Faculty of Social Sciences
University of Helsinki

Dissertationes Universitatis Helsingiensis
424/2026

STANDING UP TO MISINFORMATION

Observing corrective agency on social media

Tuomas Heikkilä

ACADEMIC DISSERTATION

To be presented, with the permission of the faculty of Social Sciences
of the University of Helsinki, for public examination in
lecture room 6 (B316), Metsätalo, on 2 October 2026, at 13.

Helsinki 2026

Preliminary examiners

Associate Professor Michael Hameleers, University of Amsterdam

Associate Professor Ethan Porter, George Washington University

Custos

Adjunct Professor, Senior Researcher Salla-Maaria Laaksonen, University of Helsinki

Supervisor

Professor Mervi Pantti, University of Helsinki

Professor Emeritus Esa Väliverronen, University of Helsinki

Adjunct Professor, Senior Researcher Salla-Maaria Laaksonen, University of Helsinki

Opponent

Professor Emily Vraga, University of Minnesota

Publisher: University of Helsinki
Series: Dissertationes Universitatis Helsingiensis 424/2026

© Tuomas Heikkilä
Cover Image: Strindbergin huoneissa #16, 2017, © Marja Pirilä

ISBN 978-952-84-2965-4 (paperback)
ISBN 978-952-84-2964-7 (PDF)
ISSN 2954-2898 (print)
ISSN 2954-2952 (PDF)

http://urn.fi/URN:ISBN:978-952-84-2964-7

Unigrafia / PunaMusta
Joensuu 2026

Abstract

Concerns of misinformation around the world have renewed academic interest in methods that could stymie its worst effects. Much of this work has centred on testing the effectiveness of corrections, or messages designed to encourage people to reject misinformation in favour of more accurate accounts of the world. Consequently, we have a great deal of experimental evidence of the type of corrections that work in principle. Yet, surprisingly little is known, of corrections as they are enacted 'in the wild': why and how often they occur, where they can be found, and what forms they take.

In this dissertation, I address this gap by examining the everyday correction of misinformation on social media platforms. Drawing of different understandings of correction in psychology and communication studies, I introduce the concept of 'corrective agency', or a capacity to publicly confront people for making claims that the corrector rejects as misinformation, presumably to convince others to adopt their own understanding of facts. Through this lens, I highlight that 'setting the record straight' requires more than favourable message characteristics or communication contexts. Rather, corrections emerge through a series of decisions to recognise misinformation (verification) and publicly respond to it (confrontation), all of which are shaped by broader social and technological conditions, personal goals, and practical considerations.

The dissertation comprises a comprehensive summary and three research articles, that analyse different aspects of corrective agency and the actors involved in it. Recognising that the framing of a problem shapes its proposed solutions, Article I reviews the ongoing debate concerning the severity and nature of the current 'misinformation problem'. The remaining articles use found digital data to examine how correctors justify their actions in response to misinformation (Article II), and how such confrontations actually unfold on social media platforms (Article III).

Taken together, the dissertation demonstrates the tensions between real-world corrections and idealised accounts of effective and responsible misinformation mitigation. On the one hand, real-world corrections appear to be prevalent and typically follow the recommendations of experimentally validated 'best practices'. They also reflect relatively sound and principled epistemic standards the aim of which is to promote factual accuracy in public communication. However, commitment to correction often appears shallow, ambivalent and, at times, opportunistic. Corrections are also frequently undertaken without a clearly articulated understanding of how misinformation operates or how it is expected to be mitigated. The comprehensive summary concludes by proposing a research agenda for reimagining misinformation mitigation through a descriptive and naturalistic approach. Finally, it considers how this agency might be sustained and developed both on its own terms and in dialogue with democratic institutions.

Tiivistelmä

Maailmanlaajuinen huoli misinformaatiosta on herättänyt keskustelua keinoista, joiden avulla sen pahimpia vaikutuksia voitaisiin hillitä. Aihetta käsittelevä tutkimus on keskittynyt erityisesti testaamaan misinformaatiota oikaisevien korjausten tehokkuutta, eli viestejä, jotka tarjoavat misinformaation tilalle oikeaa tietoa ja tällä tavoin kannustavat väärän tiedon hylkäämiseen. Kokeellisen näytön perusteella tunnetaankin periaatteessa jo verrattain hyvin, milloin korjaaminen toimii parhaimmillaan. Tästä huolimatta tiedämme yllättävän vähän oikaisuista luonnollisessa ympäristössä, eli siitä miksi ja kuinka usein oikaisuja ilmenee, mistä niitä löytää, ja millaisia muotoja ne lopulta saavat.

Tässä väitöskirjassa tutkin, miten väärää tietoa todella korjataan sosiaalisen median alustoilla. Psykologian ja viestinnän tutkimukseen pohjautuen käsitteellistän *korjaavan toimijuuden* viestijän kyvyksi kiistellä julkisesti tosiasioista muiden, väärää tietoa esittävien viestijöiden kanssa, oletettavasti saadakseen nämä omaksumaan oman käsityksensä tosiasioista. Tästä näkökulmasta tarkasteltuna korjaaminen on toimintaa, joka edellyttää muutakin kuin sommiteltujen viestien välittämistä suotuisissa olosuhteissa: oikaisut syntyvät sen sijaan lukuisten, misinformaation tunnistamiseen tähtäävien valintojen (varmentaminen) kautta, ja huipentuvat misinformaatioon kohdistuvaan julkiseen puuttumiseen (yhteenotto) sosiaalisten ja teknologisten rakenteiden, henkilökohtaisten tavoitteiden sekä käytännön näkökohtien puitteissa.

Väitöskirja koostuu yhteenvedosta ja kolmesta tutkimusartikkelista, jotka tarkastelevat korjaavaa toimijuutta ja korjaamiseen osallistuvia toimijoita eri näkökulmista. Artikkeli I on kirjallisuuskatsaus, joka syntetisoi ajankohtaista keskustelua “misinformaatio-ongelman” vakavuudesta ja luonteesta lähtökohtanaan huomio siitä, että mainitun ongelman julkinen määrittely muovaa myös siihen esitettyjä ratkaisuja. Kaksi muuta artikkelia ovat empiirisiä tutkimuksia, jotka hyödyntävät luonnollista digitaalista aineistoa käsitelläkseen sitä, miten väärän tiedon korjaajat perustelevat toimintaansa (Artikkeli II) ja miten tällaiset yhteenotot todella ilmenevät sosiaalisen median alustoilla (Artikkeli III).

Yhteenvetona väitöskirja osoittaa jännitteen, joka vallitsee sosiaalisen median alustoilla havaittujen oikaisujen ja korjaamisen ihanteiden välillä. Yhtäältä oikaisut vaikuttavat olevan verrattain yleisiä ja noudattelevan usein ajantasaisen tutkimuksen suosittamia parhaita käytäntöjä. Ne myös edustavat suhteellisen järkeviä ja periaatteellisia tiedollisia normeja, ja pyrkivät edistämään julkisen viestinnän todenperäisyyttä. Toisaalta sitoutuminen oikaisemiseen on usein pinnallista, ristiriitaista, ja jopa opportunistista. Korjaamisen taustalta vaikuttaa ajoittain myös puuttuvan selkeä näkemys siitä, miten misinformaatio vaikuttaa ja miten sen korjaamisen odotetaan toimivan. Näiden löydösten pohjalta yhteenveto päättyy hahmotelmaan tutkimusohjelman, joka jäsentää väärän tiedon torjuntaa kuvailevan ja luonnollisia tilanteita korostavan lähestymistavan kautta. Lopuksi pohditaan, miten toimijuutta voitaisiin vaalia ja kehittää sekä omilla ehdoillaan että vuoropuhelussa demokraattisten instituutioiden kanssa.

Acknowledgements

I owe this dissertation to the support and goodwill of countless people.

First and foremost, I want to thank my stellar combination of supervisors Esa Väliverronen, Mervi Pantti and Salla-Maaria Laaksonen for taking care of me during this academic journey. You have been my constant source of wisdom, cheer and support. It is through your mentoring and effort that I have built the confidence, skills and integrity to see this dissertation through. I will be always grateful for believing in me.

I am deeply humbled by the calibre of the experts who donated their time to examine and review this work before its publication. I thank my pre-examiners, Associate Professors Michael Hameleers and Ethan Porter, for your valuable encouragement and constructive suggestions. The distinguished members of my thesis committee – Associate Professor Mia-Marie Hammarlin, Adjunct Professor Minna Horowitz, University Researcher Aleksi Knuutila, and Doctoral Researcher and Editor-in-Chief Joonas Pörsti – were also instrumental in identifying areas for improvement and in providing overall support for this work in its final stages. Finally, I would like to express my deepest gratitude to Professor Emily Vraga for agreeing to act as the opponent at my public defence. I am hard pressed to think of anyone who has had a greater impact on my thinking about real-world corrections from the outset.

