufficient, despite the positive effects observed for quality certifications and badges on platforms (e.g., Dewan et al. 2023, Hui et al. 2007, Oezpolat et al. 2013). This also hints at a lack of discoverability of a signal if only presented on the content-level but not in the search results. Indirectly, the findings provide evidence for the hypothesis that labels might not be effective in reducing engagement with misinformation, especially when not accompanied by appropriate user training (Bradshaw et al. 2021, Kim et al. 2019, Moravec et al. 2023). In this instance, it cannot be assessed whether two-sided labeling – i.e., not only labeling authoritative but also non-authoritative channels – would have been more effective. Prior research suggests that one-sided interventions can create user uncertainty about their presence and non-presence, thereby reducing their effectiveness (Pennycook et al. 2020). However, the observation that the feature drives the effect suggests that facilitating the discovery of authoritative content is perhaps a promising strategy in combating misinformation. Building upon this study, future research should further develop various strategies to enhance access to authoritative content. This includes examining different algorithms for search result prioritization, user interface designs that promote reliable information, and personalized recommendation systems that favor authoritative sources. Taken together, the findings emphasize that merely labeling content as authoritative is insufficient to attract attention if users cannot easily find it.

The findings also point out that the intervention did not encourage authoritative channels to produce considerably more content, nor did it discourage non-authoritative channels. Specifically, authoritative channels increased their content production by a mere 0.8%, while non-authoritative channels did not significantly decrease theirs. These negligible changes suggest that the intervention's impact on content creation is minimal. One possible explanation for these null effects is that the observed changes in viewership were too small to serve as a strong incentive for channels to alter their production strategies. Nevertheless, the data is limited in explaining these results. However, there are hints that the quality of contributions declines as authoritative channels relax their efforts after obtaining the certification. Future research should further understand these aspects by surveying channels about their decision-making processes or conducting in-depth interviews.

4.7. Conclusion

Considering the vast number of countermeasures against fakery that focus on pruning fakery, the second empirical study of this dissertation investigated whether the complementary approach of promoting credible content positively affects its consumption and production. The YouTube Authoritative Health Information program introduced in Germany in 2023 presents the empirical setting where authoritative channels are featured in the search results and receive a label for their videos indicating content authoritativeness. The findings indicate an increase in the views of authoritative content by 4.3% at the expense of non-authoritative content driven by the search feature, not the label. In contrast, there is no considerable effect on content production. These findings show a small overall impact of the program but highlight the need to place credible content more dominantly. As the search feature seems promising, further measures should be explored to make access to credible content more intuitive.

There are two final remarks on the implementation of the intervention. One angle concerns the costs. The intervention comes with several expenses, such as reviewing the applications for the program, quality monitoring, and reputational risk. Thus, the positive effects observed here also need to be evaluated in light of the program's costs. Another angle concerns the eligibility criteria. A general trade-off for platform firms will reside regarding how to define eligibility for such a program and to navigate a potentially thin line between protection from false information and free speech (Van Alstyne et al. 2023). Given the eligibility criteria defined in YouTube's intervention in particular, only a fraction of channels get certified as authoritative, which raises the question of whether there are more – but not eligible – channels that do produce factually correct information. This could also affect the variety of authoritative content.

4.8. Appendix for Chapter

4.8.1. Tables & Figures

Table A4.1: Effects of YouTube's Authoritative Health Information Program on Channel Views

Note: The table tests for the effects on channels' views. Column (1) tests for the baseline estimation. Column (2) excludes outliers of the dependent variable. Column (3) winsorizes the dependent variable. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)
	Baseline	Outliers removed	Winsorized
After	-0.154***	-0.152***	-0.148***
	(0.021)	(0.019)	(0.019)
Treat × After	-0.047*	-0.044*	-0.048**
	(0.020)	(0.018)	(0.018)
Authoritative × Treat × After	0.089*	0.080*	0.085*
	(0.039)	(0.038)	(0.038)
Constant	4.170***	3.896***	4.112***
	(0.013)	(0.012)	(0.012)
Observations	433,968	412,272	433,968
Channel Fixed Effects	X	X	X
Time Fixed Effects	X	X	X
Adjusted-Within R²	0.028	0.032	0.030
F-Test	77.272***	68.328***	63.919***

Table A4.2: Mechanism of the Effect on Channel Views: Feature vs. Label
Note: The table tests for the mechanism by exploiting the effect of the search feature. The search feature only applies to a subset of channels, as it was only introduced for certain search terms. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1) Log(Views)
After	-0.160***
	(0.022)
Treat × After	-0.042^{+}
	(0.026)
Authoritative × Treat × After	-0.038
	(0.070)
Search feature × After	0.012
	(0.026)
Search feature × Treat × After	-0.011
	(0.041)
Search feature × Authoritative × Treat × After	0.203*
	(0.084)
Constant	4.170***
	(0.013)
Observations	433,968
Channel Fixed Effects	X
Time Fixed Effects	X
Adjusted-Within R²	0.029
F-Test	75.813***

Table A4.3: Downstream Consequences of YouTube's Authoritative Health Information Program on Channel Subscribers and Videos

Note: The table tests for the effects on channels' subscribers and videos. Column (1) tests for the number of subscribers per view. Column (2) tests for the number of videos. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1) Log(Subscribers)	(2) Log(Videos)
After	0.349***	-0.007
	(0.021)	(0.005)
Treat × After	0.040^{+}	-0.001
	(0.021)	(0.003)
Authoritative × Treat × After	-0.124**	0.009*
	(0.040)	(0.004)
Constant	2.446***	0.058***
	(0.013)	(0.003)
Observations	401,878	434,214
Channel Fixed Effects	X	X
Time Fixed Effects	X	X
Adjusted-Within R²	0.046	0.002
F-Test	86.598***	2.711***

Table A4.4: Robustness Checks: Channel Popularity
Note: The table tests for the effects on channels' subscribers and likes. Column (1) tests for the absolute number of subscribers. Column (2) tests for the daily number of likes in relation to the number of views (based on the last 20 videos). OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1) Log(Subscribers)	(2) Log(Likes)
After	0.170***	-0.026
	(0.007)	(0.031)
Treat × After	-0.011^{+}	0.036*
	(0.006)	(0.017)
Authoritative × Treat × After	-0.024**	-0.128**
	(0.007)	(0.040)
Constant	6.615***	-3.471***
	(0.003)	(0.019)
Observations	445,421	127,588
Channel Fixed Effects	X	X
Time Fixed Effects	X	X
Adjusted-Within R²	0.169	0.009
F-Test	9.130***	5.521***

Table A4.5: Robustness Checks: Selection Bias
Note: The table excludes channels that had to apply for the program. Column (1) tests for the number of views. Column (2) tests for the mechanism. Column (3) tests for the number of subscribers per view. Column (4) tests for the number of videos. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)
	Log(Views)	Log(Views)	Log(Subscribers)	Log(Videos)
After	-0.157***	-0.162***	0.352***	-0.007
	(0.021)	(0.022)	(0.021)	(0.005)
Treat × After	-0.047*	-0.042^{+}	0.040^{+}	-0.001
	(0.020)	(0.026)	(0.021)	(0.003)
Authoritative × Treat × After	0.071^{+}	-0.038	-0.115**	0.010**
	(0.041)	(0.070)	(0.042)	(0.004)
Search feature × After		0.012		
		(0.026)		
Search feature × Treat × After		-0.011		
		(0.041)		
Search feature × Authoritative × Treat × After		0.186*		
		(0.085)		
Constant	4.155***	4.155***	2.448***	0.058***
	(0.013)	(0.013)	(0.013)	(0.003)
Observations	431,733	431,733	399,643	431,983
Channel Fixed Effects	X	X	X	X
Time Fixed Effects	X	X	X	X
Adjusted-Within R²	0.028	0.028	0.046	0.002
F-Test	76.486***	75.020***	85.735***	2.699***

Table A4.6: Test for Group Differences for Authoritative Channels
Note: The table tests for differences in means between treated and control channels before and after the matching only for authoritative channels. Column “Difference in Means” reports the t-test. Column “Difference in Trends” reports a regression estimate for the difference in time trend before the treatment. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	Before Matching				After Matching			
	Control	Treatment	Difference in Means	Difference in Trends	Control	Treatment	Difference in Means	Difference in Trends
Subscribers	17,048.077	15,683.677	1,364.400	3.037	4,477.332	3,412.021	1,065.311	3.697
Views	2,010,968.110	3,017,516.033	-1,006,547.923	-3,652.312	745,492.200	1,050,707.475	-305,215.276	276.585
Videos	131.883	113.014	18.869	-0.052*	90.101	93.626	-3.526	-0.017
Age	2,192.058	2,521.706	-329.648***	-	2,453.933	2,494.643	-40.710	-
Obs.	1,560	228			224	224		

Table A4.7: Test for Group Differences for Non-Authoritative Channels
Note: The table tests for differences in means between treated and control channels before and after the matching only for non-authoritative channels. Column "Difference in Means" reports the t-test. Column "Difference in Trends" reports a regression estimate for the difference in time trend before the treatment. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	Before Matching				**After Matching**			
	Control	Treatment	Difference in Means	Difference in Trends	Control	Treatment	Difference in Means	Difference in Trends
Subscribers	17,048.077	12,143.118	4,904.959^{+}	-1.009	11,051.032	8,936.558	2,114.474	-0.216
Views	2,010,968.110	2,381,874.343	-370,906.233	833.823	1,187,953.037	1,378,423.204	-190,470.167	480.044
Videos	131.883	137.667	-5.785	0.011	125.519	126.939	-1.420	0.020
Age	2,192.058	2,286.824	-94.766*	-	2,193.703	2,192.831	0.872	-
Obs.	1,560	1,931			1,513	1,513		

Table A4.8: Robustness Checks: Within-Group Matching (Authoritative Channels)
Note: The table tests for the separate matching for authoritative channels. Column (1) tests for the number of views. Column (2) tests for the mechanism. Column (3) tests for the number of subscribers per view. Column (4) tests for the number of videos. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}\ p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

	(1)	(2)	(3)	(4)
	Log(Views)	Log(Views)	Log(Subscribers)	Log(Videos)
After	-0.241***	-0.208***	0.354***	-0.012
	(0.048)	(0.052)	(0.051)	(0.011)
Treat × After	0.089*	-0.019	-0.112*	0.014**
	(0.042)	(0.068)	(0.044)	(0.005)
Search feature × After		-0.081		
		(0.061)		
Search feature × Treat × After		0.206*		
		(0.089)		
Constant	4.284***	4.284***	2.279***	0.032***
	(0.032)	(0.032)	(0.033)	(0.006)
Observations	63,864	63,864	60,561	63,886
Channel Fixed Effects	X	X	X	X
Time Fixed Effects	X	X	X	X
Adjusted-Within R²	0.046	0.048	0.060	0.005
F-Test	42.090***	41.634***	45.298***	2.266***

Table A4.9: Robustness Checks: Within-Group Matching (Non-Authoritative Channels)
Note: The table tests for the separate matching for non-authoritative channels. Column (1) tests for the number of views. Column (2) tests for the mechanism (p=0.136 for Treat × After). Column (3) tests for the number of subscribers per view (p=0.162 for Treat × After). Column (4) tests for the number of videos. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1)	(2)	(3)	(4)
	Log(Views)	Log(Views)	Log(Subscribers)	Log(Videos)
After	-0.150***	-0.156***	0.346***	-0.007
	(0.021)	(0.023)	(0.021)	(0.005)
Treat × After	-0.039*	-0.037	0.028	-0.001
	(0.019)	(0.025)	(0.020)	(0.003)
Search feature × After		0.013		
		(0.026)		
Search feature × Treat × After		-0.006		
		(0.040)		
Constant	4.196***	4.196***	2.479***	0.060***
	(0.013)	(0.013)	(0.013)	(0.003)
Observations	430,524	430,524	399,870	430,762
Channel Fixed Effects	X	X	X	X
Time Fixed Effects	X	X	X	X
Adjusted-Within R²	0.026	0.026	0.043	0.002
F-Test	77.850***	76.814***	86.164***	2.564***

Table A4.10: Test for Group Differences with Alternative Control Group
Note: The table tests for differences in means between treated and control channels before and after the matching only for a control group consisting of Spanish and Italian channels. Column “Difference in Means” reports the t-test. Column “Difference in Trends” reports a regression estimate for the difference in time trend before the treatment. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

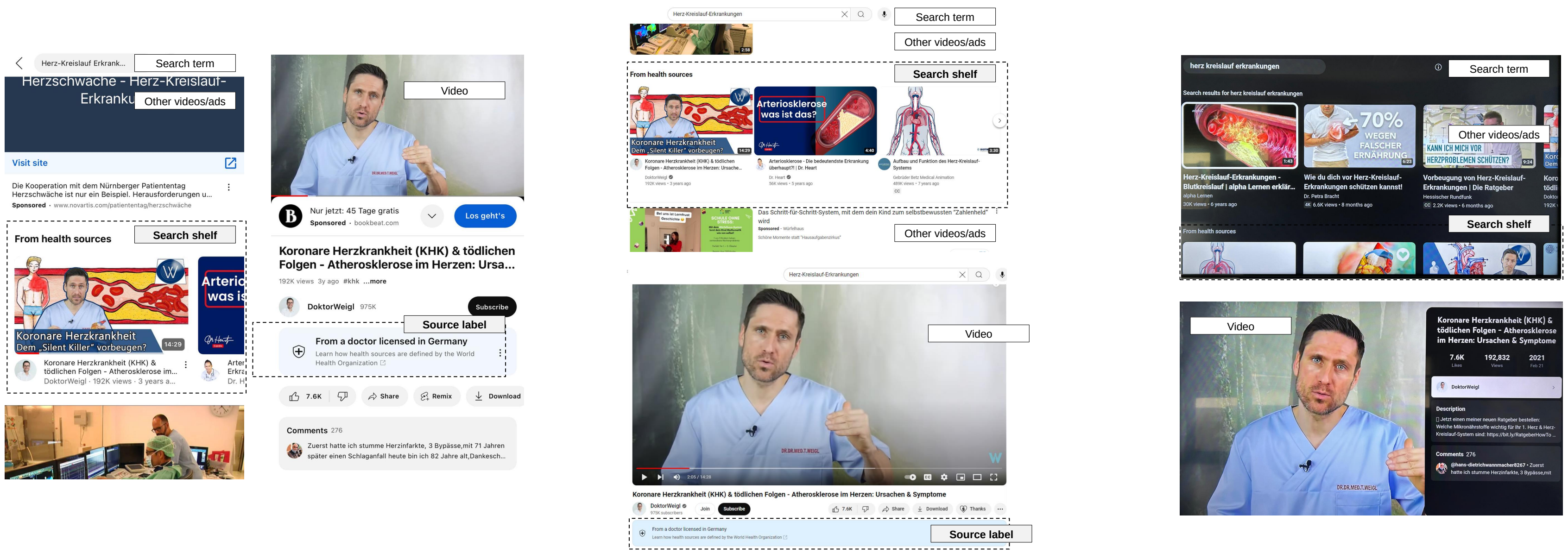
	Before Matching				**After Matching**			
	Control	Treatment	Difference in Means	Difference in Trends	Control	Treatment	Difference in Means	Difference in Trends
Subscribers	26,419.915	12,506.884	13,913.031***	-8.979**	10,591.771	8,902.634	1,689.137	-1.465
Views	3,642,109.926	2,446,814.249	1,195,295.677^{+}	-574.056	1,590,569.301	1,764,643.548	-174,074.248	846.501
Videos	171.763	135.031	36.732***	-0.032*	122.028	122.184	-0.156	-0.012
Age	2,486.758	2,311.385	175.373***	-	2,311.845	2,310.732	1.113	-
Obs.	2,792	2,161			2,114	2,114		

Table A4.11: Robustness Checks: Choice of Control Group (ES/IT)
Note: The table tests for an alternative control group of Spanish and Italian health channels. Column (1) tests for the number of views. Column (2) tests for the mechanism (p=0.174 for Treat × After; p=0.110 for Search feature × Authoritative × Treat × After). Column (3) tests for the number of subscribers per view (p=0.134 for Treat × After). Column (4) tests for the number of videos. OLS estimates. Heteroskedasticity-robust standard errors in parentheses. $^{+}\ p < 0.1$, $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

	(1)	(2)	(3)	(4)
	Log(Views)	Log(Views)	Log(Subscribers)	Log(Videos)
After	-0.141***	-0.153***	0.327***	-0.006
	(0.019)	(0.022)	(0.020)	(0.004)
Treat × After	-0.037*	-0.031	0.028	0.002
	(0.018)	(0.023)	(0.019)	(0.003)
Authoritative × Treat × After	0.080*	0.010	-0.110***	0.008*
	(0.033)	(0.059)	(0.033)	(0.003)
Search feature × After		0.028		
		(0.027)		
Search feature × Treat × After		-0.016		
		(0.037)		
Search feature × Authoritative × Treat × After		0.113		
		(0.070)		
Constant	4.219***	4.226***	2.362***	0.066***
	(0.012)	(0.012)	(0.012)	(0.003)
Observations	603,241	600,905	556,740	603,484
Channel Fixed Effects	X	X	X	X
Time Fixed Effects	X	X	X	X
Adjusted-Within R²	0.025	0.025	0.039	0.003
F-Test	121.334***	119.232***	128.783***	3.713***

Figure A4.1: Authoritative Health Information Program on Various Devices

Note: The figure illustrates the intervention with the search feature and the label for the different devices. Panel (A) shows the intervention for mobile devices with a health carousel and a label below the video. Panel (B) shows the intervention for desktops with a health carousel and a label below the video. Panel (C) shows the intervention for TVs with a health carousel but no video label.