I want to sincerely thank Matti Pohjonen and Salla for co-authoring an article that forms a part dissertation. Moreover, I would like to acknowledge the numerous projects that were instrumental in conducting this research: *UnCoCo Unconventional Communicators in Corona Crisis*, *MEX - Mediating Expertise* consortium, and *MAPS - Media platforms and social accountability* to whose meetings and seminars I was always welcomed to attend. It was a true joy and a privilege to meet, learn from, and collaborate with the members of these research teams: Essi Pöyry, Hanna Reinikainen, Juho Pääkkönen, Minna Vigren, Niko Hatakka, Sampsa Saikkonen, Hilma Anttila and Iida Lähdekorpi.

I thank The Finnish Cultural Foundation and the Doctoral Programme in Social Sciences for the extraordinary opportunity to work on this dissertation full-time over the past four years. I am also grateful to HYMY Doctoral School Travel Support for funding my conference travels. Proofreading of this dissertation was made possible by the Faculty of Social Sciences: Thank you to Mark Waller for his expert work.

I have had the exceptional opportunity to present earlier stages of this dissertation at numerous international conferences, workshops and seminars. I would like to express my sincere gratitude to the following mentors, session chairs and opponents for their valuable feedback: Sofya Glazunova, Niki Cheong, Broughton Micova, Sally Naíde Müller, Guobin Yang, Anette Grønning, Tobias Schmid, Niels Mede, Katja Valaskivi and Simon Lindgren. Special thanks go to Risto Kunelius and the other organisers of the 4th Lisbon Winter School for the Study of Communication, which played a crucial role in developing the ideas that later became my Article I. Above all, these travels gave me the opportunity to meet wonderful colleagues around the world: Julia Kling, Ottavio Vinhas, Lukas Mozdeika, Laura Aniot, Kata Horváth, Philip Schimchen, and Simona Bisiani – thank you for the camaraderie!

To my dear gang of colleagues: Anna-Liisa Heino, Elina Uutela, Tatu Matilainen, Aino Koskenniemi, Stefanie Sirén-Heikel, Svenja Scholz, Nuppu Pelevina, Viljami Vaarala, Outi Puukko, Arto Kekkonen, Annaliina Niitamo, Linda Pentikäinen, Joonas Koivukoski and Kaisa Tiusanen – thank you for making the workplace infinitely better. I could not have hoped for more fun, empathetic, wise, and kooky contemporaries. My warmest thanks go to the people of media communication studies. Special thanks go to Anu Kantola, Salli Hakala and Minttu Tikka who gave me my first opportunities in research assistance and teaching even before – and undoubtedly leading up to – my doctoral studies. Conversations with Kari Karppinen, Markus Ojala, and Timo Harjuniemi have helped me to develop and refine the ideas presented in the following pages. Finally, the Rajapinta association has given me an intellectual home. Thank you to Salla, Matti Nelimarkka, Maake Salonen and Elisa Kannasto for stewarding this community and to Essi, Hanna, Juho, Nuppu, Viljami, Minttu, Janne Matikainen, Erkki Mervaala, Esko Nieminen, Liisa Kääntä, Henna Paakki, Henri Pullinen, Sari Kivijärvi, Virpi Salojärvi, Jesse Haapoja, Tuukka Lehtiniemi, Merja Porttikivi, Elli Lillqvist, among many others, for your inspiring company and always being a great hang.

Finally, thank you to family and friends for always being there, and for sharing experiences that are so much bigger and better than this one project. Thank you to my relatives, Vilkunat ja Hollolan perustajien jälkeläiset, for modelling a life dedicated to the arts and sciences. Mom and Dad, you have always supported me in my academic (mis)adventures, even when your trust felt totally unearned. Thank you for raising me right. I love you very much.

To Emmi: thank you for sharing your life with me, tempering my nonsense, and correcting me when I most need it.

Helsinki, 3 September 2026

Tuomas Heikkilä

CONTENTS

Abstract iii

Tiivistelmä iv

Acknowledgements v

List of original publications ix

1 Introduction 1

1.1 Research problem and research question 4

2 Three perspectives on misinformation 6

2.1 Misinformation as an 'information disorder' 6

2.2 Misinformation as a psychological phenomenon 9

2.3 Misinformation as a discourse 11

3 Conceptual framework 14

3.1 Characterising corrections 14

3.1.1 Post-publication error-correction 17

3.1.2 Fact-checking 19

3.2 Elements of corrective agency 21

3.2.1 Verification 21

3.2.2 Confrontation 23

3.3 Who corrects misinformation on social media? 25

3.3.1 Moderating misinformation 26

3.3.2 User corrections 30

4 Research approach 32

4.1 Situating the research approach 32

4.2 Identifying corrections 'in the wild' 35

4.3 Research materials and their operationalisations 37

4.4 Ethical considerations 39

5 Key findings 41

5.1 Article I: Substantive critique of the misinformation problem as exaggerated, ambiguous, short-sighted and purposive presents new possibilities to reimagine its potential solutions 41

5.2 Article II: Moderating COVID-19 misinformation serves platform companies' strategic interests 43

5.3 Article III: User corrections on Twitter are common and sophisticated 44

6 Discussion 47
6.1 Conceptualising corrective agency 'in the wild' 48
6.2 Lessons from observationally grounded understanding of misinformation interventions 51
6.3 Conclusion and ways forward 54
References 58

List of original publications

This thesis is based on the following publications:

I Heikkilä, T. (2025). The contested problems of misleading communication: An integrative review of critical perspectives challenging the social problems of mis- and disinformation. Originally published in Finnish at *Media & viestintä*, *48*(1), 112–140. https://doi.org/10.23983/mv.148332

II Heikkilä, T., Laaksonen, S.-M., & Pohjonen, M. (2025). Reluctant arbiters of truth: Discursive legitimation of platform interventions against COVID-19 misinformation. *New Media & Society* (online first article). https://doi.org/10.1177/14614448251365270

III Heikkilä, T. (2026). Stop pushing junk science! User corrections of misinformation in science are common and sophisticated. *Public Understanding of Science* (online first article). https://doi.org/10.1177/09636625261470329

The publications are referred to in the text by their roman numerals.

1 INTRODUCTION

> *We cannot solve problems of public health, social inequality, or climate change without also addressing the growing problem of misinformation.*
>
> West and Bergstrom (2021, p. 1)

> *In a world relying on intertwined local and global digital infrastructures for processing and circulating information, fake news is here to stay. Fake news are communicative actions we must learn to live by.*
>
> Andersen and Søe (2020, p. 137)

Is misinformation a problem to be solved?

When I started working on this project in 2021, the troubles of online misinformation seemed beyond question. From the local QAnon supporters in Finland and the insurrection on Capitol Hill in the United States to the attacks on 5G network masts in the United Kingdom, the ivermectin craze in Peru and the anti-lockdown 'Querdenken' movement in Germany, it was easy to point to a common culprit: unsettling events everywhere were being driven by 'politically active falsehoods' (Hochschild & Einstein, 2015) that had emerged from and supported by the 'unedited public sphere' (Bimber & Gil de Zúñiga, 2020) of major digital platforms. In the space of just a few years, the idea of social media as a democratising force (Boulianne, 2019) had all but disappeared. It was replaced by a far more pessimistic view centred on the damage it causes to individuals and institutions (Lorenz-Spreen et al., 2023; Quandt, 2018). Since then, these concerns have perhaps grown even stronger worldwide (Altay, 2026a). Terms such as 'infodemics', 'disinformation campaigns' and 'deepfakes' have become mainstays in debates about the harms of social media platforms, joining a long list of previous pathologies, ranging from trolling and incivility to toxic anonymity and hate speech.

Academic research has been in lockstep with this general trend. After remaining relatively dormant for decades, attention to misleading and deceptive media techniques has grown exponentially every year since 2016 (Broda & Strömbäck, 2024), with little indication of slowing down (Figure 1). This rapid expansion has not come without challenges. While many scholars remain optimistic that this huge influx of multidisciplinary work could still form a coherent field of misinformation studies (Camargo & Simon, 2022), there is also growing concern about the

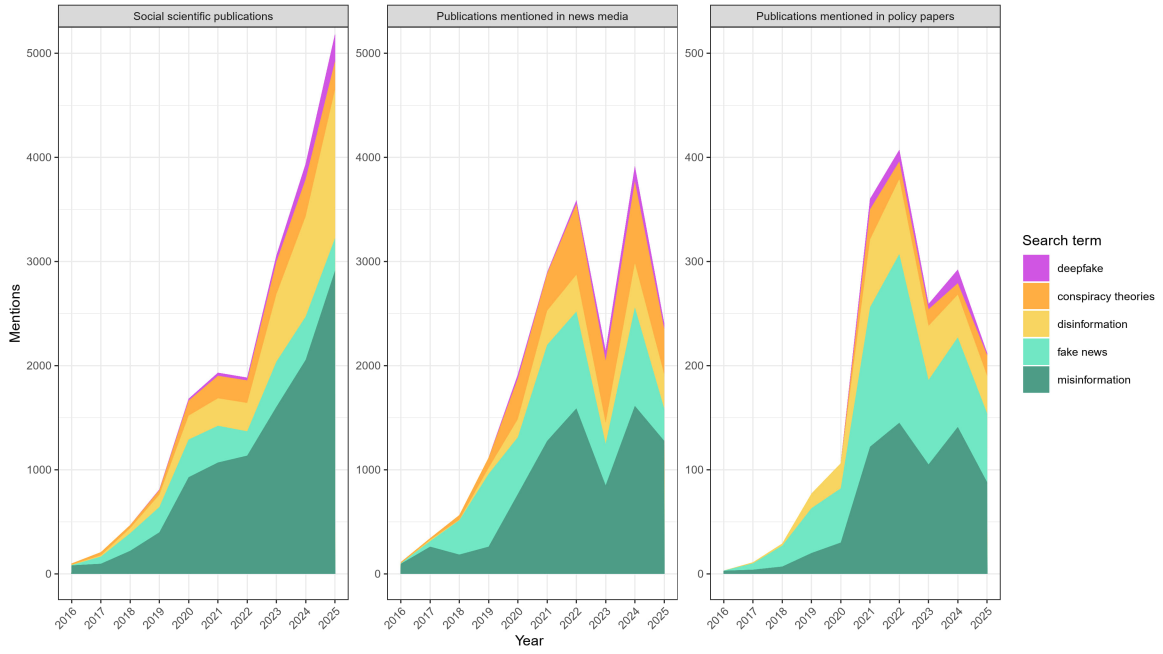


Figure 1 Growth of the scientific literature on misinformation from 2016 to 2025, adapted from Article I. The left panel shows the number of social science publications containing misinformation-related keywords in their title or abstract (data from Semantic Scholar). The middle and right panels show the number of times such publications were mentioned in news media and policy papers, respectively (data from Altmetric Scholar).

methodological (Murphy et al., 2023), conceptual (Altay et al., 2023) and epistemological (Uscinski et al., 2024) disagreements that continue to undermine this common research agenda. Despite these disputes, however, it is fair to say that there is at least one foundational assumption that unites many disparate scholars and their funding opportunities: misinformation is studied not only to understand it, but also to identify ways *to counter it more efficiently*.[1]

There are, of course, many valid reasons to believe that misinformation poses harms which warrant mitigation. For instance, it is well known that powerful interests can mislead communities about the risks they face (Oreskes & Conway, 2011), and it could be argued that governments that withhold accurate information from the public are failing in their duties to their citizens (Horowitz et al., 2024). Escaping into cosy fantasies is not a helpful strategy for improving societies (Blake-Turner, 2025), as this could devalue fact-based discourse for everyone (van Doorn, 2023). We could also argue that most people do not want to be wrong (Kunda, 1990) and those who are willing to lie to us are far from the best of us (Littrell et al., 2023). There is little doubt that misinformation provides a means of channelling dark

[1] A small but telling sign of this dual focus can be seen in the titles of notable scholarly works on misinformation research, including *Understanding And Combating Misinformation* (Arechar et al., 2023), *Why We Accept Misinformation And How To Fight It* (Berinsky, 2023), *Using Psychological Science to Understand and Fight Misinformation* (Van Der Linden et al., 2023), or simply *Misinformation And Its Correction* (Lewandowsky et al., 2012; Swire & Ecker, 2018; Wittenberg & Berinsky, 2020).

human impulses (Uscinski et al., 2022), and that holding fringe factual beliefs can alienate people from their loved ones or keep them invested in causes that exploit their gullibility.

However, misinformation research over the past few decades has focused on a much narrower set of pathologies (Adams et al., 2023), with two particularly salient concerns. The first argument, often attributed to Arendt (1967), is premised on the idea that all functioning democracies depend on a baseline of shared factual agreement. If citizens are not merely ignorant but confidently misinformed about verifiable facts (Kuklinski et al., 2000), misinformation spells distrust, polarisation and political unrest. These undermine democratic decision-making and the public response to urgent issues such as pandemics and climate change. A more pessimistic argument suggests that widespread misinformation erodes our collective ability to identify and value truth. As information environments become saturated with misinformation, our social experience is expected to literally split into parallel realities (Lewandowsky et al., 2017; Sunstein, 2017). Pushed to its logical conclusion, this line of reasoning culminates in a kind of 'epistemic apocalypse' in which no fact remains beyond dispute (Habgood-Coote, 2023).

Recently, both of these accounts have been subjected to increasing scrutiny, as new evidence has repeatedly shown that misinformation is neither as widespread nor as influential as is commonly believed (Acerbi et al., 2022; Allen et al., 2020; Budak et al., 2024). This has, in turn, reignited the debate over the urgency of tackling epistemic harms and the most effective means of doing so. While many current approaches, such as fact-checking, information literacy initiatives, and behavioural nudges towards accurate information, have been shown to successfully mitigate the effects of misinformation in experimental settings (Kozyreva et al., 2024), it is not a given that alternative approaches, such as promoting reliable information and strengthening trust in credible sources, would be any less effective (Altay, 2026a). Others have echoed this criticism, pointing out that many of the current recommendations place too much emphasis on technical interventions (Habgood-Coote, 2019; Tromble & McGregor, 2019) and an overly rational view of human beings (Munn, 2024), while overlooking the broader social structures that generate and perpetuate misleading narratives (Jungherr & Rauchfleisch, 2024; Kuo & Marwick, 2021). Still others caution that framing misinformation as a 'solvable' problem risks overlooking its complexity (Jack, 2019) and undermining the role of factual disagreement in societies governed by democratic norms (Freiling et al., 2023). This does not mean that nothing should be done about misinformation; it is simply difficult to imagine a time when misleading and deceptive information will be fully eradicated, even with sweeping measures. Therefore, misinformation is likely to persist as a communicative act that societies must learn to live with (Andersen & Søe, 2020).

1.1 Research problem and research question

The cracks in the normative assumptions of misinformation research present an interesting opportunity to shift the focus from the increasingly controversial question of how misinformation can best be mitigated, to how it is in fact mitigated. That is, while an impressive body of work has prescribed remedies that *should* work against misinformation, very few studies have examined what is *actually* done to counter it in the real world. This gap exists partly by design: correction research remains dominated by experimental studies that, quite understandably, focus on questions such as 'What works and to Whom' (Walter et al., 2020, title) but are less inclined to tackle 'Who', 'Why' and 'How'. Nevertheless, the growing availability of observable user activity on social media is making it increasingly feasible to address these latter questions directly. Indeed, there is a growing awareness that various actors attempt to counter misinformation outside of controlled research settings. However, our understanding of what these everyday corrective interactions entail, how prevalent they are, and what motivates them is still limited.

This project aims to reveal the everyday process of correcting misinformation visible as it happens online. Rather than treating correction primarily as a specialised technique, typically understood as the presentation of information designed to rebut an inaccurate claim or misperception (Chan et al., 2017; Lewandowsky et al., 2012; Vraga & Bode, 2020a), I am interested in how people come to counter misinformation, what they use corrections for and what this communicative act comprises. I therefore look beyond addressing factual errors within specific institutional frameworks, such as journalism (Graves, 2016), to identify actors who do not necessarily consider themselves to be 'fact-checkers', yet who occasionally assume a similar role in ameliorating disputed realities. To develop this approach, I focus on social media platforms and their users and ask *how these actors understand and carry out actions to counter misinformation in everyday contexts*, and *what type of 'corrective agency' emerges from these efforts*.