Panel (A): Mobile

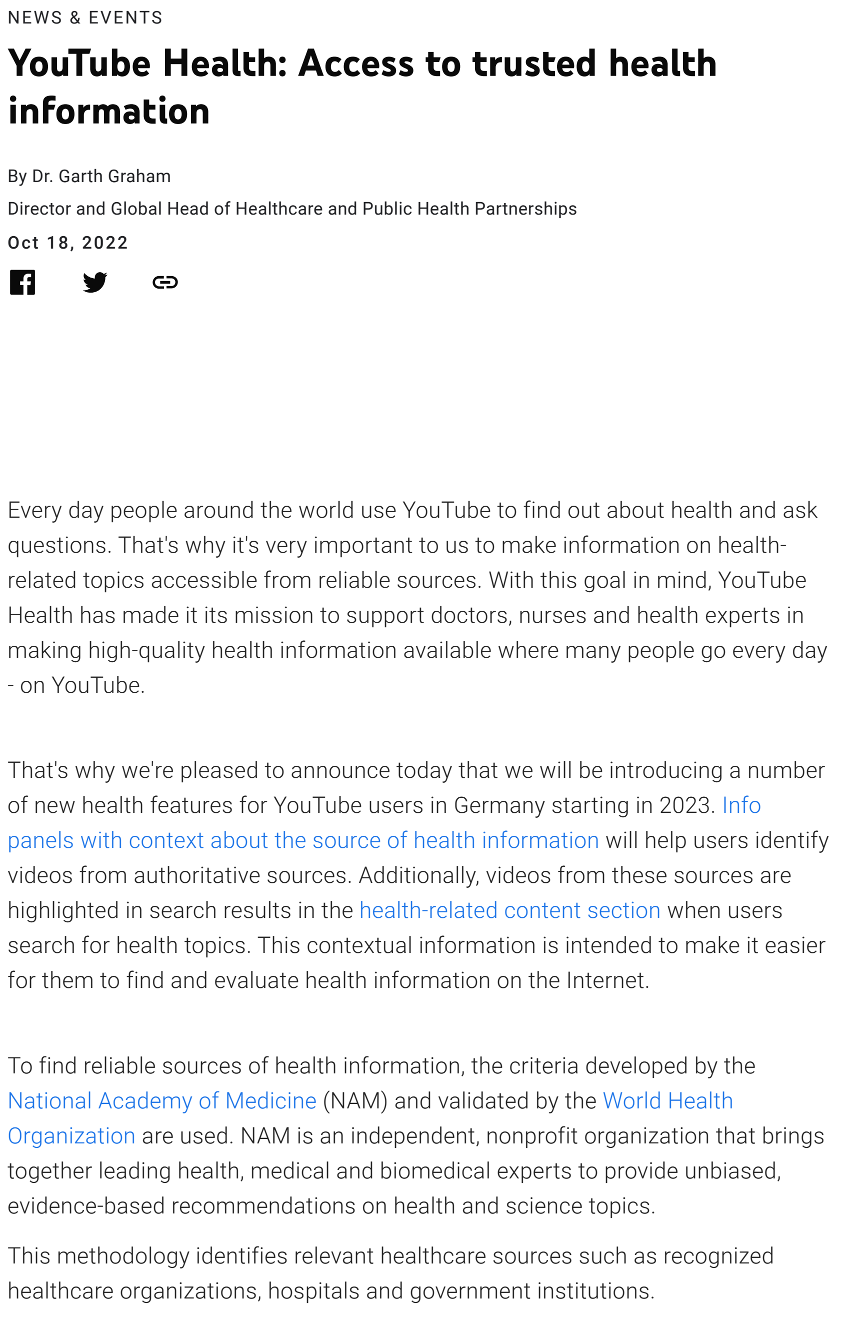
Panel (B): Desktop

Panel (C): TV

Search term
Other videos/ads
Search shelf
Video
Source label
From health sources
Visit site
From a doctor licensed in Germany
Learn how health sources are defined by the World Health Organization
Koronare Herzkrankheit (KHK) & tödlichen Folgen - Atherosklerose im Herzen: Ursa...
DoktorWeigl
Subscribe
Share
Remix
Download
Comments 276
herz kreislauf erkrankungen
Search results for herz kreislauf erkrankungen

4.8.2. Further Information on Method and Data

B4.1: Blog Post on YouTube's Authoritative Health Information Program



NEWS & EVENTS

YouTube Health: Access to trusted health information

By Dr. Garth Graham

Director and Global Head of Healthcare and Public Health Partnerships

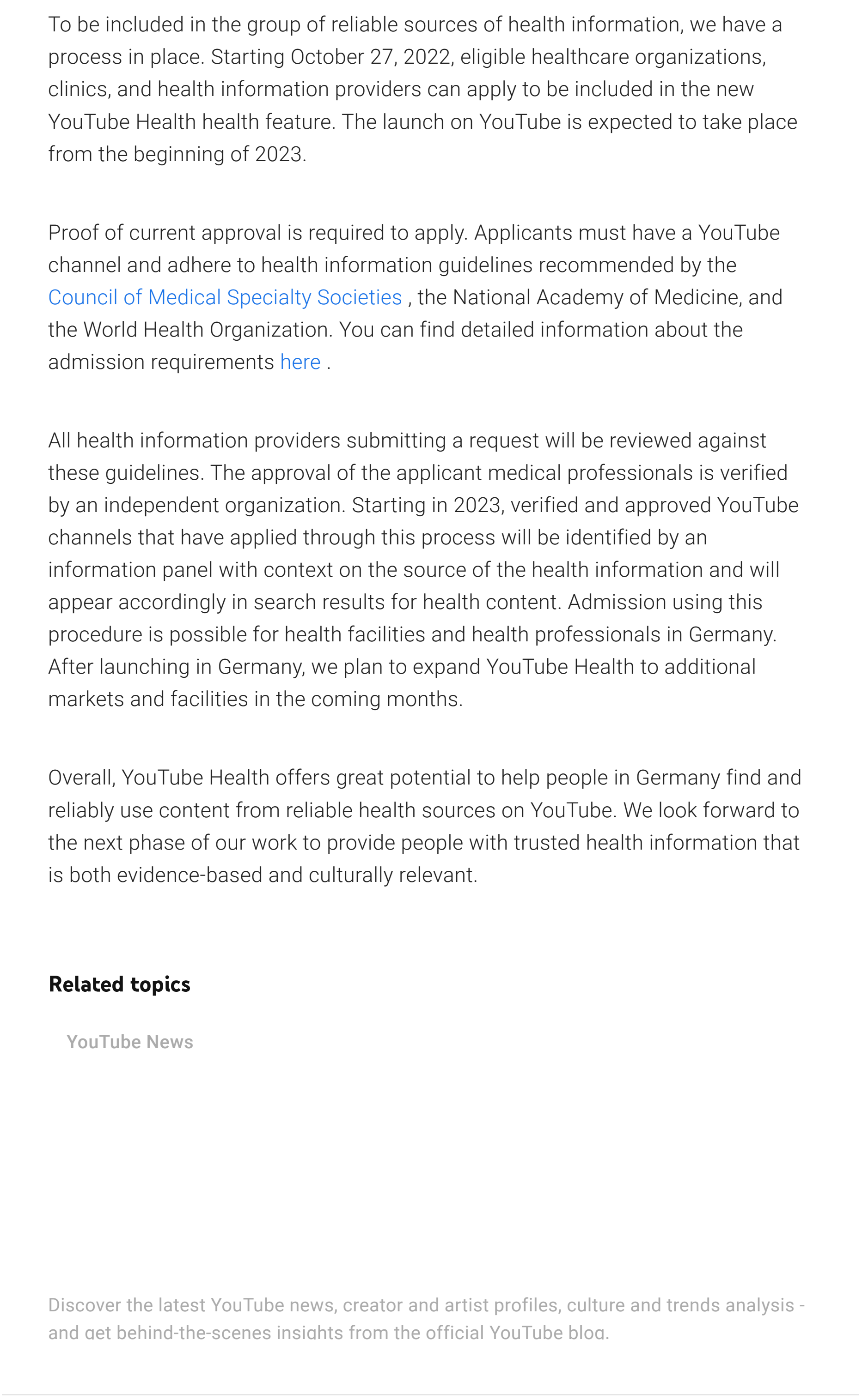
Oct 18, 2022

Every day people around the world use YouTube to find out about health and ask questions. That's why it's very important to us to make information on health-related topics accessible from reliable sources. With this goal in mind, YouTube Health has made it its mission to support doctors, nurses and health experts in making high-quality health information available where many people go every day - on YouTube.

That's why we're pleased to announce today that we will be introducing a number of new health features for YouTube users in Germany starting in 2023. Info panels with context about the source of health information will help users identify videos from authoritative sources. Additionally, videos from these sources are highlighted in search results in the health-related content section when users search for health topics. This contextual information is intended to make it easier for them to find and evaluate health information on the Internet.

To find reliable sources of health information, the criteria developed by the National Academy of Medicine (NAM) and validated by the World Health Organization are used. NAM is an independent, nonprofit organization that brings together leading health, medical and biomedical experts to provide unbiased, evidence-based recommendations on health and science topics.

This methodology identifies relevant healthcare sources such as recognized healthcare organizations, hospitals and government institutions.



To be included in the group of reliable sources of health information, we have a process in place. Starting October 27, 2022, eligible healthcare organizations, clinics, and health information providers can apply to be included in the new YouTube Health health feature. The launch on YouTube is expected to take place from the beginning of 2023.

Proof of current approval is required to apply. Applicants must have a YouTube channel and adhere to health information guidelines recommended by the Council of Medical Specialty Societies , the National Academy of Medicine, and the World Health Organization. You can find detailed information about the admission requirements here .

All health information providers submitting a request will be reviewed against these guidelines. The approval of the applicant medical professionals is verified by an independent organization. Starting in 2023, verified and approved YouTube channels that have applied through this process will be identified by an information panel with context on the source of the health information and will appear accordingly in search results for health content. Admission using this procedure is possible for health facilities and health professionals in Germany. After launching in Germany, we plan to expand YouTube Health to additional markets and facilities in the coming months.

Overall, YouTube Health offers great potential to help people in Germany find and reliably use content from reliable health sources on YouTube. We look forward to the next phase of our work to provide people with trusted health information that is both evidence-based and culturally relevant.

Related topics

YouTube News

Discover the latest YouTube news, creator and artist profiles, culture and trends analysis - and get behind-the-scenes insights from the official YouTube blog.

5. Discussion

5.1. Main Findings of this Dissertation

This dissertation comprises a literature review and two follow-up empirical studies contributing to the discussion on platform fakery in various ways. The main findings from the different studies are as follows.

The literature review provides a holistic understanding of platform fakery and identifies open questions in the research field. The study aimed to structure the existing literature along the sender-receiver framework (Shannon and Weaver 1949) and formulate open research questions. Overall, the study identifies an increasing research interest in fakery on digital platforms with a strong focus on empirical studies. The content-related insights include (1) the forms of fakery, (2) the senders of fakery, (3) the receivers of fakery, (4) the consequences of fakery, and (5) the countermeasures against fakery. Various research questions remain open that refer, among others, to fake accounts and technological advancements that are responsible for new dimensions of fakery, the role of platforms in tolerating and enhancing fakery, characteristics of sources, targets, and messages in the context of fakery, unintended fakery and unintended consequences of measures to combat, and countermeasures that can complement existing measures. The empirical studies in this dissertation investigate the research gaps on (1) the use of fake followers and its implications and (2) the effectiveness of a novel countermeasure for promoting credible content. The choice for these research questions is driven by the lack of scientific evidence, but also the public discourse related to these topics.

The first empirical study on fake followers tackles the research gap of missing evidence on fake followers and their associated risks. While news reporting highlights the challenge of fake followers on digital platforms, current research does not fully understand the phenomenon and how it manifests. This study uses Twitter's purge of fake followers in July 2018 as a unique empirical opportunity to investigate the extent to which fake followers are used by firms, as well as the potential risks associated with fake followers. Based on a data set of firms of the S&P 1500 Composite Index, only about 1.2% of firms' followers are fake, with a higher proportion of fake followers for smaller firms and firms facing greater competitive pressure. Investigating the risk of fake followers for shareholders, the cumulative abnormal returns to a firm's stock price decline by 0.078% for every percentage point in revealed fake follower shares as investors adapt their valuation once they become aware of the actual social media reach of firms. Opposing the general public perception, this study shows that fake follower use is less pronounced than expected, and the risk associated with it is relatively low.

The second empirical study on a novel countermeasure for misinformation, namely the promotion of credible content, tackles the lack of understanding of solutions that improve access to credible content instead of pruning inaccurate content. YouTube introduced the Authoritative Health Information program in Germany in February 2023. This intervention manifests in (1) the featuring of authoritative channels on top of the search results and (2) the provision of labels of authoritativeness on the video level. Using this intervention as a treatment, the study investigates its effectiveness in shifting views toward credible content. The findings from the difference-in-difference-in-differences estimation reveal that views for authoritative channels increase by 4.3% after the intervention while they decline by 4.6% for non-authoritative channels. This is driven by the ease of discovery via the search feature, not the video label. However, the intervention does not alter video production—while there is no effect for non-authoritative channels, there is a negligible increase by 0.8% for the videos of authoritative channels but a decline in subscribers by views by 8%. These findings show a small positive effect of the intervention, with ease of discovery being the most promising approach, and indicate its complementary nature to other measures. At the same time, they indicate a negligible effect on the quantity of content production, but a potentially negative effect on content quality.

Considering the literature review as the starting point of this dissertation, the studies provide new insights into the overall challenge of fakery on digital platforms. In particular, the two empirical studies provide evidence that fake followers present a minor challenge for distorting shareholders and that it requires the combination of various countermeasures to combat fakery. As the literature review hints at future research concerning new types of misinformation, countermeasures that make access to credible content more intuitive can be valuable strategies to complement approaches that focus on pruning inaccurate content. However, future research is required to identify how such measures can be designed to increase their effectiveness.

5.2. Theoretical Contributions

This dissertation advances various streams of the information systems literature with the new dimension of fakery on digital platforms.

First, the dissertation contributes to the literature on platform management and governance (e.g., Boudreau 2010, Foerderer 2020, Katz and Shapiro 1985). Following the platform literature, the value of a platform is derived from a large user base, therefore substantially benefiting from the integration of various users (Engert et al. 2023, Gandal et al. 2000, Katz and Shapiro 1985). However, in line with prior studies, lower control can lead to harm to the platform quality (Boudreau 2010, Candogan and

Drakopoulos 2020, Huang et al. 2022b). These studies require platforms and policymakers to better understand the extent of the challenge of fakery to assess the need for intervention. At the same time, it becomes increasingly important to understand how platforms can balance openness that fosters free speech and platform credibility.