In addition to corrections, this dissertation is informed by two key concepts: social media and misinformation. Within the broader 'platform ecosystem' (van Dijck et al., 2018, p. 4)[2], social media are communicative spaces that 'host, organize, and circulate users' shared content or social interactions' (Gillespie, 2018). For this dissertation, these platforms notably facilitate multiple flows of communication (Thorson & Wells, 2016). On the one hand, social media are highly fine-tuned data infrastructures where online expressions are ultimately governed by the whims of

[2] For the sake of brevity, I use the term 'platform' throughout this thesis to refer to both individual online services (e.g. Facebook) and the entities that operate them (Meta Platforms, Inc.). Despite the flaws and strategic connotations of the 'platform' metaphor, I agree with Gorwa (2019b) that the term has become too well-established to be dispensed entirely.

extremely wealthy US corporations (Klonick, 2018; Napoli & Caplan, 2017; Plantin et al., 2018). Conversely, the experiences and cultures that emerge on social media are actively shaped by how people personally curate their own information (Wu-Ouyang, 2026) and develop new media habits through interaction with other members of their extended social networks (Bucher & Helmond, 2018; Gibbs et al., 2015; Messing & Westwood, 2014). The dynamics of sharing and addressing misinformation on social media are therefore expected to involve various interacting actors and incentives that can either complement or conflict with each other.

In my view, misinformation, is best understood in terms of the 'epistemically suspect' (Uscinski et al., 2024) aspects of public communication and their growing significance in the digital media environment. While untruthful communication is an age-old phenomenon, 'misinformation' as it is currently understood, is a relatively recent term that has been established as an umbrella designation for various types of misleading and deceptive communication online. Consequentely, academics still disagree on the parameters of this 'misinformation zoo', even though many name fake news, conspiracy theories, rumours and clickbait as some of its most paradigmatic examples (Altay, Berriche, Heuer, et al., 2023). Similarly, and perhaps owing to the distinct academic traditions associated with these terms, attempts to formulate a unifying definition that encompasses all of them have proved contentious. I will next turn my attention to this issue.

2 THREE PERSPECTIVES ON MISINFORMATION

To understand the role of corrections in tackling misinformation, it is helpful to begin by briefly discussing what misinformation is. Unfortunately, despite the extensive literature that has emerged around this question, a clear answer remains elusive (Camargo & Simon, 2022; Uscinski et al., 2024; Weeks & Gil de Zúñiga, 2021). This is perhaps unsurprising, given that the subject of misleading online communication has proven to be far more complex and interconnected than many initially thought (Kupferschmidt, 2024). In this section, I attempt to make sense of this sprawling research area by focusing on three prominent perspectives: misinformation as an 'information disorder', misinformation as a psychological phenomenon, and misinformation as a discourse.

2.1 Misinformation as an 'information disorder'

The term *misinformation* began to gain academic traction in 2017 with the publication of two influential works: *Information Disorder* by Claire Wardle and Hossein Derakhshan, and *Lexicon of Lies* by Caroline Jack. Both works sought to clarify public debates surrounding the newly popularised neologism 'fake news', arguing that it had quickly become an imprecise label for phenomena that were not necessarily 'fake' nor 'news' (see also Habgood-Coote, 2019). Though the concept of misinformation had earlier roots – the *Oxford English Dictionary* dates its meaning as 'wrong or misleading information' back to the sixteenth century – scholars of the time had made very little use of it outside the specialised study of false memories (e.g. Johnson & Seifert, 1994). It was only in the early 2010s that public health researchers such as Kata (2010) and Betsch and Sachse (2012) began to use the term to describe misleading online content. The lasting contribution of Wardle, Derakhshan, and Jack was therefore not so much in coining a term, which already featured in a range of scattered and often implicit academic definitions (e.g. Fox, 1983; Hochschild & Einstein, 2015; Karlova & Fisher, 2013; Smith et al., 2011; Southwell & Thorson, 2015; Stahl, 2006), but to devise a coherent framework for understanding misinformation as a distinct *online phenomenon*. I group these insights under the following axioms:

- *Quality*. Misinformation is any information that lacks a factual basis. Although the extent to which an assertion is 'inaccurate, incorrect, or misleading' (Jack, 2017, p. 2) may vary, this quality is nonetheless assumed to be ontological and independent of the observer.
- *Focus*. Beyond its quality, misinformation is not analytically limited to any single format or source (fake news, trolls or bots), narrative genre (conspiracy theories, rumours, myths), modality (texts, images, videos) or context (such as tweets, memes or websites).
- *Intentionality*. Unlike straight-up lying or 'bullshitting' (Frankfurt, 2005), misinformation is communicated without the intention (or disregard) of misleading audiences (e.g. when a news outlet publishes a story containing factual errors, Jack, 2017). 'Disinformation', in contrast, describes a subset of inaccurate or distorted information that is also intended to deceive audiences regardless of their well-being (Wardle & Derakshan, 2017). For the sake of completeness, Wardle and Derakshan (2017) also propose a third category of 'malinformation', which contains all information that is factual, yet intentionally harmful. This could plausible include some types of propaganda (Jack, 2017).
- *Degree*. Not all misinformation is created equal. Its epistemic quality and deceptive aims can vary independently of each other, ranging from satire (which is mostly accurate and relatively harmless deception) to deliberate fabrications or the misrepresentation of accurate facts taken out of context (Wardle & Derakshan, 2017).
- *Pathology*. Misinformation has an intrinsic pathological nature, independent of the express intention to deceive. It is 'problematically inaccurate information' (Jack, 2017, p. 1), in a special sense which presumably distinguishes it from other types of fictional information. While the cause of this pathology is not explicitly explored, it is heavily implied that it is tied to the negative social influences that misinformation exerts on individuals and collectives (e.g. 'the complexity and scale of information pollution in our digitally-connected world presents an unprecedented challenge', Wardle & Derakshan, 2017, p. 10).

This 'information disorder' framework has proved tremendously influential among and beyond more academic attempts to conceptualise the source of informational threats that emerged around 2016 (e.g. Benkler et al., 2018; Bennett & Livingston, 2018; Lazer et al., 2018; Tandoc, Lim, et al., 2018; Woolley & Howard, 2018). This is partly due to its flexibility in explaining the production, dissemination and consumption of misleading content in a variety of contexts and domains: whether political (Weeks & Gil de Zúñiga, 2021), scientific (Southwell et al., 2022) or health-related (Nan et al., 2023); visual (Dan et al., 2021) or multimodal (Molina, 2025); participatory (Starbird et al., 2019) or orchestrated (Broniatowski et al., 2018). Conceptually, the framework has also aligned well with a number of important theoretical observations, ranging from high-choice information environment (Prior, 2007) and audience fragmentation (Sunstein, 2017) to algorithmic curation (Thorson & Wells, 2016) and the transformation of media's gatekeeping role (Bennett & Pfetsch, 2018). Its analytical tools have also been widely adaptod by large-scale studies using digital materials and computational techniques (Venturini & Rogers, 2019) to observe and model information flows in online networks (Del

Vicario et al., 2016). In short, it has provided scholars from various disciplines with the fundamental vocabulary to begin asking questions such as how disinformation campaigns operate (Woolley & Howard, 2018), who shares fake news online (Chadwick et al., 2018), and how prevalent misinformation is on social media as a whole (Guess et al., 2019). This legacy is perhaps best encapsulated by one of the most often quoted statements in recent social science: 'Falsehood diffuses significantly farther, faster, deeper, and more broadly than the truth' (Vosoughi et al., 2018, p. 5)[3].

Notwithstanding these successes, the body of literature based on the work of Wardle, Derakhshan, and Jack still has its shortcomings (Altay, Berriche, & Acerbi, 2023; Broda & Strömbäck, 2024; Camargo & Simon, 2022; Pasquetto et al., 2024), a fact that its authors also fully acknowledge (Wardle, 2023). For example, although most contemporary research recognises that misinformation is not confined to a specific medium, format or modality, surprisingly little research has ventured outside its most obvious and egregious manifestations (Steenbuch Traberg, 2022). Likewise, relatively few studies have examined the broader narratives underlying individual pieces of false information (Kuo & Marwick, 2021) or attempted to situate misinformation within more established theoretical traditions (Harsin, 2024). Conversely, a burgeoning literature on 'fake news' suggests that efforts to abandon the term have been unsuccessful, regardless of conceptual (Weeks & Gil de Zúñiga, 2021) and methodological limitations (Rogers, 2020), as well as its seemingly low prevalence (Acerbi et al., 2022; Allen et al., 2020). At the same time, the term 'disinformation' has taken on a life of its own, even as rigorous efforts to define it as a distinct concept remain rare (but see Chadwick & Stanyer, 2022; Hameleers, 2023a). More fundamentally, there is still no definitive answer to how we know that misinformation is, in fact, false or misleading (Uscinski et al., 2024). One of the best attempts to address this shortcoming is to define misinformation narrowly as 'information considered incorrect based on the best available evidence from relevant experts at the time' (Vraga & Bode, 2020b, p. 138). However, this definition raises the obvious challenge of identifying such experts and placing trust in their infallibility (see Krause et al., 2022). In practice, then, misinformation appears to remain a category with no clear epistemic baseline beyond more or less convincing, context-specific operationalisations (Pasquetto et al., 2024).

[3] The influence of this one quote cannot be overstated. To date, the article by Vosoughi, Roy, and Aral has been cited, conservatively, over 7,908 times. Even more impressively, it has achieved the highest Altmetric Attention Score – a metric that aggregates the volume of mentions in both legacy and social media – of any social science paper ever indexed by the service. In their public statements, the authors have since expressed some regret about the scope of their proclamation. Rather than measuring all online 'falsehoods', the study was limited to those Twitter shares of news articles that had been contested by leading US-based fact-checking organisations.

2.2 Misinformation as a psychological phenomenon

While the information disorder framework is based on the pathology of misinformation, it is rather vague about the mechanisms by which misinformation causes problems in the information environment. Although an influx of low-quality information can dilute the salience of hard facts, social media is already full of opinionated content that is not strictly factual either. It is also inappropriate to group misinformation with abusive online expressions, such as harassment and hate speech, as these directly affront basic human dignity and the sense of security. Misinformation is sometimes inherently linked to oppressive narratives (see Kuo & Marwick, 2021) and can be used to falsely accuse innocent parties of events beyond their control (Uscinski & Enders, 2023). Despite this, misinformation may also offer genuinely positive social uses. For instance, certain untrue statements may help people make sense of distressing events (DiFonzo & Bordia, 2007; Douglas et al., 2019), express justified suspicions (Gagliardone et al., 2021), and serve as a form of self-expression (Sihvonen, 2025). In fact, some would argue that misinformation can be informative, despite being inherently irrational (Karlova & Fisher, 2013).

One convincing argument for clarifying misinformation's harmful effects comes from the field of psychology, which studies how false claims promote false beliefs, or how 'misinformation misinforms' (Lewandowsky et al., 2017, p. 355). Building on the argument that misinformation spreads 'under the guise of informative content' (Guess & Lyons, 2020, p. 10), this suggests that some individuals mistakenly accept it as *bona fide* facts (Stahl, 2006). In aggregate, such misbeliefs at the individual level provide a plausible mechanism for linking misinformation to broader downstream effects, such as declining trust in public institutions (Bennett & Livingston, 2018). It is therefore worthwhile to understand socio-psychological factors that drive these *misperceptions*, or 'cases in which people's beliefs about factual matters are not supported by clear evidence and expert opinion' (Nyhan & Reifler, 2010, p. 306).[4]

The psychology of misinformation is studied within several theoretical traditions (see Ecker et al., 2022; and Pennycook & Rand, 2021 for recent reviews), which are inspired, broadly speaking, by three key observations about the

[4] I should note that disinformation, defined as 'a context-bound deliberate act for which actors *covertly deceive* recipients by de-contextualizing, manipulating or fabricating information *to maximize utility* with the (targeted) outcome of misleading recipients' (Hameleers, 2023a, p. 6, emphasis added), may present additional problems. We can condemn deceit and manipulation whether or not they achieve their intended goals. Furthermore, as Hameleers et al. (e.g. 2020, 2022) compellingly demonstrate, the experience of being deliberately deceived by disinformation is worse than accidentally consuming misinformation. This perception alone may contribute to some unique downstream effects (see also McKay & Tenove, 2021). Regarding the argument that disinformation also inflicts particular harms on society, I am inclined to attribute many of its issues instead to propaganda, which employs deception as one means of ruthlessly shaping public opinion in one's favour (but see Chadwick & Stanyer, 2022).

limitations of human memory and cognition. The first of these is *truth bias*. If people were intuitively able to detect dishonesty, misinformation would not be a major concern. Unfortunately, doubting information does not come naturally. The perception that others are honest in their communications appears to be not only a normative ideal, defended by the likes of Grice (1975) and Habermas (1984), but also an empirically supported fact (Brashier & Marsh, 2020; Levine, 2014). All things being equal, people tend to accept incoming information at face value, and only later engage in secondary processes to verify its authenticity (Lewandowsky et al., 2012). Consequently, substantial evidence suggests that exposing individuals to misinformation reduces the accuracy of their beliefs (e.g. Chan & Albarracín, 2023; Ecker et al., 2022; Li & Yang, 2025), also in scenarios that mimic social media feeds.

Misinformation can also flourish due to directionally *motivated reasoning*, whereby personal and group goals are prioritised over accuracy (Kunda, 1990; Taber & Lodge, 2006). Rather than considering all incoming information equally, people tend to weigh new evidence based on their existing beliefs. Examples of this include confirmation and disconfirmation biases, whereby people seek (or dismiss) information depending on whether it supports (or challenges) their prior views (Taber & Lodge, 2006). It is particularly concerning under conditions of heightened ideological polarisation and partisanship (see Altay, Berriche, Heuer, et al., 2023), that people may consider certain (mis)information to be orthodox simply because of their social, religious or political affiliations (e.g. because it is endorsed by elite members of their in-groups, Flynn et al., 2017) is particularly concerning under conditions of heightened ideological polarisation and partisanship. In principle, misinformation that is relevant to one's in-group may therefore gain particular traction on social media, where users can, to some extent, flock among like-minded individuals (Nyhan et al., 2023) and interact with algorithmic recommendations to further increases their exposure to pro-attitudinal content (Shin, 2020).[5]

Finally, recent studies based on the dual-process models of reasoning (Evans, 2003) suggest that people may fall prey to misinformation not only due to their partisanship, but simply because of a *lack of effort* (Pennycook & Rand, 2021). In other words, relying on quick but lazy intuitive thinking makes people more susceptible to false beliefs than slow, analytical reasoning does. Spotting misinformation amid other online content requires deliberate effort, which may be lacking when people casually browse social media (Pennycook & Rand, 2019). Compounding this issue, people tend to navigate online spaces by inferring credibility based on a number of cognitive heuristics, or mental shortcuts (Metzger & Flanagin, 2013). While it is possible to use these indicators to reliably judge the

[5] Selective exposure and social homophily are well-documented phenomena, but neither necessarily implies the existence of so-called 'filter bubbles' or 'echo chambers' which have recently been disputed (Bruns, 2019; Dubois & Blank, 2018) and complicated (Nyhan et al., 2023; Törnberg, 2022).

trustworthiness of an online source, social media platforms have historically done little to help their users to separate credible sources from low-quality content (Bimber & Gil de Zúñiga, 2020). Consequently, it seems all too easy to fall for misinformation simply because it is popular, endorsed by celebrities or carries other social cues that people mistakenly perceive as a proxy for factual accuracy (Molina et al., 2022).

Although psychological research has convincingly established that false claims can reduce the accuracy of beliefs, it has had less success in conclusively demonstrating other real-world effects (Lee & Albarracín, 2026; Osman, 2025; Uscinski et al., 2022). While there is no shortage of studies that posit a causal connection between misinformation and various behavioural outcomes, research directly testing this link remains rare (e.g. Murphy et al., 2023)[6]. Moreover, as Li and Yang (2025) point out, some of the most compelling evidence connecting misinformation to later behaviours (e.g. Tan et al., 2015) still relies on self-reported measures. On the other hand, the few studies that have directly examined the relationship between misinformation exposure and real-world outcomes, such as vote choice (Eady et al., 2023) and vaccination intentions (Porter et al., 2022), have found no clear effect. This does not imply that such effects are implausible (Tay et al., 2024), particularly given the extensive correlational evidence supporting them (Lorenz-Spreen et al., 2023). However, as it stands, the observed association may also run in the opposite direction, whereby individuals who are predisposed to non-normative behaviours may simply be more likely to seek out misinformation as well (Budak et al., 2024; Littrell et al., 2023; Osman, 2025; Uscinski et al., 2022).

2.3 Misinformation as a discourse

The most recent, and perhaps still the least organised, research perspective moves beyond treating misinformation as a stable analytical concept; instead, it views misinformation as a socially constructed label that can reflect and shape public understanding of problematic communication. As Farkas and Schou (2018) observe, terms such as 'fake news' perform important discursive work in the service of diverse strategic objectives. Populist political actors, for example, tend to exploit this label to undermine established media companies rhetorically and legislatively (Hameleers & Minihold, 2022; Kluknavská & Eisele, 2023). The very same legacy media organisations may in turn incorporate similar terms into their own meta-discourses, using misinformation as a useful foil that distinguishes journalism as a principled and accountable truth-seeking institution (Carlson, 2020).

[6] Admittedly, devising such research designs is extremely difficult. However, as Murphy et al. (2023) explicitly point out, such methodological challenges hardly justify remarks about behavioural outcomes that were never measured.