The dissertation advances this literature in two ways. Regarding the phenomenon of fake followers that has received little attention from research in the past, the dissertation shows that only a minor proportion of firms' followers are fake. In addition, it shows that the financial risk associated with fake followers in terms of shareholder reactions is limited. This makes fake followers a minor topic in the information systems discussion. Considering the trade-off between openness and quality control, the study on YouTube indicates that the ease of finding credible content enhances its consumption. However, the study also provides evidence that it is not a universal solution and should only be used as a complementary measure. In particular, it sheds light on the need to set incentives accordingly in order to enhance the production of high-quality content.

The findings of this dissertation highlight that the existence of fakery is both a problem of platform governance and user behavior on these platforms. On the one hand, platform providers do not sufficiently monitor and detect fakery, so it remains on the platform. On the other hand, users do not fully incorporate quality signals provided by the platforms, e.g., quality labels that indicate content credibility. This dissertation suggests that besides the explicit statement for quality, platforms need to implicitly nudge users toward credible content to make it easier and intuitive to access.

Second, the dissertation contributes to the literature on social media use and engagement (e.g., Chen et al. 2015a, 2018, Lee et al. 2018a, Luo et al. 2013, Yan and Tan 2014). The dissertation explores content consumers and providers regarding the use of and reaction to fakery. From the consumer side, digital platforms, in particular social media, are among others used to obtain information (Khan 2017, Yan and Tan 2014). Fakery becomes increasingly challenging to tackle in such networks as their novel nature makes users likely to disseminate them further in the network (London Jr et al. 2022, Vosoughi et al. 2018). This is why the dissertation assesses how measures can reduce exposure to fakery, proactively reducing adverse effects. From the supply side, fakery can be exploited to benefit from a misinformed community that relies on the information and takes decisions in favor of the sender of fakery (e.g. Clarke et al. 2021, Lappas et al. 2016, Ullah et al. 2014). Considering the incentives in line with the positive effects of social media on content consumers' behavior (e.g. Deng et al. 2018, Hennig-Thurau et al. 2015, Kumar et al. 2016, Lacka et al. 2022, Wang et al. 2021b), this dissertation investigates the incentives to engage in fakery and the extent to which this manifests on digital platforms.

The dissertation complements the discussion on social media engagement with the perspective of fakery. Considering the consumption of fakery, the dissertation shows that investors' reactions to fake followers are minor, making them less susceptible to this deceptive behavior. Regarding health misinformation, the study does not directly assess the consumption of fakery but shows that the consumption of non-promoted content declines once credible content is promoted. Also, there are some insights into the supply of fakery. In the empirical study on Twitter, firms are shown to use fake followers to a small degree, with a higher proportion of fake follower use for firms facing higher competitive intensity and small firms. In contrast, no shift is observed in content production once platforms provide incentives for content contributors by placing them more prominently. However, this is potentially related to the small degree of incentives.

The findings of this dissertation bring forward the understanding of the supply and demand for fakery. Users react to fakery in different ways. They incorporate fakery in their decision-making; however, in the sense of fake followers, this presents only a minor risk. In addition, an implicit nudge that shifts users' attention can be a promising means to reduce the negative effect of fakery on decision-making. From a supply perspective, the dissertation highlights an incentive for content creators to engage in fakery, particularly when their positioning is relevant from an economic standpoint. However, providing incentives to promote more credible content is not effective, potentially driven by the low degree to which the incentives are provided.

Third, the dissertation refers to the societal impact of digital platforms and regulatory intervention. Various studies show the adverse effects of fakery for both individuals and society as a whole (Barone and Miniard 1999, Cantarella et al. 2023, Cho et al. 2011, Shi et al. 2022). In the focus of prior literature are platform-owned approaches to overcome the challenge (e.g., Kim et al. 2019, Moravec et al. 2020, 2023, Pennycook and Rand 2019). However, despite effectiveness in experiments, it is unclear whether platforms reliably integrate such measures, mainly because there are incentives to tolerate fakery (Candogan and Drakopoulos 2020). This dissertation investigates whether the platform self-regulation is effective or whether regulatory pressure is required.

The dissertation adds to this stream of literature by empirically assessing two examples of platform interventions intended to combat forms of fakery. Both interventions, the purge of fake accounts on Twitter and the promotion of authoritative channels on YouTube, have significant effects on platform quality—the purge of Twitter accounts leads to the removal of tens of millions of accounts, and the health information effectively shifts attention from non-authoritative to authoritative channels. However, both interventions are preceded by some external pressure, i.e., events that made it necessary for the platform firms to intervene. This hints at the need for extrinsic motivation for platforms to overcome the challenge of fakery.

The findings of this dissertation provide evidence that platforms do not necessarily engage in countermeasures for misinformation voluntarily but that some external pressure is often required to reach positive outcomes associated with platform initiatives. In that sense, the incentives for platforms should be aligned with societal goals to make them intervene. At the same time, having access to data and tracing the content on platforms is essential to validate the outcomes of interventions.

5.3. Practical Contributions

This dissertation provides implications that practitioners can use to shape the environment around digital platforms.

First, this dissertation informs platform providers in terms of governance and design. Based on the insights of the first empirical study, there is only a limited need to investigate further attempts to identify fake followers as both their extent and implications seem limited. However, as prior research indicates, during highly controversial societal situations, there can be some need to counter social bots that seem more powerful in shaping online discourse. In contrast to the findings on fake followers, the second empirical study sheds light on a complementary intervention to combat misinformation. There is some understanding of how users access content, indicating that platforms should promote credible content to make its consumption more intuitive and of lower effort. This helps to ultimately increase the demand for credible content and the overall platform quality. At the same time, apart from the formal qualifications of channels as a quality signal, platform providers should consider further eligibility criteria to maintain a diversity of topics on the platform.

Second, this dissertation informs platform users about the challenges associated with fakery. The literature review thoroughly explains how different forms of fakery emerge and how they influence behavior. Assessing the source of content and validating its expertise can be a promising measure to distinguish credible content from fakery from a platform perspective, however, requires users to be aware of such signals and trust them. On this account, this dissertation encourages social media users to rely on quality cues implemented by platforms to improve the quality of content they access. At the same time, it outlines that these measures are insufficient in tackling the challenge of misinformation and that users' literacy is required to curb the spread of fakery. Platform users should critically question content and not solely rely on the expertise of platform providers to monitor.

Third, this dissertation provides policymakers insights regarding platform regulation and user literacy. The findings indicate the tendency of platforms to intervene against fakery as a response to outside pressure. As a result, there is a need to define an environment in which platform providers' incentives are aligned with societal goals, potentially by the regulatory body. In addition, as users are constantly exposed to fakery and it is impossible to monitor every piece of information on the platform, there is a need to tackle the challenge of literacy on digital platforms. Policymakers should investigate how society can be equipped with the resources to differentiate between high- and low-quality content on digital platforms.

5.4. Limitations and Future Research

This dissertation has some limitations regarding contents and methodology and paves the way for future research.

First, the literature review indicates several gaps in the current body of literature, and this dissertation could only investigate two major gaps. With fakery being an emerging topic in the literature and a constantly evolving challenge in practice, the pace of technological advancements continually requires the revision of prior work and an extension of existing theoretical and practical insights. New forms of fakery that are increasingly difficult to detect require updating governance measures to combat them. At the same time, some questions that need to be answered to better understand the stakeholders involved in the uncertainty setting on the platform remain open. Future research should advance information systems research area by taking the literature review as a starting point for investigating open questions in the field.

Second, the methodological focus of this dissertation is limited to specific platforms, and findings could have slightly different characteristics when carried out in other settings, e.g., because the applications differ depending on the respective platform (Pelletier et al. 2020). Facing the large variety of platforms on the market that each fulfill different purposes, the generalizability of the findings is given to a limited extent. Even though the studies align with theory and advance current literature, there is still the possibility that some behavior is unique for a specific purpose of the platform. For example, while Twitter is mainly used for information purposes, Instagram or Facebook serve mostly social desires (Pelletier et al. 2020). The studies in this dissertation investigate the informational component of digital platforms. However, it is unclear how the more hedonic nature of some platforms influences the validity of the findings, mainly if individuals are not interested in the credibility of information or are unwilling to double-check. Follow-up studies should consider the methodological variety and investigate the topic from various angles.

Third, the choice of the research design is a critical factor in deriving the findings for the studies. The literature review relies on a scoping review but only includes the FT50 journals, the journals in the Senior Scholars' Basket of Eight, and the journal Business & Information Systems Engineering. This selection of journals limits the papers considered for review to a manageable number. Yet, it also bears the risk that some open questions have already been answered in studies beyond the presented set of interest. In particular, more novel topics might have been tackled in conference papers that were not included in the review.

The first empirical study relies on a panel data study design that compares the number of followers before vs. after the purge. However, this design comes with three disadvantages. First, it cannot be observed whether the trend from the pre-period can be assumed for the post-period without the intervention, and it is more challenging to draw causal conclusions when a control group is absent. The study tries to overcome this challenge by relying on a historical control group to measure the extent of fake follower use and builds control groups from unaffected firms to assess shareholder reactions; however, this cannot entirely rule out confounding events. However, further robustness checks and investigations of press releases and public news make it highly likely to assume that the purge of followers drives the effect. Second, the purge of fake accounts is carried out by Twitter itself, and it cannot be ensured that the platform reliably detects and removes all fake accounts. If not all fake followers were removed, this would lead to an estimate that is too small in magnitude. However, resulting from a substantial database and outside pressure, Twitter likely acted in its best interest. Third, it is unclear whether the fake followers are purchased by firms or whether they have a different origin, as the research design is unable to provide insights about the source of the accounts, raising questions in terms of interpretation of the estimates. Despite this limitation, the estimates from the study provide robust evidence of the overall existence of fake followers and their risks, no matter who is responsible for them. In addition, the theory provides plausible reasons for why firms would be likely to use fake followers, while such incentives are less clear for other sources. Overall, the study advances the understanding of fake followers, which has previously been unexplored; however, a more diversified angle can help investigate this area of fakery even further, mainly because the effects only present a lower bound.

The second empirical study uses the implementation of the YouTube Authoritative Health Information program to assess the effect of promoting credible content. There are four major limitations in the design of the study. First, due to the delay between the announcement of the program and the start of the data collection after setting up the related code, it cannot be observed whether there has already been some change in channel behavior in response to the announcement. However, there is substantial confidence that the changes in demand only manifest after the promotion of content comes into place. Users would not know prior whether channels would be promoted and, in particular, would not yet see the search feature and the video label. In line, as channels only gain visibility once YouTube promotes them, changes in supply are likely to manifest only after the actual introduction. Second, in line with the computational resources, the data collection ended after 16 weeks, and the long-term effects could not be shown. However, changes can likely be observed within that time frame as it is sufficiently long to observe changes and adapt. Third, the study only relies on channel data but has no more profound insights into users and videos. Based on the data, it is impossible to assess users' behavior apart from their behavior toward health channels, i.e., whether they move to alternative platforms or whether there are adverse spillover effects for other topics. Also, the data does not allow for tracking the qualitative specificities of the channels' videos, which would enable evaluating changes in content due to the intervention. This would ultimately allow to better understand the quality of content, thus complementing the analysis of channels' subscribers. Fourth, YouTube relies on personalized algorithms according to which users potentially do not see the platform in the same way. As a result, the appearance of the search feature can differ for different individuals. However, differentiating only between authoritative and non-authoritative channels, using channel fixed effects, and balancing the results over extensive data gives some confidence that this would not be a major problem for the findings of the study. Against this backdrop, future studies should further explore design decisions as this seems promising for altering users' behavior.

6. Conclusion

This dissertation adds to the broader question of how fakery emerges and spreads on digital platforms and what related consequences and countermeasures are. This dissertation is three-fold: First, it provides a scoping review of the literature on fakery on digital platforms along the sender-receiver framework, presenting the foundation for the empirical studies. The review indicates rising interest in platform fakery, with some open questions remaining for future research. Second, with fake followers being identified as a public concern and a major gap in the literature, the first empirical study conducts a panel data study on the use and implications of fake followers with the 2018 Great Purge on Twitter as an intervention. During this purge, Twitter removed tens of millions of accounts from their platform if they showed suspicious behavior. The study shows that fake follower use is smaller than assumed, with minor implications for investors. Third, the second empirical study complements the insights on countermeasures against misinformation and conducts a difference-in-difference-in-differences estimation with the introduction of the YouTube Authoritative Health Information program in Germany in 2023 as the treatment. The platform features channels on top of the search results and provides them with a label on the video level. The study shows that the label is ineffective in altering views but that the search feature effectively shifts views from non-authoritative to authoritative channels. However, the effects on content production are negligible. These studies provide insights for platform providers, users, and policymakers.

References

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., and Nunamaker Jr, J. F. 2010. “Detecting Fake Websites: The Contribution of Statistical Learning Theory,” MIS Quarterly (34:3), pp. 435–461.

Abdulqader, M., Namoun, A., and Alsaawy, Y. 2022. “Fake Online Reviews: A Unified Detection Model Using Deception Theories,” IEEE Access (10), pp. 128622–128655.

Ackoff, R. L. 1989. “From Data to Wisdom,” Journal of Applied Systems Analysis (16), pp. 3–9.

Adomavicius, G., Bockstedt, J. C., Curley, S. P., and Zhang, J. 2013. “Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects,” Information Systems Research (24:4), pp. 956–975.

Aguiar, L., Claussen, J., and Peukert, C. 2018. “Catch Me If You Can: Effectiveness and Consequences of Online Copyright Enforcement,” Information Systems Research (29:3), pp. 656–678.

Ahmad, W., Sen, A., Eesley, C., and Brynjolfsson, E. 2024. “The Role of Advertisers and Platforms in Monetizing Misinformation: Descriptive and Experimental Evidence,” NBER Working Paper.

Akar, E., Hakyemez, T. C., Bozanta, A., and Akar, S. 2021. “What Sells on the Fake News Market? Examining the Impact of Contextualized Rhetorical Features on the Popularity of Fake Tweets,” Online Journal of Communication and Media Technologies (12:1), e202201.

Algarni, A., Xu, Y., and Chan, T. 2017. “An Empirical Study on the Susceptibility to Social Engineering in Social Networking Sites: The Case of Facebook,” European Journal of Information Systems (26:6), pp. 661–687.

Allcott, H., and Gentzkow, M. 2017. “Social Media and Fake News in the 2016 Election,” Journal of Economic Perspectives (31:2), pp. 211–236.

Allcott, H., Gentzkow, M., and Yu, C. 2019. “Trends in the Diffusion of Misinformation on Social Media,” Research & Politics (6:2).

Ananthakrishnan, U. M., Li, B., and Smith, M. D. 2020. “A Tangled Web: Should Online Review Portals Display Fraudulent Reviews?,” Information Systems Research (31:3), pp. 950–971.

Anderson, E. T., and Simester, D. I. 2014. “Reviews Without a Purchase: Low Ratings, Loyal Customers, and Deception,” Journal of Marketing Research (51:3), pp. 249–269.

Angrist, J. D., and Pischke, J.-S. 2008. Mostly Harmless Econometrics: An Empiricist’s Companion, Princeton, NJ: Princeton University Press.

Appan, R., and Browne, G. J. 2012. “The Impact of Analyst-Induced Misinformation on the Requirements Elicitation Process,” MIS Quarterly (36:1), pp. 85–106.

ARD, and ZDF 2019. Verteilung der Nutzer von Videoportalen, wie z.B. YouTube, nach Endgeräten in Deutschland im Jahr 2019. Statista (October 10). Retrieved Janaury 10, 2024, from https://de.statista.com/statistik/daten/studie/1073657/umfrage/verteilung-der-nutzer-von-videoportalen-nach-endgeraeten-in-deutschland/.

Armstrong, G. M., Gurol, M. N., and Russ, F. A. 1979. “Detecting and Correcting Deceptive Advertising,” Journal of Consumer Research (6:3), pp. 237–246.