Emerging research on media discourses about misinformation has found patterns of emphasising its threats far more than available evidence suggests (Bratich, 2020; Carlson, 2020), a fact acknowledged by some journalists themselves (Farkas & Schousboe, 2024). Farkas (2023) identifies several ways in which such meanings are constructed in legacy news. One of these involves actively anticipating and preparing for what misinformation *might* do: the 'fearful futures' (Wahl-Jorgensen & Carlson, 2021) of severe societal disruptions that seem imminent regardless of the evidence. Misinformation is also typically constructed as a radically 'foreign' influence, or 'incursion in the media ecology' (Cover et al., 2023, p. 218), contrasting with the well-meaning domestic actors. The threat of misinformation is also frequently seen as inseparable from high levels of technological sophistication, which can, in principle, victimise any internet user (also Habgood-Coote, 2023). Finally, misinformation is increasingly portrayed as a national security threat, prompting calls for national security solutions (surveillance, military intelligence), even if this could lead to legitimate factual disagreements being discouraged or even penalised (also Monsees, 2020).

The research community also contributes to dominant misinformation discourses. This is particularly evident in studies that use and maintain popular metaphors such as 'infodemics', implicitly or explicitly suggesting that misinformation behaves like and can be addressed as a communicable disease[7]. An obvious problem with this shorthand is that it oversimplifies or entirely neglects some of the communicative and psychological dynamics that I have introduced in the previous sections (see also Anderson, 2021; Simon & Camargo, 2021). Similar criticism has been levelled at related buzzwords such as 'post-truth'. Much scholarship has rightfully questioned the assumption that the political events and technological developments witnessed around 2016 marked a decisive shift away from fact-based public discourse (e.g. Habgood-Coote, 2019) given that the American 'epistemic polarisation' (Friedman, 2023), fragmented epistemic communities (Mejia et al., 2018), and political weaponisation of truth claims (also Arendt, 1967) all date back decades. Framing the present moment as an 'information disorder' also has its flaws. Nature-related metaphors of order and pollution imply that there is, in fact, an objective and balanced 'proper order' to the functioning of an information environment, and that perhaps the current 'crisis is to be resolved by a reinvigoration of enlightenment epistemic norms, which principally manifests in a defence of establishment knowledge-generating

[7] To the best of my knowledge, research that treats misinformation as an 'infodemic' in a literal sense is still limited to the field of public health (e.g. Briand et al., 2021) whereas the term tends to be used more often rhetorically than analytically. However, there are some notable exceptions. For example,. For example, Cinelli et al. (2020) developed an epidemiological model to calculate the 'basic reproduction number R0' for the spread of unreliable information about COVID-19 on various social media platforms. According to their model, '$R_0 > 1$ indicates the emergence of an infodemic' (p. 3).

institutions (major newspapers, broadcasting corporations, fact-checkers)' (Habgood-Coote, 2023, p. 16).

The way misinformation is framed in public discourse matters, as these perceptions can measurably influence how people approach digital information in general (Hameleers, 2025). As concerns about misinformation have spread globally, including to contexts where exposure appears relatively low (Altay, 2025; Knuutila et al., 2022), some of their unintended consequences are also coming into focus. Those who perceive themselves as having been misled may respond with coping strategies that fuel public distrust, disengagement, and news avoidance (Hameleers et al., 2024). Accordingly, alarmist public discourse around misinformation has also been linked to cynicism and declining satisfaction with democracy, particularly in the United States (Jones-Jang et al., 2021; Jungherr & Rauchfleisch, 2024). This poses a distinct challenge for interventions that frequently emphasise the threat of misinformation, and thereby risk exacerbating these negative spillover effects (Hameleers, 2023b; Hoes et al., 2024).

3 CONCEPTUAL FRAMEWORK

> *Correction of misinformation likely entails substantial effort not only in terms of message construction but also simply in terms of ensuring adequate message exposure. (…) Any campaign to correct misinformation, even if poised to succeed in terms of message content, likely requires considerable planning to reach all relevant members of a population in question. Misinformation is thus challenging to remedy not only because of how humans process information but also because of the complexity of audience reach in contemporary society.*
>
> Southwell and Thorson (2015, p. 590)

The correction of misinformation has fascinated the social sciences for decades. Nevertheless, a clear conceptual framework for describing corrections 'in the wild' has remained elusive. In this section, I review key literature on countering misinformation, ranging from experimental work on the effectiveness of correcting misperceptions to observational mass media research on practices such as post-publication error correction and fact-checking. Building on this work, I proceed to sketch a holistic and interdisciplinary understanding of corrections as communicative acts comprising distinct elements. Finally, I consider the two main groups of individuals involved in correcting misinformation on social media and summarise how existing research has explained their ability to identify and challenge it.

3.1 Characterising corrections

The idea that false beliefs can be refuted by 'setting the record straight' can be traced back to Allport and Lepkin's (1945) seminal work on wartime rumours. Such attempts to counter inaccurate information with accurate information are nowadays commonly referred to as corrections, consistent with the everyday usage of 'correcting someone' to mean pointing out a factual error in their reasoning (Takahashi & Beebe, 1993). Yet, correction remains a surprisingly theoretically underdeveloped concept. This is perhaps most evident in the psychological literature, the discipline most closely associated with correction research, where the key object of study almost never defined, and correction is typically conflated with other, analytical concepts such as retraction (e.g. Ecker et al., 2017), refutation(Amazeen & Krishna, 2024) and debiasing (e.g. Schwarz et al., 2007) or

Table 1 Definitions of correction and related terms in psychological literature.

Main functions	Example definitions or assumption
To expose information as misinformation	'People usually cannot recognize that a piece of information is incorrect until they receive a correction or retraction' (Lewandowsky et al., 2012, p. 111)
	'[Debunking is] presenting a corrective message that establishes that the prior message was misinformation' (Chan et al., 2017, p. 1532)
	'Traditionally, to qualify as a refutation, messages must expose misinformation and provide explanations as to why it is inaccurate' (Amazeen & Krishna, 2024, p. 1)
To contradict misinformation	'[Correction involves] the presentation of information designed to rebut an inaccurate claim or a misperception' (Vraga & Bode, 2020a, p. S278)
	'[Mis]information might be subsequently challenged by a correction, which can take the form of a retraction (a simple negation, such as 'it is not true that arson caused the fire') or a refutation (a more detailed correction that explains why the misinformation was false).' (Ecker et al., 2022, p. 15)
	'We regard factual corrections as interventions produced by journalistic and/or media organizations that challenge misinformation by presenting countervailing, verifiable evidence and arguments' (Porter et al., 2023, p. 2)
To counteract the effects of misinformation	'Ideally, one might want a correction to diminish or eliminate subsequent effects from such misinformation so that one understands and reasons about it like those not exposed to the misinformation' (Johnson & Seifert, 1994, p. 1420)

even journalistic fact-checking (Walter & Salovich, 2021). Of the few definitions that have been offered (see Table 1), 'correction' is perhaps most readily understood as 'the presentation of information designed to rebut an inaccurate claim or a misperception' (Vraga & Bode, 2020a, p. S278). However, different authors also tend to emphasise aspects such as exposing or unmasking misinformation (Lewandowsky et al., 2012), contradicting it (Porter et al., 2023) or educating audiences so that they can reject misinformation themselves (Johnson & Seifert, 1994). The qualifier *corrective* also has several meanings. It is sometimes used to describe efforts to communicate a correction, such as corrective advertising (Aikin et al., 2015), but sometimes refers to intended or observed impact (the corrective effect). Common synonyms such as *debunking* are not much more informative. Some authors associate it with the length and detail of a corrective message (versus flat 'rebuttals', Lewandowsky et al., 2012), while others define the in terms of timing (versus pre-emptive 'prebunking', Brashier et al., 2021) and still others with specific communicative strategies (versus the practice of 'fact-checking', Roozenbeek et al., 2023).

Regardless of this ambiguity, experimental studies tend to approach correction fairly similarly. Conceptually, it is common to differentiate between such factors as correction source (who issues the correction), recipient (whose belief change is tracked), message (what the correction is composed of) and mechanism (how is it delivered, see Chan & Albarracín, 2023; Walter, Brooks, et al., 2021; Walter &

Murphy, 2018). A common laboratory stimulus[8], for example, might present a carefully worded response to misinformation within a journalistic fact-checking article to measure whether participants accept the correction (by displaying a corrective effect) or reject it (retaining their baseline misbeliefs). These factors are then systematically manipulated to address the central question of psychological correction research: what types of correction are most effective in improving belief accuracy?