Asimovic, N., Nagler, J., Bonneau, R., and Tucker, J. A. 2021. “Testing the Effects of Facebook Usage in an Ethnically Polarized Setting,” Proceedings of the National Academy of Sciences (118:25), e2022819118.

Association for Information Systems 2011. Senior Scholars’ Basket of Journals. Retrieved January 30, 2023, from https://aisnet.org/page/SeniorScholarBasket.

Attas, D. 1999. “What’s Wrong with “Deceptive” Advertising?,” Journal of Business Ethics (21:1), pp. 49–59.

Awan, T. M., Aziz, M., Sharif, A., Ch, T. R., Jasam, T., and Alvi, Y. 2022. “Fake News During the Pandemic Times: A Systematic Literature Review using PRISMA,” Open Information Science (6:1), pp. 49–60.

Badrinathan, S. 2021. “Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India,” American Political Science Review (115:4), pp. 1325–1341.

Baptista, J. P., and Gradim, A. 2020. “Understanding Fake News Consumption: A Review,” Social Sciences (9:10), 185.

Barney, J. 1991. “Firm Resources and Sustained Competitive Advantage,” Journal of Management (17:1), pp. 99–120.

Barone, M. J., and Miniard, P. W. 1999. “How and When Factual Ad Claims Mislead Consumers: Examining the Deceptive Consequences of Copy × Copy Interactions for Partial Comparative Advertisements,” Journal of Marketing Research (36:1), pp. 58–74.

Barrera, O., Guriev, S., Henry, E., and Zhuravskaya, E. 2020. “Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics,” Journal of Public Economics (182), 104123.

Becker, G. S., and Lewis, H. G. 1973. “On the Interaction Between the Quantity and Quality of Children,” Journal of Political Economy (81:2, Part 2), pp. 279–288.

Beisecker, S., Schlereth, C., and Hein, S. 2024. “Shades of Fake News: How Fallacies Influence Consumers’ Perception,” European Journal of Information Systems (33:1), pp. 41–60.

Bello Rinaudo, N., Matook, S., and Dennis, A. R. 2022. “Social Media’s Stockholm Syndrome: A Literature Review of User’s Love and Hate,” in Proceedings of the 33rd Australasian Conference on Information Systems, Melbourne, Australia.

Benbasat, I., and Wang, W. 2005. “Trust in and Adoption of Online Recommendation Agents,” Journal of the Association for Information Systems (6:3), 4.

Benjamin, V., and Raghu, T. 2023. “Augmenting Social Bot Detection with Crowd-Generated Labels,” Information Systems Research (34:2), pp. 487–507.

Berger, L. M., Kerkhof, A., Mindl, F., and Münster, J. 2023. “Debunking “Fake News” on Social Media: Short- and Long-Term Effects of Fact Checking and Media Literacy Interventions,” CESifo Working Paper.

Bessi, A., and Ferrara, E. 2016. “Social Bots Distort the 2016 US Presidential Election Online Discussion,” First Monday (21:11-7).

Bharadwaj, A. S., Bharadwaj, S. G., and Konsynski, B. R. 1999. “Information Technology Effects on Firm Performance as Measured by Tobin’s q,” Management Science (45:7), pp. 1008–1024.

Bhargava, H. K. 2022. “The Creator Economy: Managing Ecosystem Supply, Revenue Sharing, and Platform Design,” Management Science (68:7), pp. 5233–5251.

Blackwell, M., Iacus, S., King, G., and Porro, G. 2009. “cem: Coarsened Exact Matching in Stata,” The Stata Journal (9:4), pp. 524–546.

Bloomberg 2015. Bloomberg and Twitter Sign Data Licensing Agreement. September 16. Retrieved October 25, 2022, from https://www.bloomberg.com/company/press/bloomberg-and-twitter-sign-data-licensing-agreement/.

Bockstedt, J., and Goh, K. H. 2011. “Seller Strategies for Differentiation in Highly Competitive Online Auction Markets,” Journal of Management Information Systems (28:3), pp. 235–268.

Borwankar, S., Zheng, J., and Kannan, K. N. 2022. “Democratization of Misinformation Monitoring: The Impact of Twitter’s Birdwatch Program,” Working Paper.

Boudreau, K. 2010. “Open Platform Strategies and Innovation: Granting Access vs. Devolving Control,” Management Science (56:10), pp. 1849–1872.

Boudreau, K. J., and Hagiu, A. 2009. “Platform Rules: Multi-Sided Platforms as Regulators,” in Platforms, Markets and Innovation, Gawer, A. (ed.), Edward Elgar Publishing, pp. 163–191.

Bradshaw, S., Elswah, M., and Perini, A. 2021. “Look Who’s Watching: Platform Labels and User Engagement on State-Backed Media Outlets,” American Behavioral Scientist (Forthcoming).

BrightLocal 2024. Share of Consumers Confident They’ve Seen Fake Reviews on Amazon in the U.S. from 2022 to 2024. Statista (March 6). Retrieved June 24, 2024, from https://www.statista.com/statistics/997026/amazon-shopping-categories-largest-share-fake-product-reviews/.

Brown, S. J., and Warner, J. B. 1985. “Using Daily Stock Returns: The Case of Event Studies,” Journal of Financial Economics (14:1), pp. 3–31.

Browne, R. 2018. Jack Dorsey Loses 200,000 Followers on Twitter After Fake User Purge. CNBC (July 13). Retrieved February 10, 2022, from https://www.cnbc.com/2018/07/13/jack-dorsey-loses-200000-followers-on-twitter-after-fake-user-purge.html.

Burstin, H., Curry, S., Ranney, M. L., Arora, V., Wachler, B. B., Chou, W.-Y. S., Correa, R., Cryer, D., Dizon, D., Flores, E., Harmon, G., Jain, A., Johnson, K., Laine, C., Leininger, L., McMahon, G., Michaelis, L., Minhas, R., Mularski, R., Oldham, J., Padman, R., Pinnock, C., Rivera, J., Southwell, B., Villarruel, A., and Wallace, K. 2023. “Identifying Credible Sources of Health Information in Social Media: Phase 2—Considerations for Non-Accredited Nonprofit Organizations, For-Profit Entities, and Individual Sources,” NAM Perspectives (PMC10617996).

Burtch, G., He, Q., Hong, Y., and Lee, D. 2022. “How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit,” Management Science (68:5), pp. 3488–3506.

Burtch, G., Hong, Y., Bapna, R., and Griskevicius, V. 2018. “Stimulating Online Reviews by Combining Financial Incentives and Social Norms,” Management Science (64:5), pp. 2065–2082.

Calderon, E. D. V., James, T. L., and Lowry, P. B. 2023. “How Facebook’s Newsfeed Algorithm Shapes Childhood Vaccine Hesitancy: An Algorithmic Fairness, Accountability, and Transparency (FAT) Perspective,” Data and Information Management (7:3), 100042.

Calzada, J., and Gil, R. 2020. “What Do News Aggregators Do? Evidence from Google News in Spain and Germany,” Marketing Science (39:1), pp. 134–167.

Campbell, D. W., and Shang, R. 2022. “Tone at the Bottom: Measuring Corporate Misconduct Risk from the Text of Employee Reviews,” Management Science (68:9), pp. 7034–7053.

Candogan, O., and Drakopoulos, K. 2020. “Optimal Signaling of Content Accuracy: Engagement vs. Misinformation,” Operations Research (68:2), pp. 497–515.

Cantarella, M., Fraccaroli, N., and Volpe, R. 2023. “Does Fake News Affect Voting Behaviour?,” Research Policy (52:1), 104628.

Cao, Z., Zhu, Y., Li, G., and Qiu, L. 2023. “Consequences of Information Feed Integration on User Engagement and Contribution: A Natural Experiment in an Online Knowledge-Sharing Community,” Information Systems Research (Forthcoming).

Carhart, M. M. 1997. “On Persistence in Mutual Fund Performance,” The Journal of Finance (52:1), pp. 57–82.

Carson, D. J. 1985. “The Evolution of Marketing in Small Firms,” European Journal of Marketing (19:5), pp. 7–16.

Carson, T. L., Wokutch, R. E., and Cox, J. E. 1985. “An Ethical Analysis of Deception in Advertising,” Journal of Business Ethics (4:2), pp. 93–104.

Caruccio, L., Desiato, D., and Polese, G. 2018. “Fake Account Identification in Social Networks,” in Proceedings of the 6th IEEE International Conference on Big Data, Seattle, WA, pp. 5078–5085.

Caruso, D. L. 2023. Santé: Comment YouTube Accélère le Grand Ménage dans les Vidéos de Désinformation Médicale. LeParisien (September 7). Retrieved October 12, 2023, from https://www.leparisien.fr/high-tech/sante-comment-youtube-accelere-le-grand-menage-dans-les-videos-de-desinformation-medicale-07-09-2023-IAZQKHPUBRBDXNE5FBRKUJB6E4.php.

Cavusoglu, H., Li, Z., and Kim, S. H. 2021. “How Do Virtual Badges Incentivize Voluntary Contributions to Online Communities?,” Information & Management (58:5), 103483.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. 2010. “Measuring User Influence in Twitter: The Million Follower Fallacy,” in Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, Washington, DC.

Chatain, O. 2022. “The Business Model of Content-Sharing Platforms and the Supply of Content Moderation: Implications for Combating Information Manipulations,” Report 99, IRSEM.

Chaxel, A.-S. 2022. “How Misinformation Taints our Belief System: A Focus on Belief Updating and Relational Reasoning,” Journal of Consumer Psychology (32:2), pp.

370–373.

Chen, C., Bai, Y., and Wang, R. 2019a. “Online Political Efficacy and Political Participation: A Mediation Analysis Based on the Evidence from Taiwan,” New Media & Society (21:8), pp. 1667–1696.

Chen, C.-Y., Kearney, M., and Chang, S.-L. 2021. “Belief in or Identification of False News According to the Elaboration Likelihood Model,” International Journal of Communication (15), pp. 1263–1285.

Chen, H., De, P., and Hu, Y. J. 2015a. “IT-Enabled Broadcasting in Social Media: An Empirical Study of Artists’ Activities and Music Sales,” Information Systems Research (26:3), pp. 513–531.

Chen, H., Hu, Y. J., and Huang, S. 2019b. “Monetary Incentive and Stock Opinions on Social Media,” Journal of Management Information Systems (36:2), pp. 391–417.

Chen, J., Guo, Z., and Huang, J. 2022. “An Economic Analysis of Rebates Conditional on Positive Reviews,” Information Systems Research (33:1), pp. 224–243.

Chen, K., Wang, M., Huang, C., Kinney, P. L., and Anastas, P. T. 2020. “Air Pollution Reduction and Mortality Benefit During the COVID-19 Outbreak in China,” The Lancet Planetary Health (4:6), pp. 210–212.

Chen, L., and Papanastasiou, Y. 2021. “Seeding the Herd: Pricing and Welfare Effects of Social Learning Manipulation,” Management Science (67:11), pp. 6734–6750.

Chen, M., Jacob, V. S., Radhakrishnan, S., and Ryu, Y. U. 2015b. “Can Payment-Per-Click Induce Improvements in Click Fraud Identification Technologies?,” Information Systems Research (26:4), pp. 754–772.

Chen, S., Xiao, L., and Kumar, A. 2023. “Spread of Misinformation on Social Media: What Contributes to It and How to Combat It,” Computers in Human Behavior (141), 107643.

Chen, W., Wei, X., and Zhu, K. 2018. “Engaging Voluntary Contributions in Online Communities: A Hidden Markov Model,” MIS Quarterly (42:1), pp. 83–100.

Cheung, C. M.-Y., Sia, C.-L., and Kuan, K. K. 2012. “Is this Review Believable? A Study of Factors Affecting the Credibility of Online Consumer Reviews from an ELM Perspective,” Journal of the Association for Information Systems (13:8), pp. 618–635.

Chevalier, J. A., and Mayzlin, D. 2006. “The Effect of Word of Mouth on Sales: Online Book Reviews,” Journal of Marketing Research (43:3), pp. 345–354.

Cho, C. H., Martens, M. L., Kim, H., and Rodrigue, M. 2011. “Astroturfing Global Warming: It Isn’t Always Greener on the Other Side of the Fence,” Journal of Business Ethics (104:4), pp. 571–587.

Chua, C. E. H., Wareham, J., and Robey, D. 2007. “The Role of Online Trading Communities in Managing Internet Auction Fraud,” MIS Quarterly (31:4), pp. 759–781.

Chung, S., Animesh, A., Han, K., and Pinsonneault, A. 2020. “Financial Returns to Firms’ Communication Actions on Firm-Initiated Social Media: Evidence from Facebook Business Pages,” Information Systems Research (31:1), pp. 258–285.

Clarke, J., Chen, H., Du, D., and Hu, Y. J. 2021. “Fake News, Investor Attention, and Market Reaction,” Information Systems Research (32:1), pp. 35–52.

Claussen, J., Kretschmer, T., and Mayrhofer, P. 2013. “The Effects of Rewarding User Engagement: The Case of Facebook Apps,” Information Systems Research (24:1), pp. 186–200.

Cohn, A., Gesche, T., and Maréchal, M. A. 2022. “Honesty in the Digital Age,” Management Science (68:2), pp. 827–845.

Colomina, C., Margalef, H. S., and Youngs, R. 2021. “The Impact of Disinformation on Democratic Processes and Human Rights in the World,” Study Requested by the European Parliament’s Subcommittee on Human Rights.

Confessore, N., and Dance, G. J. 2018. Battling Fake Accounts, Twitter to Slash Millions of Followers. The New York Times (July 11). Retrieved September 1, 2021, from https://www.nytimes.com/2018/07/11/technology/twitter-fake-followers.html.

Confessore, N., Dance, G. J. X., Harris, R., and Hansen, M. 2018. The Follower Factory. The New York Times (January 27). Retrieved July 18, 2022, from https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html.

Connelly, B. L., Certo, S. T., Ireland, R. D., and Reutzel, C. R. 2011. “Signaling Theory: A Review and Assessment,” Journal of Management (37:1), pp. 39–67.

Cowley, E., and Janus, E. 2004. “Not Necessarily Better, But Certainly Different: A Limit to the Advertising Misinformation Effect on Memory,” Journal of Consumer Research (31:1), pp. 229–235.

Craig, A. W., Loureiro, Y. K., Wood, S., and Vendemia, J. M. 2012. “Suspicious Minds: Exploring Neural Processes During Exposure to Deceptive Advertising,” Journal of Marketing Research (49:3), pp. 361–372.

Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., and Tesconi, M. 2015. “Fame for Sale: Efficient Detection of Fake Twitter Followers,” Decision Support Systems (80), pp. 56–71.

Darke, P. R., Ashworth, L., and Main, K. J. 2010. “Great Expectations and Broken Promises: Misleading Claims, Product Failure, Expectancy Disconfirmation and Consumer Distrust,” Journal of the Academy of Marketing Science (38:3), pp. 347–362.

Darke, P. R., and Ritchie, R. J. 2007. “The Defensive Consumer: Advertising Deception, Defensive Processing, and Distrust,” Journal of Marketing Research (44:1), pp. 114–127.

DataReportal, We Are Social, Google, and Meltwater 2022. Leading Countries Based on YouTube Audience Size as of April 2022. Statista (January 31). Retrieved December 9, 2022, from https://www.statista.com/statistics/280685/number-of-monthly-unique-youtube-users/.

Dave, P. 2018. Twitter Cuts Suspect Users From Follower Counts Again, Blames Bug. Reuters (November 10). Retrieved December 30, 2021, from https://www.reuters.com/article/us-twitter-followers-idUSKCN1NE2K3.

Davenport, T. H., and Beck, J. C. 2001. The Attention Economy: Understanding the New Currency of Business, Boston, MA: Harvard Business School Press.

Davidson III, W. N., and Worrel, D. L. 1988. “The Impact of Announcements of Corporate Illegalities on Shareholder Returns,” Academy of Management Journal (31:1), pp. 195–200.

De Beaumont 2021. Study: Americans Who Get COVID-19 Information From Social

Media More Likely To Believe Misinformation, Less Likely To Be Vaccinated. November 4. Retrieved June 24, 2024, from https://debeaumont.org/news/2021/social-media-misinformation-poll/.

De Oliveira Santini, F., Ladeira, W. J., Pinto, D. C., Herter, M. M., Sampaio, C. H., and Babin, B. J. 2020. “Customer Engagement in Social Media: A Framework and Meta-Analysis,” Journal of the Academy of Marketing Science (48:6), pp. 1211–1228.

De Vries, L., Gensler, S., and Leeflang, P. S. 2017. “Effects of Traditional Advertising and Social Messages on Brand-Building Metrics and Customer Acquisition,” Journal of Marketing (81:5), pp. 1–15.

Dehning, B., Richardson, V. J., and Zmud, R. W. 2003. “The Value Relevance of Announcements of Transformational Information Technology Investments,” MIS Quarterly (27:4), pp. 637–656.

Dellarocas, C. 2006a. “Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms,” Management Science (52:10), pp. 1577–1593.

Dellarocas, C. 2006b. “Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms,” Management Science (52:10), pp. 1577–1593.

Deng, B., and Chau, M. 2021. “The Effect of the Expressed Anger and Sadness on Online News Believability,” Journal of Management Information Systems (38:4), pp. 959–988.

Deng, S., Huang, Z. J., Sinha, A. P., and Zhao, H. 2018. “The Interaction Between Microblog Sentiment and Stock Return: An Empirical Examination,” MIS Quarterly (42:3), pp. 895–918.

Deng, X., Kahn, B. E., Unnava, H. R., and Lee, H. 2016. “A “Wide” Variety: Effects of Horizontal versus Vertical Display on Assortment Processing, Perceived Variety, and Choice,” Journal of Marketing Research (53:5), pp. 682–698.

Deng, Y., Zheng, J., Khern-am-nuai, W., and Kannan, K. N. 2022. “More than the Quantity: The value of Editorial Reviews for a UGC Platform,” Management Science (68:9), pp. 6865–6888.

Deng, Z., Liesch, P. W., and Wang, Z. 2021. “Deceptive Signaling on Globalized Digital Platforms: Institutional Hypnosis and Firm Internationalization,” Journal of International Business Studies (52:6), pp. 1096–1120.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. 2003. “Cues to Deception,” Psychological Bulletin (129:1), pp. 74–118.

Dewan, S., Kim, J., and Nian, T. 2023. “Economic Impacts of Platform-Endorsed Quality Certification: Evidence from Airbnb,” MIS Quarterly (47:3), pp. 1353–1368.

Dimoka, A., Hong, Y., and Pavlou, P. A. 2012. “On Product Uncertainty in Online Markets: Theory and Evidence,” MIS Quarterly (36:2), pp. 395–426.

Donaker, G., Kim, H., Luca, M., and Weber, M. 2019. “Designing Better Online Review Systems,” Harvard Business Review (97:6), pp. 122–129.

Duffy, C. 2022. More Than 80 Fact-Checking Organizations Call Out YouTube’s ‘Insufficient’ Response to Misinformation. CNN Business (January 12). Retrieved December 7, 2022, from https://edition.cnn.com/2022/01/12/tech/youtube-fact-checkers-letter/index.html.

Edelman, B. 2009. “How to Combat Online Ad Fraud,” Harvard Business Review (87:12), pp. 24–25.

El-Komboz, L. A., Kerkhof, A., and Loh, J. 2023. “Platform Partnership Programs and Content Supply: Evidence from the YouTube “Adpocalypse”,” CESifo Working Paper.

Elisabeth, N.-N. 1974. “The Spiral of Silence: A Theory of Public Opinion,” Journal of Communication (24:2), pp. 43–51.

Engert, M., Evers, J., Hein, A., and Krcmar, H. 2023. “Sustaining Complementor Engagement in Digital Platform Ecosystems: Antecedents, Behaviours and Engagement Trajectories,” Information Systems Journal (33:5), pp. 1151–1185.

Eurostat 2021. One in Two EU Citizens Look for Health Information Online. April 6. Retrieved December 9, 2022, from https://ec.europa.eu/eurostat/de/web/products-eurostat-news/-/edn-20210406-1.

Evans, D. S., and Schmalensee, R. 2010. “Failure to Launch: Critical Mass in Platform Businesses,” Review of Network Economics (9:4), 1.

Facebook 2024. Actioned Fake Accounts on Facebook Worldwide From 4th Quarter 2017 to 4th Quarter 2023. Statista (March 1). Retrieved June 24, 2024, from https://www.statista.com/statistics/1013474/facebook-fake-account-removal-quarter/.

Fama, E. F. 1970. “Efficient Capital Markets: A Review of Theory and Empirical Work,” The Journal of Finance (25:2), pp. 383–417.

Fama, E. F., Fisher, L., Jensen, M. C., and Roll, R. 1969. “The Adjustment of Stock Prices to New Information,” International Economic Review (10:1), pp. 1–21.

Federal Trade Commission 2023. Federal Trade Commission Announces Proposed Rule Banning Fake Reviews and Testimonials. June 30. Retrieved October 6, 2023, from https://www.ftc.gov/news-events/news/press-releases/2023/06/federal-trade-commission-announces-proposed-rule-banning-fake-reviews-testimonials.

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. 2016. “The Rise of Social Bots,” Communications of the ACM (59:7), pp. 96–104.

Foerderer, J. 2020. “Interfirm Exchange and Innovation in Platform Ecosystems: Evidence from Apple’s Worldwide Developers Conference,” Management Science (66:10), pp. 4772–4787.

Foerderer, J., Kude, T., Mithas, S., and Heinzl, A. 2018. “Does Platform Owner’s Entry Crowd Out Innovation? Evidence from Google Photos,” Information Systems Research (29:2), pp. 444–460.

Foerderer, J., Lueker, N., and Heinzl, A. 2021. “And The Winner Is…? The Desirable and Undesirable Effects of Platform Awards,” Information Systems Research (32:4), pp. 1155–1172.

Foerderer, J., and Schuetz, S. W. 2022. “Data Breach Announcements and Stock Market Reactions: A Matter of Timing?,” Management Science (68:10), pp. 7298–7322.

French, A. M., Storey, V. C., and Wallace, L. 2023. “The Impact of Cognitive Biases on the Believability of Fake News,” European Journal of Information Systems (Forthcoming).

Gaeth, G. J., and Heath, T. B. 1987. “The Cognitive Processing of Misleading Advertising in Young and Old Adults: Assessment and Training,” Journal of Consumer

Research (14:1), pp. 43–54.

Gallus, J. 2017. “Fostering Public Good Contributions With Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia,” Management Science (63:12), pp. 3999–4015.

Gallwitz, F., and Kreil, M. 2021. “The Rise and Fall of ‘Social Bot’ Research,” Working Paper.

Gandal, N., Kende, M., and Rob, R. 2000. “The Dynamics of Technological Adoption in Hardware/Software Systems: The Case of Compact Disc Players,” The RAND Journal of Economics (31:1), pp. 43–61.

Gardner, D. M. 1975. “Deception in Advertising: A Conceptual Approach,” Journal of Marketing (39:1), pp. 40–46.

Garg, R., Kim, J. H. J., and Lee, S. K. 2023. “The Price of Losing Trust: An Empirical Analysis of Social Misconduct by YouTube Creators,” Working Paper.

Gelper, S., Van der Lans, R., and Van Bruggen, G. 2021. “Competition for Attention in Online Social Networks: Implications for Seeding Strategies,” Management Science (67:2), pp. 1026–1047.

George, J., Gerhart, N., and Torres, R. 2021. “Uncovering the Truth about Fake News: A Research Model Grounded in Multi-Disciplinary Literature,” Journal of Management Information Systems (38:4), pp. 1067–1094.

George, J. F., Gupta, M., Giordano, G., Mills, A. M., Tennant, V. M., and Lewis, C. C. 2018. “The Effects of Communication Media and Culture on Deception Detection Accuracy,” MIS Quarterly (42:2), pp. 551–575.

George, J. F., Marett, K., and Tilley, P. A. 2008. “The Effects of Warnings, Computer-Based Media, and Probing Activity on Successful Lie Detection,” IEEE Transactions on Professional Communication (51:1), pp. 1–17.

George, J. F., and Robb, A. 2008. “Deception and Computer-Mediated Communication in Daily Life,” Communication Reports (21:2), pp. 92–103.

Gerlitz, C., and Rieder, B. 2013. “Mining One Percent of Twitter: Collections, Baselines, Sampling,” M/C Journal (16:2).

Germano, F., Gómez, V., and Sobbrio, F. 2022. “Crowding Out The Truth? A Simple Model of Misinformation, Polarization and Meaningful Social Interactions,” CESifo Working Paper 10011.

Ghose, A., Goldfarb, A., and Han, S. P. 2013. “How is the Mobile Internet Different? Search Costs and Local Activities,” Information Systems Research (24:3), pp. 613–631.

Gimpel, H., Heger, S., Olenberger, C., and Utz, L. 2021. “The Effectiveness of Social Norms in Fighting Fake News on Social Media,” Journal of Management Information Systems (38:1), pp. 196–221.

Goes, P. B., Lin, M., and Au Yeung, C.-m. 2014. ““Popularity Effect” in User-Generated Content: Evidence from Online Product Reviews,” Information Systems Research (25:2), pp. 222–238.

Goldhaber, M. H. 1997. “The Attention Economy and the Net,” First Monday (2:4).

Goodrow, C. 2021. On YouTube’s Recommendation System. YouTube Official Blog

(September 15). Retrieved April 2, 2024, from https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/.

Graham, B., and Zweig, J. 2003. The Intelligent Investor, New York, NY: Harper Business.

Graham, G. 2022a. Answering your Health Questions in Brazil, India, and Japan. YouTube Official Blog (March 24). Retrieved October 12, 2023, from https://blog.youtube/news-and-events/answering-your-health-questions-brazil-india-and-japan/.

Graham, G. 2022b. New Ways to Answer Your Health Questions in the United Kingdom. YouTube Official Blog (June 15). Retrieved October 12, 2023, from https://blog.youtube/news-and-events/new-ways-to-answer-your-health-questions-in-the-united-kingdom/.

Graham, G. 2022c. YouTube Health: Zugang zu Zuverlässigen Gesundheitsinformationen. YouTube Official Blog (October 18). Retrieved December 7, 2022, from https://blog.youtube/intl/de-de/news-and-events/youtube-health-zugang-zu-zuverlassigen-gesundheitsinformationen/.

Greve, H. R., Palmer, D., and Pozner, J.-E. 2010. "Organizations Gone Wild: The Causes, Processes, and Consequences of Organizational Misconduct," Academy of Management Annals (4:1), pp. 53–107.

Gutt, D., Neumann, J., Zimmermann, S., Kundisch, D., and Chen, J. 2019. "Design of Review Systems–A Strategic Instrument to Shape Online Reviewing Behavior and Economic Outcomes," The Journal of Strategic Information Systems (28:2), pp. 104–117.

Hagen, L., Neely, S., Keller, T. E., Scharf, R., and Vasquez, F. E. 2022. "Rise of the Machines? Examining the Influence of Social Bots on a Political Discussion Network," Social Science Computer Review (40:2), pp. 264–287.

Halckenhaeusser, A., Foerderer, J., and Heinzl, A. 2020. "Platform Governance Mechanisms: An Integrated Literature Review and Research Directions," in Proceedings of the 28th European Conference on Information Systems, Virtual.

Hamby, A., Ecker, U., and Brinberg, D. 2020. "How Stories in Memory Perpetuate the Continued Influence of False Information," Journal of Consumer Psychology (30:2), pp. 240–259.

Harrison, A. 2018. "The Effects of Media Capabilities on the Rationalization of Online Consumer Fraud," Journal of the Association for Information Systems (19:5), pp. 408–440.

Hawlitschek, F., Stofberg, N., Teubner, T., Tu, P., and Weinhardt, C. 2018. "How Corporate Sharewashing Practices Undermine Consumer Trust," Sustainability (10:8), 2638.

He, F., Du, H., and Yu, B. 2022a. "Corporate ESG Performance and Manager Misconduct: Evidence from China," International Review of Financial Analysis (82), 102201.

He, L., Luo, J., Tang, Y., Wu, Z., and Zhang, H. 2023. "Motivating User-Generated Content: Unintended Consequences of Incentive Thresholds," MIS Quarterly (47:3), pp. 1015–1044.

He, S., Hollenbeck, B., and Proserpio, D. 2022b. “The Market for Fake Reviews,” Marketing Science (41:5), pp. 896–921.

Heese, J., Pérez-Cavazos, G., and Peter, C. D. 2022. “When the Local Newspaper Leaves Town: The Effects of Local Newspaper Closures on Corporate Misconduct,” Journal of Financial Economics (145:2-Part-B), pp. 445–463.

Hennig-Thurau, T., Wiertz, C., and Feldhaus, F. 2015. “Does Twitter Matter? The Impact of Microblogging Word of Mouth on Consumers’ Adoption of New Movies,” Journal of the Academy of Marketing Science (43:3), pp. 375–394.

Hernon, P. 1995. “Disinformation and Misinformation Through the Internet: Findings of an Exploratory Study,” Government Information Quarterly (12:2), pp. 133–139.

Herzberg, A., and Jbara, A. 2008. “Security and Identification Indicators for Browsers Against Spoofing and Phishing Attacks,” ACM Transactions on Internet Technology (8:4), 16.

Hinz, O., Skiera, B., Barrot, C., and Becker, J. U. 2011. “Seeding Strategies for Viral Marketing: An Empirical Comparison,” Journal of Marketing (75:6), pp. 55–71.

Hirshleifer, D., and Teoh, S. 2009. “Thought and Behavioral Contagion in Capital Markets,” in Handbook of Financial Markets: Dynamics and Evolution, Hens, T., and Schenk-Hoppé, K. R. (eds.), Amsterdam, Netherlands: North Holland, pp. 1–56.

Ho, S. M., Hancock, J. T., Booth, C., and Liu, X. 2016. “Computer-Mediated Deception: Strategies Revealed by Language-Action Cues in Spontaneous Communication,” Journal of Management Information Systems (33:2), pp. 393–420.

Horner, C. G., Galletta, D., Crawford, J., and Shirsat, A. 2021. “Emotions: The Unexplored Fuel of Fake News on Social Media,” Journal of Management Information Systems (38:4), pp. 1039–1066.

Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., and Maziad, M. 2011. “Opening Closed Regimes: What was the Role of Social Media During the Arab Spring?,” Working Paper.

Hu, N., Bose, I., Koh, N. S., and Liu, L. 2012. “Manipulation of Online Reviews: An Analysis of Ratings, Readability, and Sentiments,” Decision Support Systems (52:3), pp. 674–684.

Huang, J. T., Kaul, R., and Narayanan, S. 2022a. “The Causal Effect of Attention and Recognition on the Nature of User-Generated Content: Experimental Results from an Image-Sharing Social Network,” Stanford University Graduate School of Business Research Paper.

Huang, P., Lyu, G., and Xu, Y. 2022b. “Quality Regulation on Two-Sided Platforms: Exclusion, Subsidization, and First-Party Applications,” Management Science (68:6), pp. 4415–4434.

Huang, P., Tafti, A. R., and Mithas, S. 2018. “Platform Sponsor’s Investments and User Contributions in Knowledge Communities: The Role of Knowledge Seeding,” MIS Quarterly (42:1), pp. 213–240.

Huang, Z., and Liu, D. 2023. “Economics of Social Media Fake Accounts,” Working Paper.

Hui, K.-L., Teo, H. H., and Lee, S.-Y. T. 2007. “The Value of Privacy Assurance: An

Exploratory Field Experiment," MIS Quarterly (31:1), pp. 19–33.

Hutto, C. J., and Gilbert, E. E. 2014. "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Proceedings of the 8th International AAAI Conference on Web and Social Media, Ann Arbor, MI, pp. 216–225.

Hwang, E. H., and Lee, S. 2024. "A Nudge to Credible Information as a Countermeasure to Misinformation: Evidence from Twitter," Information Systems Research (Forthcoming).

Iacus, S. M., King, G., and Porro, G. 2012. "Causal Inference Without Balance Checking: Coarsened Exact Matching," Political Analysis (20:1), pp. 1–24.