While correction literature lacks a useful description for identifying a correction outside of controlled conditions, it is very instructive in defining the components of successful – meaning accuracy-boosting – corrections. Much of this work stems from decades of pessimistic evidence suggesting that misinformation is particularly difficult to remove from memory, a phenomenon commonly known as the continued influence effect (Lewandowsky et al., 2012; Walter & Tukachinsky, 2020). Adding to these concerns is the possibility that corrections may perversely 'backfire' and reinforce rather than reduce false beliefs, inadvertently spreading the original misinformation to new audiences (Schwarz et al., 2007; but see Ecker et al., 2020), or provoking resistance when correction challenges the worldviews of their recipients (Nyhan & Reifler, 2010; but see Wood & Porter, 2019; Nyhan, 2021). In response, considerable effort has been devoted to identifying the social and cognitive barriers that impede belief updating and to developing strategies for overcoming these barriers (see Ecker et al., 2022 for a review).

Since the mid-2010s, a series of meta-analyses have addressed some of these concerns. According to the current state-of-the-art research, corrections are at least moderately effective in improving belief accuracy across diverse experimental settings, and do not pose a significant risk of backfiring (Chan et al., 2017; Janmohamed et al., 2021; Walter et al., 2020; Walter & Murphy, 2018). However, this work has also established that correction effectiveness varies considerably depending on the source, message design and context. This has prompted continued efforts to develop more 'enhanced refutations' (MacFarlane et al., 2021). While the 'best correction practices' are still evolving, the most recent guidelines emphasise factors such as clarity, factual elaboration, and an affirmative tone. Research also agrees on designing corrective messages according to specific 'debunking strategies', such as providing a plausible alternative explanation to

[8] Ecker, Hogan, and Lewandowsky (2017) describe how researchers have 'traditionally used a paradigm in which individuals read a (fictional) news report or scenario that includes a piece of critical information that subsequently is or is not retracted' (p. 108). In other settings, correction may be as simple as having a member of the research team present to tell research participant to disregard initial misinformation as inaccurate (Chan & Albarracín, 2023, p. 8). Nyhan and Reifler (2010) critique this design choice as apparently unrealistic, writing that 'such authoritative statements of fact (such as those provided by a survey interviewer to a subject) are not reflective of how citizens typically receive information' as 'people typically receive corrective information within "objective" news reports pitting two sides of an argument against each other, which is significantly more ambiguous than receiving a correct answer from an omniscient source' (p. 306).

refute misinformation (Bode et al., 2024; Ecker et al., 2022; Kozyreva et al., 2024; Prike & Ecker, 2023).

Although recent evidence is encouraging, expectations of how well corrections work, even under the best of circumstances, remain modest. For example, corrections proven to be effective against fictional misinformation in controlled settings, seem far less impressive when studies test them against misinformation about real issues in realistic scenarios (Chan & Albarracín, 2023; Walter et al., 2020). Corrections are also less effective when the issue is politically polarised, possibly because recipients actively counter information that threatens their identity (Chan & Albarracín, 2023). As with misinformation, online users do not accept or share corrections impartially, but view them through the lens of their ideological biases (Amazeen et al., 2019). Lastly, there is growing recognition that even the most effective interventions only have a short-lived impact, with improvements to belief accuracy deteriorating significantly within weeks (Carey et al., 2022) and suggesting that corrections should be repeated at regular intervals (Craig & Vijaykumar, 2023).

3.1.1 Post-publication error-correction

While the psychological research is interested in the persuasive influence of corrections, mass media studies have traditionally viewed them primarily as a means of acknowledging and addressing errors. As fallible institutions, media companies (and their advertisers (Aikin et al., 2015) tend to issue corrections 'following mistakes and errors' to 'declare their ambition for high standards and reaffirm internal processes of quality which may prevent the recurrence of flaws' (Kampf & Daskal, 2014, p. 165). In other words, corrections aim to 'amend the record when [factual] errors occur' (Hettinga, 2013, p. 30). In journalism, dedicated correction pages and notices that acknowledge and apologise for found mistakes function therefore as 'accountability products' that signal professional principles of responsibility, transparency, and trustworthiness (Kampf & Daskal, 2014).

The normative ideals of self-correction, as codified in numerous journalistic codes of ethics (Appelman & Hettinga, 2021), establish rather clear-cut requirements regarding correction *speed* ('Essentially incorrect information must be corrected without delay', Council for Mass Media in Finland, 2024 as cited in Appelman & Hettinga, 2021) and *impact*, suggesting that they should be 'presented in the way most likely *to offset whatever damage was done by the error*' (Seib & Fitzpatrick, 1997 in Bugeja, 2007, p. 49, emphasis added). As Bugeja (2007) argues, this process invokes the concept of 'making whole', primarily to *restore* trust and integrity (i.e. the expectation that the outlet will improve its quality in future), but also to *mitigate* the damage caused by the publication of factually false, misleading or incomplete information. While journalism studies largely remain silent on what

such damage might entail, advertising research provides some concrete examples. For instance, Aikin et al. (2015), advocate corrections as a way to retract any false beliefs generated from a misleading account and substitute them with more accurate information.

Studies have shown that admitting errors in reporting helps media outlets to maintain transparency without causing significant damage to its credibility (Appelman & Hettinga, 2015). However, newsrooms do not necessarily have consistent procedures for identifying and correcting mistakes. This leaves journalists to decide for themselves whether they are willing to acknowledge errors, despite the potential professional consequences (Hettinga, 2013). Such organisational dynamics may help explain why corrections are relatively rare and usually limited to addressing trivial mistakes (Barkin & Levy, 1983; Hettinga & Appelman, 2014). Consequently, these normative and practical aspects seem to overshadow the open question on whether the purpose of a correction is also to inform readers effectively and repair the harm that inaccurate reporting has caused to them, as Silverman (2011, para. 6) notes:

> The point of an online corrections page is to have a centralized place where readers can see the latest mistakes and corrections. It gives them the opportunity to discover if a recent article they read, or reporting they heard or saw, has been updated or corrected. [...] A dedicated page makes corrections more visible and accessible, and it increases the likelihood that people will receive the corrected information. After all, that's the point of making correction in the first place. Yet correction pages are the exception, not the rule.

Similar findings emerge from research on corrections in scientific publishing. In keeping with the ideal of self-correcting science (Popper, 1965), most academic journals apply editorial policies that allow them to revisit and potentially 'purge their record' of evidence that is no longer considered 'valid as a source of knowledge' (Kühberger et al., 2022, p. 1; Resnik et al., 2015). Such errors are typically reported by aggrieved colleagues and remorseful authors (Williams & Wager, 2013), or increasingly discovered by volunteer 'science sleuths' (Garisto, 2025) and online communities who 'peer-review' published work amongst themselves (Mandavilli, 2011; Peng et al., 2022). Of the various ways to amend published research, retractions are considered the ultimate self-correcting mechanism as they enable editors – although hardly consistently (Frampton et al., 2021) – to deface, remove or update the original work to reflect its issues[9]. Such interventions are usually accompanied by retraction notices that further elaborate on the reasons and justifications for the retraction, such as whether the correction is due to minor

[9] Electronic journals and repositories offer even more possibilities to indicate retractions, such as digital banners, warning labels and updated headings (e.g. 'RETRACTED: [original title]'). When embedded in publication metadata, these signals can also travel with the article across platforms, including some reference management software such as Zotero and social media link previews.

errors, such as duplicate publications or improper and questionable research practices, which are the leading cause of retractions, historically (Brainard, 2018; Fang et al., 2012; Grieneisen & Zhang, 2012). As such, they are routinely hailed as *potential warnings* to readers of scientific literature, as well as being used as a means of accountability and maintaining consistent quality.

The extent to which retractions stymie the circulation of demonstrably flawed science is unclear. Although self-correction is not uncommon, concerns persist that editorial action is still used too sparingly (Besançon et al., 2022; Brainard, 2018; Grieneisen & Zhang, 2012), particularly as scientific misconduct gets increasingly systematic and sophisticated (e.g. Richardson et al., 2025). The reasons remain poorly understood, not least because the retraction process appears to vary considerably from one journal to another. Editors may hesitate to retract papers due to concerns about stigmatising authors (especially junior scholars whose careers can be seriously affected by retraction, Memon et al., 2025), or because they wish to avoid the disputes that tend to accompany retractions (Brainard, 2018; Marcus & Oransky, 2014; Williams & Wager, 2013). Another common critique is that the investigations move slowly, with serious problems sometimes remaining unaddressed for years (Fang et al., 2012; Kühberger et al., 2022; Trikalinos et al., 2008). At the same time, numerous bibliometric analyses show that retracted articles continue to receive inappropriately favourable citations for years, albeit at lower rates (Bar-Ilan & Halevi, 2017; Korpela, 2010). It is therefore unsurprising that the 'scarlet letter' of retraction often fails to inform wider audiences. This seems to be true for the members of the public who struggle to make sense of retractions (Greitemeyer, 2014), science journalists who seldom follow up on stories based on flawed research (Valinciute & Halffman, 2025), and even generative AI tools that appear unable identify retracted work (Thelwall et al., 2025).