Im, K. S., Dow, K. E., and Grover, V. 2001. "A Reexamination of IT Investment and the Market Value of the Firm – An Event Study Methodology," Information Systems Research (12:1), pp. 103–117.

Jabr, W. 2022. "Review Credibility as a Safeguard Against Fakery: The Case of Amazon," European Journal of Information Systems (31:4), pp. 525–545.

Jabr, W., and Zheng, Z. 2014. "Know Yourself and Know Your Enemy: An Analysis of Firm Recommendations and Consumer Reviews in a Competitive Environment," MIS Quarterly (38:3), pp. 635–654.

Jacobs, J. 2018. In Twitter Purge, Top Accounts Lose Millions of Followers. The New York Times (July 12). Retrieved February 10, 2022, from https://www.nytimes.com/2018/07/12/technology/twitter-followers-nyt.html.

Jain, S., Jain, P., and Rezaee, Z. 2010. "Stock Market Reactions to Regulatory Investigations: Evidence from Options Backdating," Research in Accounting Regulation (22:1), pp. 52–57.

Jegadeesh, N., and Titman, S. 1993. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," The Journal of Finance (48:1), pp. 65–91.

Jensen, M. C., and Meckling, W. H. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," Journal of Financial Economics (3:4), pp. 305–360.

Jensen, M. L., Averbeck, J. M., Zhang, Z., and Wright, K. B. 2013. "Credibility of Anonymous Online Product Reviews: A Language Expectancy Perspective," Journal of Management Information Systems (30:1), pp. 293–324.

Jia, W., Redigolo, G., Shu, S., and Zhao, J. 2020. "Can Social Media Distort Price Discovery? Evidence from Merger Rumors," Journal of Accounting and Economics (70:1), 101334.

Jin, C., Yang, L., and Hosanagar, K. 2023. "To Brush or Not to Brush: Product Rankings, Consumer Search, and Fake Orders," Information Systems Research (34:2), pp. 532–552.

Johar, G. V. 1995. "Consumer Involvement and Deception from Implied Advertising Claims," Journal of Marketing Research (32:3), pp. 267–279.

Johar, G. V. 1996. "Intended and Unintended Effects of Corrective Advertising on Beliefs and Evaluations: An Exploratory Analysis," Journal of Consumer Psychology (5:3), pp. 209–230.

Johar, G. V. 2022. "Untangling the Web of Misinformation and False Beliefs," Journal of Consumer Psychology (32:2), pp. 374–383.

Johar, G. V., and Roggeveen, A. L. 2007. “Changing False Beliefs from Repeated Advertising: The Role of Claim-Refutation Alignment,” Journal of Consumer Psychology (17:2), pp. 118–127.

Kaplan, A. M., and Haenlein, M. 2012. “The Britney Spears Universe: Social Media and Viral Marketing at Its Best,” Business Horizons (55:1), pp. 27–31.

Kartal, M., and Tyran, J.-R. 2022. “Fake News, Voter Overconfidence, and the Quality of Democratic Choice,” American Economic Review (112:10), pp. 3367–3397.

Katz, M. L., and Shapiro, C. 1985. “Network Externalities, Competition, and Compatibility,” The American Economic Review (75:3), pp. 424–440.

Keppo, J., Kim, M. J., and Zhang, X. 2022. “Learning Manipulation Through Information Dissemination,” Operations Research (70:6), pp. 3490–3510.

Kerkhof, A. 2024. “Advertising and Content Differentiation: Evidence from YouTube,” .

Khan, A., Brohman, K., and Addas, S. 2022a. “The Anatomy of ‘Fake News’: Studying False Messages as Digital Objects,” Journal of Information Technology (37:2), pp. 122–143.

Khan, M. B., Goel, S., Katar Anandan, J., Zhao, J., and Naik, R. R. 2022b. “Deepfake Audio Detection,” in Proceedings of the 28th American Conference on Information Systems, Minneapolis, MN.

Khan, M. L. 2017. “Social Media Engagement: What Motivates User Participation and Consumption on YouTube?,” Computers in Human Behavior (66), pp. 236–247.

Khern-am nuai, W., Kannan, K., and Ghasemkhani, H. 2018. “Extrinsic versus Intrinsic Rewards for Contributing Reviews in an Online Platform,” Information Systems Research (29:4), pp. 871–892.

Kim, A., and Dennis, A. R. 2019. “Says Who? The Effects of Presentation Format and Source Rating on Fake News in Social Media,” MIS Quarterly (43:3), pp. 1025–1039.

Kim, A., Moravec, P. L., and Dennis, A. R. 2019. “Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings,” Journal of Management Information Systems (36:3), pp. 931–968.

Kim, E.-H., and Youm, Y. N. 2017. “How Do Social Media Affect Analyst Stock Recommendations? Evidence from S&P 500 Electric Power Companies’ Twitter Accounts,” Strategic Management Journal (38:13), pp. 2599–2622.

King, D., and Auschaitrakul, S. 2020. “Symbolic Sequence Effects on Consumers’ Judgments of Truth for Brand Claims,” Journal of Consumer Psychology (30:2), pp. 304–313.

King, K. K., Wang, B., Escobari, D., and Oraby, T. 2021. “Dynamic Effects of Falsehoods and Corrections on Social Media: A Theoretical Modeling and Empirical Evidence,” Journal of Management Information Systems (38:4), pp. 989–1010.

Kington, R., Arnesen, S., Chou, W.-Y. S., Curry, S., Lazer, D., and Villarruel, A. 2021. “Identifying Credible Sources of Health Information in Social Media: Principles and Attributes,” NAM Perspectives (PMC8486420).

Kircher, T., and Foerderer, J. 2024. “Ban Targeted Advertising? An Empirical Investigation of the Consequences for App Development,” Management Science (70:2), pp. 1070–1092.

Kirmani, A., and Zhu, R. 2007. “Vigilant Against Manipulation: The Effect of Regulatory Focus on the Use of Persuasion Knowledge,” Journal of Marketing Research (44:4), pp. 688–701.

Kitchenham, B., and Charters, S. 2007. “Guidelines for Performing Systematic Literature Reviews in Software Engineering,” EBSE Technical Report (EBSE-2007-01).

Knittel, C. R., and Stango, V. 2014. “Celebrity Endorsements, Firm Value, and Reputation Risk: Evidence from the Tiger Woods Scandal,” Management Science (60:1), pp. 21–37.

Kodura, M. 2023. YouTube ist bevorzugte Infoquelle für Health-Produkte. Healthcare Marketing (June 9). Retrieved January 10, 2024, from https://www.healthcaremarketing.eu/_rubric/detail.php?rubric=Mediennr=9591495914.

Kogan, S., Moskowitz, T. J., and Niessner, M. 2023. “Social Media and Financial News Manipulation,” Review of Finance (27:4), pp. 1229–1268.

Kohli, R., Devaraj, S., and Ow, T. T. 2012. “Does Information Technology Investment Influence a Firm’s Market Value? A Case of Non-Publicly Traded Healthcare Firms,” MIS Quarterly (36:4), pp. 1145–1163.

Kokkodis, M., Lappas, T., and Kane, G. C. 2022. “Optional Purchase Verification in E-Commerce Platforms: More Representative Product Ratings and Higher Quality Reviews,” Production and Operations Management (31:7), pp. 2943–2961.

Konchitchki, Y., and O’Leary, D. E. 2011. “Event Study Methodologies in Information Systems Research,” International Journal of Accounting Information Systems (12:2), pp. 99–115.

Koohikamali, M., and Sidorova, A. 2017. “Information Re-Sharing on Social Network Sites in the Age of Fake News,” Informing Science: The International Journal of an Emerging Transdiscipline (20), pp. 215–235.

Kovács, B., and Sharkey, A. J. 2014. “The Paradox of Publicity: How Awards Can Negatively Affect the Evaluation of Quality,” Administrative Science Quarterly (59:1), pp. 1–33.

Kreienbrink, M. 2023. Warum Liebscher & Bracht mehr als hundert ihrer Videos gelöscht haben. Spiegel (March 4). Retrieved March 14, 2024, from https://www.spiegel.de/netzwelt/web/liebscher-und-bracht-darum-haben-die-youtube-stars-mehr-als-hundert-ihrer-videos-geloescht-a-5f0df8b8-d71f-4a56-abdd-33fb44263470.

Kronrod, A., Gordeliy, I., and Lee, J. K. 2023. “Been There, Done That: How Episodic and Semantic Memory Affects the Language of Authentic and Fictitious Reviews,” Journal of Consumer Research (50:2), pp. 405–425.

Kruger, J., and Dunning, D. 1999. “Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments,” Journal of Personality and Social Psychology (77:6), pp. 1121–1134.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., and Kannan, P. 2016. “From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior,” Journal of Marketing (80:1), pp. 7–25.

Kumar, N., Venugopal, D., Qiu, L., and Kumar, S. 2018. “Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning," Journal of Management Information Systems (35:1), pp. 350–380.

Kumar, N., Venugopal, D., Qiu, L., and Kumar, S. 2019. "Detecting Anomalous Online Reviewers: An Unsupervised Approach Using Mixture Models," Journal of Management Information Systems (36:4), pp. 1313–1346.

Kwieciński, D. 2017. "Measures of Competitive Intensity – Analysis Based on Literature Review," Central European Management Journal (25:1), pp. 53–77.

Laato, S., Islam, A. N., Islam, M. N., and Whelan, E. 2020. "What Drives Unverified Information Sharing and Cyberchondria During the COVID-19 Pandemic?," European Journal of Information Systems (29:3), pp. 288–305.

Lacka, E., Boyd, D. E., Ibikunle, G., and Kannan, P. 2022. "Measuring the Real-Time Stock Market Impact of Firm-Generated Content," Journal of Marketing (86:5), pp. 58–78.

Lamy, E. 2023. "Epistemic Responsibility in Business: An Integrative Framework for an Epistemic Ethics," Journal of Business Ethics (183:1), pp. 1–14.

Lappas, T., Sabnis, G., and Valkanas, G. 2016. "The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry," Information Systems Research (27:4), pp. 940–961.

Law, S., Hawkins, S. A., and Craik, F. I. 1998. "Repetition-Induced Belief in the Elderly: Rehabilitating Age-Related Memory Deficits," Journal of Consumer Research (25:2), pp. 91–107.

Lazer, D. M., 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., and Zittrain, J. L. 2018. "The Science of Fake News," Science (359:6380), pp. 1094–1096.

Lee, D., Hosanagar, K., and Nair, H. S. 2018a. "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook," Management Science (64:11), pp. 5105–5131.

Lee, J. M., Hwang, B.-H., and Chen, H. 2017. "Are Founder CEOs More Overconfident than Professional CEOs? Evidence from S&P 1500 Companies," Strategic Management Journal (38:3), pp. 751–769.

Lee, S.-Y., Qiu, L., and Whinston, A. 2018b. "Sentiment Manipulation in Online Platforms: An Analysis of Movie Tweets," Production and Operations Management (27:3), pp. 393–416.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., and Cook, J. 2012. "Misinformation and its Correction: Continued Influence and Successful Debiasing," Psychological Science in the Public Interest (13:3), pp. 106–131.

Lewandowsky, S., and Van der Linden, S. 2021. "Countering Misinformation and Fake News Through Inoculation and Prebunking," European Review of Social Psychology (32:2), pp. 348–384.

Li, H. O.-Y., Bailey, A., Huynh, D., and Chan, J. 2020. "YouTube as a Source of Information on COVID-19: A Pandemic of Misinformation?," BMJ Global Health (5:5), e002604.

Li, Y., and Xie, Y. 2020. “Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement,” Journal of Marketing Research (57:1), pp. 1–19.

Li, Y.-J., Marga, J. J., Cheung, C. M., Shen, X.-L., and Lee, M. 2022. “Health Misinformation on Social Media: A Systematic Literature Review and Future Research Directions,” AIS Transactions on Human-Computer Interaction (14:2), pp. 116–149.

Lin, Y.-K., Rai, A., and Yang, Y. 2022. “Information Control for Creator Brand Management in Subscription-Based Crowdfunding,” Information Systems Research (33:3), pp. 846–866.

Lindenmayr, M., and Foerderer, J. 2022. “Qualitätssicherung in Digitalen Plattform-Ökosystemen: Implementierung von Kontrollsystemen am Beispiel von Apple iOS,” HMD Praxis der Wirtschaftsinformatik (59:5), pp. 1312–1322.

Lindenmayr, M., and Foerderer, J. 2024. “Digitale B2B-Plattformökosysteme für Produzierende Unternehmen,” in Digitale Plattformen und Ökosysteme im B2B-Bereich, Schallmo, D. R. A., Kundisch, D., Lang, K., and Hasler, D. (eds.), Springer Gabler, pp. 161–182.

Lindenmayr, M., Kircher, T., Stolte, A., and Foerderer, J. 2022. “The Economic and Social Consequences of Digital Platforms: A Systematic and Interdisciplinary Literature Review,” in Digitalization Across Organizational Levels: New Frontiers for Information Systems Research, Dibbern, J., Foerderer, J., Kude, T., Rothlauf, F., and Spohrer, K. (eds.), Cham, Switzerland: Springer, pp. 147–178.

LinkedIn 2024. Number of Fake LinkedIn Accounts Detected and Removed Worldwide from 1st Half of 2019 to 2nd Half 2023. Statista (April 1). Retrieved June 24, 2024, from https://www.statista.com/statistics/1328849/linkedin-number-of-fake-accounts-detected-and-removed/.

Lis, B., and Neßler, C. 2014. “Electronic Word of Mouth,” Business & Information Systems Engineering (6:1), pp. 63–65.

Liu, B. 2012. “Sentiment Analysis and Opinion Mining,” Synthesis Lectures on Human Language Technologies (5:1), pp. 1–167.

Liu, X., Zhang, B., Susarla, A., and Padman, R. 2020. “Go to YouTube and Call Me in the Morning: Use of Social Media for Chronic Conditions,” MIS Quarterly (44:1b), pp. 257–283.

Liu, Y., and Feng, J. 2021. “Does Money Talk? The Impact of Monetary Incentives on User-Generated Content Contributions,” Information Systems Research (32:2), pp. 394–409.

Liu, Y., Shankar, V., and Yun, W. 2017. “Crisis Management Strategies and the Long-Term Effects of Product Recalls on Firm Value,” Journal of Marketing (81:5), pp. 30–48.

Loibl, C., and Hira, T. K. 2009. “Investor Information Search,” Journal of Economic Psychology (30:1), pp. 24–41.

London Jr, J., Li, S., and Sun, H. 2022. “Seems Legit: An Investigation of the Assessing and Sharing of Unverifiable Messages on Online Social Networks,” Information Systems Research (33:3), pp. 978–1001.

Luca, M. 2011. “Reviews, Reputation, and Revenue: The Case of Yelp.com,” Harvard Business School Working Paper (12-016).

Luca, M., and Zervas, G. 2016. “Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud,” Management Science (62:12), pp. 3412–3427.

Ludwig, S., Van Laer, T., De Ruyter, K., and Friedman, M. 2016. “Untangling a Web of Lies: Exploring Automated Detection of Deception in Computer-Mediated Communication,” Journal of Management Information Systems (33:2), pp. 511–541.

Luo, X., Zhang, J., and Duan, W. 2013. “Social Media and Firm Equity Value,” Information Systems Research (24:1), pp. 146–163.

MacKinlay, A. C. 1997. “Event Studies in Economics and Finance,” Journal of Economic Literature (35:1), pp. 13–39.

Mallipeddi, R. R., Janakiraman, R., Kumar, S., and Gupta, S. 2021. “The Effects of Social Media Content Created by Human Brands on Engagement: Evidence from Indian General Election 2014,” Information Systems Research (32:1), pp. 212–237.

Malmendier, U., and Tate, G. 2009. “Superstar CEOs,” The Quarterly Journal of Economics (124:4), pp. 1593–1638.

Martin, K. D., Borah, A., and Palmatier, R. W. 2017. “Data Privacy: Effects on Customer and Firm Performance,” Journal of Marketing (81:1), pp. 36–58.

Matook, S., Dennis, A. R., and Wang, Y. M. 2022. “User Comments in Social Media Firestorms: A Mixed-Method Study of Purpose, Tone, and Motivation,” Journal of Management Information Systems (39:3), pp. 673–705.

Mayzlin, D. 2006. “Promotional Chat on the Internet,” Marketing Science (25:2), pp. 155–163.

Mayzlin, D., Dover, Y., and Chevalier, J. 2014. “Promotional Reviews: An Empirical Investigation of Online Review Manipulation,” American Economic Review (104:8), pp. 2421–2455.

McCabe, D. 2021. Zuckerberg, Dorsey and Pichai Testify About Disinformation. The New York Times (March 25). Retrieved June 10, 2024, from https://www.nytimes.com/2021/03/25/technology/zuckerberg-dorsey-and-pichai-testify-about-disinformation.html.

McCall, J. J. 1970. “Economics of Information and Job Search,” The Quarterly Journal of Economics (84:1), pp. 113–126.

McLachlan, S., and Cooper, P. 2022. How to Get More Views on YouTube [REAL Ones]. Hootsuite (February 9). Retrieved June 10, 2024, from https://blog.hootsuite.com/get-views-youtube/.

McPherson, M., Smith-Lovin, L., and Cook, J. M. 2001. “Birds of a Feather: Homophily in Social Networks,” Annual Review of Sociology (27), pp. 415–444.

Miller, S., Menard, P., Bourrie, D., and Sittig, S. 2024. “Integrating Truth Bias and Elaboration Likelihood to Understand How Political Polarisation Impacts Disinformation Engagement on Social Media,” Information Systems Journal (34:3), pp. 642–679.

Mitkina, M., Lee, S., and Tan, Y. 2023. “Effect of Online Fitness Challenges on User Exercising Behavior: The Case of Youtube Fitness Channels,” Working Paper.

Mohammed, F., and Salam, A. F. 2021. “Me and the Other Not Me – Deepfake as Digitally Constructed Alternate Deceptive Identity: Loss of Control Over One’s Identity and Consequences,” in Proceedings of the 42nd International Conference on Information Systems, Austin, TX.

Moravec, P., Minas, R., and Dennis, A. R. 2019. “Fake News on Social Media: People Believe What They Want to Believe When it Makes No Sense At All,” MIS Quarterly (43:4), pp. 1343–1360.

Moravec, P. L., Collis, A., and Wolczynski, N. 2023. “Countering State-Controlled Media Propaganda Through Labeling: Evidence from Facebook,” Information Systems Research (Forthcoming).

Moravec, P. L., Kim, A., and 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), pp. 987–1006.

Moravec, P. L., Kim, A., Dennis, A. R., and Minas, R. K. 2022. “Do You Really Know If It’s True? How Asking Users to Rate Stories Affects Belief in Fake News on Social Media,” Information Systems Research (33:3), pp. 887–907.

Morse, S. 2016. Twitter Data and Financial Markets: Fundamental and Consumer Analysis. Twitter Blog (March 14). Retrieved October 25, 2022, from https://blog.twitter.com/official/en_us/a/2016/twitter-data-and-financial-markets-fundamental-and-consumer-analysis.html.

Mostagir, M., Ozdaglar, A., and Siderius, J. 2022. “When is Society Susceptible to Manipulation?,” Management Science (68:10), pp. 7153–7175.

Mostagir, M., and Siderius, J. 2022. “Learning in a Post-Truth World,” Management Science (68:4), pp. 2860–2868.

Mostagir, M., and Siderius, J. 2023a. “Social Inequality and the Spread of Misinformation,” Management Science (69:2), pp. 968–995.

Mostagir, M., and Siderius, J. 2023b. “Strategic Reviews,” Management Science (69:2), pp. 904–921.

Mousavi, R., and Gu, B. 2019. “The Impact of Twitter Adoption on Lawmakers’ Voting Orientations,” Information Systems Research (30:1), pp. 133–153.

Muchnik, L., Aral, S., and Taylor, S. J. 2013. “Social Influence Bias: A Randomized Experiment,” Science (341:6146), pp. 647–651.

Mullainathan, S., and Shleifer, A. 2005. “The Market for News,” American Economic Review (95:4), pp. 1031–1053.

Murphy, D. L., Shrieves, R. E., and Tibbs, S. L. 2009. “Understanding the Penalties Associated with Corporate Misconduct: An Empirical Examination of Earnings and Risk,” Journal of Financial and Quantitative Analysis (44:1), pp. 55–83.

Naeem, S. B., Bhatti, R., and Khan, A. 2021. “An Exploration of How Fake News is Taking Over Social Media and Putting Public Health at Risk,” Health Information & Libraries Journal (38:2), pp. 143–149.

Ng, K. C., Ke, P. F., So, M. K., and Tam, K. Y. 2023. “Augmenting Fake Content Detection in Online Platforms: A Domain Adaptive Transfer Learning via Adversarial Training Approach,” Production and Operations Management (32:7), pp. 2101–2122.

Ng, K. C., Tang, J., and 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), pp. 898–930.

Nie, C., Zheng, Z. E., and Sarkar, S. 2022. “Competing with the Sharing Economy: Incumbents’ Reaction on Review Manipulation,” MIS Quarterly (46:3), pp. 1573–1602.

Nielsen, F. Å. 2011. “A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs,” Working Paper.

Nikitkov, A., and Bay, D. 2008. “Online Auction Fraud: Ethical Perspective,” Journal of Business Ethics (79:3), pp. 235–244.

Nofer, M., and Hinz, O. 2015. “Using Twitter to Predict the Stock Market: Where is the Mood Effect?,” Business & Information Systems Engineering (57:4), pp. 229–242.

Nyhan, B. 2020. “Facts and Myths About Misperceptions,” Journal of Economic Perspectives (34:3), pp. 220–236.

Nyhan, B., Porter, E., Reifler, J., and Wood, T. J. 2020. “Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability,” Political Behavior (42:3), pp. 939–960.

OECD/European Union 2020. Health at a Glance: Europe 2020: State of Health in the EU Cycle, Paris, France: OECD Publishing, https://doi.org/10.1787/82129230-en.

Oestreicher-Singer, G., and Sundararajan, A. 2012. “The Visible Hand? Demand Effects of Recommendation Networks in Electronic Markets,” Management Science (58:11), pp. 1963–1981.

Oezpolat, K., Gao, G. G., Jank, W., and Viswanathan, S. 2013. “Research Note – The Value of Third-Party Assurance Seals in Online Retailing: An Empirical Investigation,” Information Systems Research (24:4), pp. 1100–1111.

Oh, O., Agrawal, M., and Rao, H. R. 2013. “Community Intelligence and Social Media Services: A Rumor Theoretic Analysis of Tweets During Social Crises,” MIS Quarterly (37:2), pp. 407–426.

Ormans, L. 2016. 50 Journals used in FT Research Rank. Financial Times. Retrieved January 30, 2023, from https://www.ft.com/content/3405a512-5cbb-11e1-8f1f-00144feabdc0.

Papanastasiou, Y. 2020. “Fake News Propagation and Detection: A Sequential Model,” Management Science (66:5), pp. 1826–1846.

Paré, G., Trudel, M.-C., Jaana, M., and Kitsiou, S. 2015. “Synthesizing Information Systems Knowledge: A Typology of Literature Reviews,” Information & Management (52:2), pp. 183–199.

Pariser, E. 2011. The Filter Bubble: What the Internet is Hiding From You, London, UK: Penguin UK.

Park, S., Xie, M., and Xie, J. 2023. “Frontiers: Framing Price Increase as Discount: A New Manipulation of Reference Price,” Marketing Science (42:1), pp. 37–47.

Parker, G., and Van Alstyne, M. 2018. “Innovation, Openness, and Platform Control,” Management Science (64:7), pp. 3015–3032.

Parker, G. G., and Van Alstyne, M. W. 2005. “Two-Sided Network Effects: A Theory of

Information Product Design," Management Science (51:10), pp. 1494–1504.

Parker, G. G., Van Alstyne, M. W., and Choudary, S. P. 2016. Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make them Work for You, New York, NY: WW Norton & Company.

Pelletier, M. J., Krallman, A., Adams, F. G., and Hancock, T. 2020. "One Size Doesn't Fit All: A Uses and Gratifications Analysis of Social Media Platforms," Journal of Research in Interactive Marketing (14:2), pp. 269–284.

Peng, J., Hahn, J., and Huang, K.-W. 2023. "Handling Missing Values in Information Systems Research: A Review of Methods and Assumptions," Information Systems Research (34:1), pp. 5–26.

Pennycook, G., Bear, A., Collins, E. T., and 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), pp. 4944–4957.

Pennycook, G., and Rand, D. G. 2019. "Fighting Misinformation on Social Media Using Crowdsourced Judgments of News Source Quality," Proceedings of the National Academy of Sciences (116:7), pp. 2521–2526.

Pennycook, G., and Rand, D. G. 2021. "The Psychology of Fake News," Trends in Cognitive Sciences (25:5), pp. 388–402.

Phelippeaux, C. 2023. YouTube Met en Avant les Professionels de Santé Grâce à un Nouvel Outil. Les Gens d'Internet (September 7). Retrieved October 12, 2023, from https://gensdinternet.fr/2023/09/07/youtube-met-en-avant-les-professionnels-de-sante-grace-a-un-nouvel-outil/.

Pian, W., Chi, J., and Ma, F. 2021. "The Causes, Impacts and Countermeasures of COVID-19 "Infodemic"': A Systematic Review Using Narrative Synthesis," Information Processing & Management (58:6), 102713.

Piccolo, S., Tedeschi, P., and Ursino, G. 2018. "Deceptive Advertising with Rational Buyers," Management Science (64:3), pp. 1291–1310.

Pierce, L., Snow, D. C., and McAfee, A. 2015. "Cleaning House: The Impact of Information Technology Monitoring on Employee Theft and Productivity," Management Science (61:10), pp. 2299–2319.

Poch, R., and Martin, B. 2015. "Effects of Intrinsic and Extrinsic Motivation on User-Generated Content," Journal of Strategic Marketing (23:4), pp. 305–317.

Porter, E., and 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.

Prawesh, S., and Padmanabhan, B. 2014. "The "Most Popular News" Recommender: Count Amplification and Manipulation Resistance," Information Systems Research (25:3), pp. 569–589.

Pu, J., Nian, T., Qiu, L., and Cheng, H. K. 2022. "Platform Policies and Sellers' Competition in Agency Selling in the Presence of Online Quality Misrepresentation," Journal of Management Information Systems (39:1), pp. 159–186.

Qiao, D., and Rui, H. 2023. "Text Performance on the Vine Stage? The Effect of Incentive on Product Review Text Quality," Information Systems Research (34:2), pp.

676–697.

Qiu, L., and Kumar, S. 2017. “Understanding Voluntary Knowledge Provision and Content Contribution Through a Social-Media-Based Prediction Market: A Field Experiment,” Information Systems Research (28:3), pp. 529–546.

Rabin, M., and Schrag, J. L. 1999. “First Impressions Matter: A Model of Confirmatory Bias,” The Quarterly Journal of Economics (114:1), pp. 37–82.

Rao, A. 2022. “Deceptive Claims Using Fake News Advertising: The Impact on Consumers,” Journal of Marketing Research (59:3), pp. 534–554.

Rao, A., and Wang, E. 2017. “Demand for “Healthy” Products: False Claims and FTC Regulation,” Journal of Marketing Research (54:6), pp. 968–989.

Rauchfleisch, A., and Kaiser, J. 2020. “The False Positive Problem of Automatic Bot Detection in Social Science Research,” PloS One (15:10), e0241045.

Rhoades, S. A. 1993. “The Herfindahl-Hirschman Index,” Federal Reserve Bulletin (79:3), pp. 188–189.

Rietveld, J., Schilling, M. A., and Bellavitis, C. 2019. “Platform Strategy: Managing Ecosystem Value Through Selective Promotion of Complements,” Organization Science (30:6), pp. 1232–1251.

Riquelme, I. P., and Román, S. 2014. “The Influence of Consumers’ Cognitive and Psychographic Traits on Perceived Deception: A Comparison Between Online and Offline Retailing Contexts,” Journal of Business Ethics (119:3), pp. 405–422.

Rockmann, K. W., and Northcraft, G. B. 2008. “To Be Or Not To Be Trusted: The Influence of Media Richness on Defection and Deception,” Organizational Behavior and Human Decision Processes (107:2), pp. 106–122.

Roggeveen, A. L., and Johar, G. V. 2002. “Perceived Source Variability Versus Familiarity: Testing Competing Explanations for the Truth Effect,” Journal of Consumer Psychology (12:2), pp. 81–91.

Román, S. 2010. “Relational Consequences of Perceived Deception in Online Shopping: The Moderating Roles of Type of Product, Consumer’s Attitude Toward the Internet and Consumer’s Demographics,” Journal of Business Ethics (95:3), pp. 373–391.

Roozenbeek, J., Van der Linden, S., Goldberg, B., Rathje, S., and Lewandowsky, S. 2022. “Psychological Inoculation Improves Resilience Against Misinformation on Social Media,” Science Advances (8:34), eabo6254.

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., and Stieglitz, S. 2019. “Are Social Bots a Real Threat? An Agent-Based Model of the Spiral of Silence to Analyse the Impact of Manipulative Actors in Social Networks,” European Journal of Information Systems (28:4), pp. 394–412.

Ross, S. A. 1973. “The Economic Theory of Agency: The Principal’s Problem,” The American Economic Review (63:2), pp. 134–139.

Rubin, J., Samek, A., and Sheremeta, R. M. 2018. “Loss Aversion and the Quantity–Quality Tradeoff,” Experimental Economics (21), pp. 292–315.

Rust, R. T., Rand, W., Huang, M.-H., Stephen, A. T., Brooks, G., and Chabuk, T. 2021. “Real-Time Brand Reputation Tracking Using Social Media,” Journal of Marketing (85:4), pp. 21–43.

Sadler, E. 2021. “A Practical Guide to Updating Beliefs from Contradictory Evidence,” Econometrica (89:1), pp. 415–436.

Salge, C. A. d. L., Karahanna, E., and Thatcher, J. B. 2022. “Algorithmic Processes of Social Alertness and Social Transmission: How Bots Disseminate Information on Twitter,” MIS Quarterly (46:1), pp. 229–259.

Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. 2020. “Detection of Novel Social Bots by Ensembles of Specialized Classifiers,” in Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, Ireland, pp. 2725–2732.

Scholten, S., and Scholten, U. 2012. “Platform-Based Innovation Management: Directing External Innovational Efforts in Platform Ecosystems,” Journal of the Knowledge Economy (3:2), pp. 164–184.

Schuetz, S. W., Sykes, T. A., and Venkatesh, V. 2021. “Combating COVID-19 Fake News on Social Media Through Fact Checking: Antecedents and Consequences,” European Journal of Information Systems (30:4), pp. 376–388.

Serra-Garcia, M., and Gneezy, U. 2021. “Mistakes, Overconfidence, and the Effect of Sharing on Detecting Lies,” American Economic Review (111:10), pp. 3160–3183.

Shalit, S. S., and Sankar, U. 1977. “The Measurement of Firm Size,” The Review of Economics and Statistics (59:3), pp. 290–298.

Shannon, C. E., and Weaver, W. 1949. The Mathematical Theory of Communication, University of Illinois Press.

Shapiro, C., Varian, H. R., Carl, S., et al. 1998. Information Rules: A Strategic Guide to the Network Economy, Boston, MA: Harvard Business School Press.

Sher, S. 2011. “A Framework for Assessing Immorally Manipulative Marketing Tactics,” Journal of Business Ethics (102:1), pp. 97–118.

Shi, Z., Liu, X., and Srinivasan, K. 2022. “Hype News Diffusion and Risk of Misinformation: The Oz Effect in Health Care,” Journal of Marketing Research (59:2), pp. 327–352.

Shirish, A., Srivastava, S. C., and Chandra, S. 2021. “Impact of Mobile Connectivity and Freedom on Fake News Propensity During the COVID-19 Pandemic: A Cross-Country Empirical Examination,” European Journal of Information Systems (30:3), pp. 322–341.

Siering, M., and Janze, C. 2019. “Information Processing on Online Review Platforms,” Journal of Management Information Systems (36:4), pp. 1347–1377.

Siering, M., Koch, J.-A., and Deokar, A. V. 2016. “Detecting Fraudulent Behavior on Crowdfunding Platforms: The Role of Linguistic and Content-Based Cues in Static and Dynamic Contexts,” Journal of Management Information Systems (33:2), pp. 421–455.

Silva, B. C., and Proksch, S.-O. 2021. “Fake It ‘Til You Make It: A Natural Experiment to Identify European Politicians’ Benefit from Twitter Bots,” American Political Science Review (115:1), pp. 316–322.

Simon, H. A. 1971. “Designing Organizations for an Information-Rich World,” in Computers, Communications, and the Public Interest, Greenberger, M. (ed.), Baltimore, MD: Johns Hopkins Press, pp. 37–72.

Sjöblom, M., Törhönen, M., Hamari, J., and Macey, J. 2019. “The Ingredients of Twitch

Streaming: Affordances of Game Streams," Computers in Human Behavior (92), pp. 20–28.

Skurnik, I., Yoon, C., Park, D. C., and Schwarz, N. 2005. "How Warnings about False Claims Become Recommendations," Journal of Consumer Research (31:4), pp. 713–724.

Social Puncher 2018. Twitter Purge: True Story – Part 1. October. Retrieved February 13, 2023, from https://socialpuncher.com/.

Song, R., Kim, H., Lee, G. M., and Jang, S. 2019. "Does Deceptive Marketing Pay? The Evolution of Consumer Sentiment Surrounding a Pseudo-Product-Harm Crisis," Journal of Business Ethics (158:3), pp. 743–761.

Sorescu, A., Warren, N. L., and Ertekin, L. 2017. "Event Study Methodology in the Marketing Literature: An Overview," Journal of the Academy of Marketing Science (45:2), pp. 186–207.

Spence, M. 1973. "Job Market Signaling," The Quarterly Journal of Economics (87:3), pp. 335–374.

Stanley, M. L., Whitehead, P. S., and Marsh, E. J. 2022. "The Cognitive Processes Underlying False Beliefs," Journal of Consumer Psychology (32:2), pp. 359–369.

Stieglitz, S., and Dang-Xuan, L. 2013. "Emotions and Information Diffusion in Social Media——Sentiment of Microblogs and Sharing Behavior," Journal of Management Information Systems (29:4), pp. 217–248.

Stiglitz, J. E. 2000. "The Contributions of the Economics of Information to Twentieth Century Economics," The Quarterly Journal of Economics (115:4), pp. 1441–1478.

Suler, J. 2004. "The Online Disinhibition Effect," Cyberpsychology & Behavior (7:3), pp. 321–326.

Sun, Y., Dong, X., and McIntyre, S. 2017. "Motivation of User-Generated Content: Social Connectedness Moderates the Effects of Monetary Rewards," Marketing Science (36:3), pp. 329–337.

Szabo, S., and Webster, J. 2021. "Perceived Greenwashing: The Effects of Green Marketing on Environmental and Product Perceptions," Journal of Business Ethics (171:4), pp. 719–739.

Tafti, A., Zotti, R., and Jank, W. 2016. "Real-Time Diffusion of Information on Twitter and the Financial Markets," PloS One (11:8), e0159226.

Tandoc Jr, E. C., Lim, Z. W., and Ling, R. 2018. "Defining "Fake News": A Typology of Scholarly Definitions," Digital Journalism (6:2), pp. 137–153.

Tellis, G. J., MacInnis, D. J., Tirunillai, S., and Zhang, Y. 2019. "What Drives Virality (Sharing) of Online Digital Content? The Critical Role of Information, Emotion, and Brand Prominence," Journal of Marketing (83:4), pp. 1–20.

Ter Huurne, M., Ronteltap, A., Corten, R., and Buskens, V. 2017. "Antecedents of Trust in the Sharing Economy: A Systematic Review," Journal of Consumer Behaviour (16:6), pp. 485–498.

Tergiman, C., and Villeval, M. C. 2023. "The Way People Lie in Markets: Detectable vs. Deniable Lies," Management Science (69:6), pp. 3340–3357.

Terlaak, A., and King, A. A. 2006. “The Effect of Certification with the ISO 9000 Quality Management Standard: A Signaling Approach,” Journal of Economic Behavior & Organization (60:4), pp. 579–602.

Teubner, T., Adam, M. T., and Hawlitschek, F. 2020. “Unlocking Online Reputation: On the Effectiveness of Cross-Platform Signaling in the Sharing Economy,” Business & Information Systems Engineering (62:6), pp. 501–513.

Teubner, T., Saade, N., Hawlitschek, F., and Weinhardt, C. 2016. “It’s Only Pixels, Badges, and Stars: On the Economic Value of Reputation on Airbnb,” in Proceedings of the 27th Australasian Conference on Information Systems, Wollongong, Australia.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. 2010. “Sentiment Strength Detection in Short Informal Text,” Journal of the American Society for Information Science and Technology (61:12), pp. 2544–2558.

TikTok 2024. Number of Fake Interactions Removed on TikTok from 2nd Quarter 2021 to 4th Quarter 2023. Statista (March 19). Retrieved June 24, 2024, from https://www.statista.com/statistics/1318295/tiktok-fake-interactions-removed/.

Tipton, M. M., Bharadwaj, S. G., and Robertson, D. C. 2009. “Regulatory Exposure of Deceptive Marketing and its Impact on Firm Value,” Journal of Marketing (73:6), pp. 227–243.

Tiwana, A., Konsynski, B., and Bush, A. A. 2010. “Research Commentary – Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics,” Information Systems Research (21:4), pp. 675–687.

Törhönen, M., Sjöblom, M., Hassan, L., and Hamari, J. 2020. “Fame and Fortune, or Just Fun? A Study on Why People Create Content on Video Platforms,” Internet Research (30:1), pp. 165–190.

Trendel, O., Mazodier, M., and Vohs, K. D. 2018. “Making Warnings About Misleading Advertising and Product Recalls More Effective: An Implicit Attitude Perspective,” Journal of Marketing Research (55:2), pp. 265–276.

Turel, O., and Osatuyi, B. 2021. “Biased Credibility and Sharing of Fake News on Social Media: Considering Peer Context and Self-Objectivity State,” Journal of Management Information Systems (38:4), pp. 931–958.

Twitter Data 2016. Twitter Data and the Financial Markets. Twitter Blog (July 28). Retrieved October 25, 2022, from https://blog.twitter.com/en_us/topics/insights/2016/twitter-data-and-the-financial-markets.

Twyman, N. W., Proudfoot, J. G., Cameron, A.-F., Case, E., Burgoon, J. K., and Twitchell, D. P. 2020. “Too Busy to be Manipulated: How Multitasking with Technology Improves Deception Detection in Collaborative Teamwork,” Journal of Management Information Systems (37:2), pp. 377–395.

Ullah, S., Massoud, N., and Scholnick, B. 2014. “The Impact of Fraudulent False Information on Equity Values,” Journal of Business Ethics (120:2), pp. 219–235.

Ursu, R. M. 2018. “The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions,” Marketing Science (37:4), pp. 530–552.

Valsesia, F., Proserpio, D., and Nunes, J. C. 2020. “The Positive Effect of Not Following

Others on Social Media," Journal of Marketing Research (57:6), pp. 1152–1168.

Van Alstyne, M., Smith, M. D., and Lin, H. 2023. "Improving Section 230, Preserving Democracy, and Protecting Free Speech," Communications of the ACM (66:4), pp. 26–28.

Van Bommel, J. 2003. "Rumors," The Journal of Finance (58:4), pp. 1499–1520.

Van Roy, B., and Yan, X. 2010. "Manipulation Robustness of Collaborative Filtering," Management Science (56:11), pp. 1911–1929.

Van Scotter, J. R., and Roglio, K. D. D. 2020. "CEO Bright and Dark Personality: Effects on Ethical Misconduct," Journal of Business Ethics (164:3), pp. 451–475.

Vasist, P. N., and Krishnan, S. 2022. "Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research," Communications of the Association for Information Systems (51), pp. 590–636.

Vom Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., and Cleven, A. 2009. "Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process," in Proceedings of the 17th European Conference on Information Systems, Verona, Italy.

Vosoughi, S., Roy, D., and Aral, S. 2018. "The Spread of True and False News Online," Science (359:6380), pp. 1146–1151.

Wang, C., Zhang, X., and Hann, I.-H. 2018. "Socially Nudged: A Quasi-Experimental Study of Friends' Social Influence in Online Product Ratings," Information Systems Research (29:3), pp. 641–655.

Wang, J., Li, Y., and Rao, H. R. 2017. "Coping Responses in Phishing Detection: An Investigation of Antecedents and Consequences," Information Systems Research (28:2), pp. 378–396.

Wang, S., Pang, M.-S., and Pavlou, P. A. 2021a. "Cure or Poison? Identity Verification and the Posting of Fake News on Social Media," Journal of Management Information Systems (38:4), pp. 1011–1038.

Wang, S., Pang, M.-S., and Pavlou, P. A. 2022. "Seeing Is Believing? How Including a Video in Fake News Influences Users' Reporting of the Fake News to Social Media Platforms," MIS Quarterly (46:3), pp. 1323–1354.

Wang, Y.-Y., Guo, C., Susarla, A., and Sambamurthy, V. 2021b. "Online to Offline: The Impact of Social Media on Offline Sales in the Automobile Industry," Information Systems Research (32:2), pp. 582–604.

Wardle, C., and Derakhshan, H. 2017. Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking, Strasbourg, FR: Council of Europe.

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," MIS Quarterly (26:2), pp. xiii–xxiii.

Wei, X., Zhang, Z., Zhang, M., Chen, W., and Zeng, D. D. 2022. "Combining Crowd and Machine Intelligence to Detect False News on Social Media," MIS Quarterly (46:2), pp. 977–1008.

Wei, Z., Xiao, M., and Rong, R. 2021. "Network Size and Content Generation on Social Media Platforms," Production and Operations Management (30:5), pp. 1406–1426.

Weinmann, M., Valacich, J., Schneider, C., Jenkins, J. L., and Hibbeln, M. T. 2022. “The Path of the Righteous: Using Trace Data to Understand Fraud Decisions in Real Time,” MIS Quarterly (46:4), pp. 2317–2336.

Weitzman, M. 1979. “Optimal Search for the Best Alternative,” Econometrica (47:3), pp. 641–654.

Weiß, E.-M. 2023. Youtube Health führt Siegel ein: Dr. Google soll sicherer werden. Heise Online (February 28). Retrieved March 16, 2023, from https://www.heise.de/news/Youtube-Health-startet-in-Deutschland-Siegel-und-Schelfe-fuer-Inhalteersteller-7530139.html.

Welt 2022. YouTube führt neues Label für Gesundheitsinformationen ein. October 18. Retrieved April 23, 2024, from https://www.welt.de/vermischtes/article241649649/YouTube-Health-Plattform-fuehrt-Label-fuer-Gesundheitsinformationen-ein.html.

Wies, S., Bleier, A., and Edeling, A. 2023. “Finding Goldilocks Influencers: How Follower Count Drives Social Media Engagement,” Journal of Marketing (87:3), pp. 383–405.

Wilbur, K. C., and Zhu, Y. 2009. “Click Fraud,” Marketing Science (28:2), pp. 293–308.

Wiles, M. A., Jain, S. P., Mishra, S., and Lindsey, C. 2010. “Stock Market Response to Regulatory Reports of Deceptive Advertising: The Moderating Effect of Omission Bias and Firm Reputation,” Marketing Science (29:5), pp. 828–845.

Wilson, A. E., Darke, P. R., and Sengupta, J. 2022. “Winning the Battle but Losing the War: Ironic Effects of Training Consumers to Detect Deceptive Advertising Tactics,” Journal of Business Ethics (181), pp. 997–1013.

Wöhner, T., Köhler, S., and Peters, R. 2015. “Managed Wikis: A New Approach for Web 2.0,” Business & Information Systems Engineering (57:3), pp. 155–166.

Wooldridge, J. M. 2019. Introductory Econometrics: A Modern Approach, 4th ed., Mason, OH: South Western, Cengage Learning.

World Health Organization 2022. Health Topics. Retrieved December 5, 2022, from https://www.who.int/health-topics.

Wu, Y., and Geylani, T. 2020. “Regulating Deceptive Advertising: False Claims and Skeptical Consumers,” Marketing Science (39:4), pp. 788–806.

Xiao, B., and Benbasat, I. 2011. “Product-Related Deception in E-Commerce: A Theoretical Perspective,” MIS Quarterly (35:1), pp. 169–195.

Xiao, B., and Benbasat, I. 2015. “Designing Warning Messages for Detecting Biased Online Product Recommendations: An Empirical Investigation,” Information Systems Research (26:4), pp. 793–811.

Xiao, Y., and Watson, M. 2019. “Guidance on Conducting a Systematic Literature Review,” Journal of Planning Education and Research (39:1), pp. 93–112.

Xie, G.-X., Chang, H., and Rank-Christman, T. 2022. “Contesting Dishonesty: When and Why Perspective-Taking Decreases Ethical Tolerance of Marketplace Deception,” Journal of Business Ethics (175:1), pp. 117–133.

Xie, G.-X., Madrigal, R., and Boush, D. M. 2015. “Disentangling the Effects of Perceived Deception and Anticipated Harm on Consumer Responses to Deceptive Advertising,” Journal of Business Ethics (129:2), pp. 281–293.

Xu, D. J., Cenfetelli, R. T., and Aquino, K. 2012. “The Influence of Media Cue Multiplicity on Deceivers and Those Who are Deceived,” Journal of Business Ethics (106:3), pp. 337–352.

Yan, L., and Tan, Y. 2014. “Feeling Blue? Go Online: An Empirical Study of Social Support Among Patients,” Information Systems Research (25:4), pp. 690–709.

Yang, S.-B., Lim, J.-H., Oh, W., Animesh, A., and Pinsonneault, A. 2012. “Research Note – Using Real Options to Investigate the Market Value of Virtual World Businesses,” Information Systems Research (23:3-Part-2), pp. 1011–1029.

Yip, J. A., and Schweitzer, M. E. 2016. “Mad and Misleading: Incidental Anger Promotes Deception,” Organizational Behavior and Human Decision Processes (137), pp. 207–217.

YouTube Help 2023. Apply to be a Source in YouTube Health Features. Retrieved March 12, 2024, from https://support.google.com/youtube/answer/12796915.

YouTube Help 2024a. Get Info on Health-Related Content. Retrieved March 12, 2024, from https://support.google.com/youtube/answer/9795167.

YouTube Help 2024b. How Engagement Metrics are Counted. Retrieved March 12, 2024, from https://support.google.com/youtube/answer/2991785?hl=en.

Zahra, S. A., Priem, R. L., and Rasheed, A. A. 2005. “The Antecedents and Consequences of Top Management Fraud,” Journal of Management (31:6), pp. 803–828.

ZEIT Online 2022. YouTube führt Label für verlässliche Gesundheitsinformationen ein. October 18. Retrieved April 23, 2024, from https://www.zeit.de/digital/internet/2022-10/youtube-gesundheit-label-quellen-google.

Zhang, D., Zhou, L., Kehoe, J. L., and Kilic, I. Y. 2016. “What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews,” Journal of Management Information Systems (33:2), pp. 456–481.

Zhang, M., and Luo, L. 2023. “Can Consumer-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence from Yelp,” Management Science (69:1), pp. 25–50.

Zhang, X., Du, Q., and Zhang, Z. 2022. “A Theory-Driven Machine Learning System for Financial Disinformation Detection,” Production and Operations Management (31:8), pp. 3160–3179.

Zhao, Y., Yang, S., Narayan, V., and Zhao, Y. 2013. “Modeling Consumer Learning from Online Product Reviews,” Marketing Science (32:1), pp. 153–169.

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., and Nunamaker Jr, J. F. 2004. “A Comparison of Classification Methods for Predicting Deception in Computer-Mediated Communication,” Journal of Management Information Systems (20:4), pp. 139–166.

Zhou, R., Khemmarat, S., Gao, L., Wan, J., and Zhang, J. 2016. “How YouTube Videos are Discovered and Its Impact on Video Views,” Multimedia Tools and Applications (75:10), pp. 6035–6058.