3.1.2 Fact-checking

In addition to addressing their own factual errors, over the past few decades journalists have established a vibrant fact-checking movement dedicated to correcting public statements made by other individuals and institutions. Following Vinhas and Bastos' (2022) broad definition, professional fact-checking involves 'publishing regular assessments about the validity of questionable claims made by a wide array of sources, including governments, politicians, institutions, news organizations, and social media users' (p. 499). Fact-checking emerged from US newsroom practices in the 2000s (Graves et al., 2016) as an attempt to promote a more epistemically principled alternative to the customary 'he said, she said' style of reporting and thereby maintain a more fact-driven public discourse (Graves & Amazeen, 2019). Today, these ideals are promoted by a global network of independent fact-checking organisations that increasingly define themselves in

contrast to the perceived shortcomings of traditional media outlets (Nieminen & Rapeli, 2019; Singer, 2021). In practice, this ethos is supported by specialised methodologies that seek to establish the accuracy of socially or politically salient claims (Walter et al., 2020), for example by tracing their origins, evaluating source credibility and verifying statements against official or authoritative information (Cazzamatta, 2025; Graves, 2017). By making their findings public, fact-checking also entails an element of *confrontation* (Graves & Amazeen, 2019), as it 'endorses or challenges the truthfulness of another individual or organization' (p. 4), thereby implicitly discouraging untruthful statements.

As fact-checking is explicitly motivated by the desire to resolve factual disputes, its performance in this regard has also attracted criticism (Vinhas & Bastos, 2022). For example, it is evident that fact-checkers cannot hope to attain the same degree of methodological rigour as social scientific research, despite both aiming to establish social facts (Uscinski & Butler, 2013). In particular, Uscinski and Butler (2013) stress that many fact-checkers attempt to reach a verdict on statements that are inherently uncheckable, either because they require predicting the future or determining counterfactuals. Another persistent problem is that fact-checks tend to assign one rating to multiple claims, making them difficult to disentangle (Nieminen & Sankari, 2021). Disagreements between different fact-checking organisations also cause contention, although many discrepancies seem to disappear when their rating systems are harmonised (Lee et al., 2023).

More pressing challenges relate to the practical aspects of producing and consuming fact-checks. As the ethos of fact-checking involves careful, evidence-based verification, the process is necessarily labour-intensive and time-consuming, and it is unrealistic to expect it to keep up with the volume and speed of dubious claims it attempts to address (Graves, 2017; Vinhas & Bastos, 2022). At the same time, the reach of fact-checks also remains limited, as available evidence suggests that only a small proportion of people exposed to misinformation also encounter the corresponding fact-checks (Guess et al., 2020). Furthermore, even openly independent fact-checking organisations routinely face accusations of political bias (Hameleers, 2024), while online audiences tend to accept or reject fact-checks based on their partisan views, for example, dismissing corrections that portray their own group in a negative light (Amazeen et al., 2019; Hameleers & van der Meer, 2020; Shin & Thorson, 2017). Finally, while fact-checks can be understood as a form of persuasion (Garrett et al., 2013), a meta-analysis by Walter et al. (2020) found that this function was at odds with the tendency to produce lengthy journalistic articles lacking lexical simplicity, which detracted from their persuasiveness.

3.2 Elements of corrective agency

The previously mentioned traditions offer complementary insights that provide a fuller understanding of correction. Psychological research shows that, when the source, message and audience factors are combined correctly, corrections can be persuasive enough to limit the cognitive effects of misinformation. On the other hand, research in media ethics, journalism and science communication emphasise the broader social conditions in which corrections occur. These include norms, values and expectations, as well as the practical considerations involved in adhering to them. For example, the ideals of scientific or journalistic integrity may be deemed to outweigh the inconvenience of correcting errors and the perceived risks to personal or institutional reputation that they pose.

Taken together, it seems that setting the record straight requires more than favourable message characteristics or communication contexts. It also depends on individual capacities to recognise misinformation and respond to misinformation by disavowing or counterarguing it. In this sense, correction presupposes a degree of agency commonly associated with concepts such as intentionality, motivation, deliberation, freedom, routines and creativity (see Emirbayer & Mische, 1998). More formally, such agency can be understood as the individual or organisational *choices* that not only convey corrective information persuasively but also '*signal* preferences and *perform* broader social functions' (Wu-Ouyang, 2026, p. 3, emphasis added). In this section, I focus on two elements of this agency identified in previous research: recognising misinformation (verification) and engaging in communicative acts to address it (confrontation).

3.2.1 Verification

Before they can persuade others to reject misinformation, people must first reject it themselves. Paquin et al. (2022) formally define this capacity as 'unyielding': recognising a claim to be false even if it is represented as true. In Hall's (1973) classic terminology, corrective actors need to adopt an oppositional rather than dominant reading of misinformation. Following Uscinski, Littrell, and Klofstad (2024), such rejection is complicated by the fact that misinformation is unlikely to ever be self-evidently false. Rather, establishing its veracity requires the means to probe its trustworthiness, that is, verification.

Yu and Shen (2023) define verification as 'actions to check whether the message exists or is true and correct' (p. 3). In most cases, this entails corroborating claims independently (see Fallis, 2004), ideally by consulting various diverse sources that testify for or against the original assertion (Sun, 2022; Torres et al., 2018). From an epistemological perspective, this approach moves away from a strictly realist understanding of misinformation, criticised by scholars such as Friedman (2023),

towards pragmatic objectivity (Suomalainen et al., 2025). In other words, while corrective actors may assume that there are true assertions that exist independently of observers, these facts are only 'made true' after being validated, verified or corroborated (James, 1907).

I should note that requiring verification as a prerequisite for correction is subject to two obvious objections. Firstly, some individuals see no reason to verify a claim because they already possess relevant facts to reject it as false. While this is true, it is also apparent that corrections are not limited to a small group of issue experts, and even experts tend to 'show [their] work' (Graves, 2017, p. 524) by presenting additional evidence in support of their position. Moreover, as I have already discussed, even experts in command of their field routinely publish errors which are typically only corrected following extensive investigations. Secondly, some factual disputes are clearly habitual and inspired by already established beliefs and ideologies associated with one's in-group. For example, accusing someone of dishonesty may require no evidence beyond the conviction that 'they all lie'. It is important not to dismiss such reflexive corrections outright. However, these are only one type of correction and are perhaps unlikely to withstand factual counterarguments or persuade people who do not already share the same worldview and affiliations.

The factors that prompt individuals to verify online claims have so far received little attention (Yu & Shen, 2023), and the evidence is still inconclusive. For instance, some studies have indicated that people verify information when they are uncertain about its accuracy (Tandoc, Ling, et al., 2018), whereas others have found that verification is more prevalent when individuals already suspect the information to be true (Edgerly et al., 2020; Walter, Edgerly, et al., 2021). At the same time, existing research shows that verification behaviour is deeply influenced by social dynamics. Jun, Meng, and Johar (2017) found that the mere perception of the presence of others can reduce verification intentions, as individuals may feel comfortable 'hiding in the crowd' and allowing others to corroborate claims on their behalf. Conversely, verification seems to be more likely when people feel that their social contacts can hold them accountable for sharing dubious claims, thereby damaging their reputation as a trustworthy source (Altay et al., 2020; Waruwu et al., 2020). Sun (2022) also demonstrates that people may be motivated to verify information out of altruistic concerns, for example, if they believe that others are susceptible to misinformation, and that authenticating claims can protect them from harm (see also Edgerly et al., 2020). These findings are particularly notable given the consistent reports of people worldwide being aware of and concerned about misinformation and its effects (Altay, 2026b; Knuutila et al., 2022).

The actual investigative techniques used to verify claims vary depending on the available resources and the institutional context. Graves (2017) notes that professional fact-checkers typically evaluate the plausibility of a claim by

comparing various strands of evidence and following a set of protocols centred on tracing the provenance of the claim – that is, its origin and spread – as well as consulting experts and official government data. These findings are consistent with Cazzamatta's (2025) large-scale content analysis of fact-checking articles worldwide. This analysis also reveals the frequent use of forensic analysis and reference to previous fact-checks, particularly when verifying online content. Similar, albeit more improvised, practices have also been identified among members of the public who lack access to such resources. Tandoc et al. (2018) refer to this as institutional verification, whereby claims are authenticated by consulting formal sources such as search engines, official data sources, and fact-checking websites. However, their study also shows that people often engage in interpersonal verification by consulting informal sources, such as friends and family members, when assessing a claim's credibility.

3.2.2 Confrontation

Even after identifying misinformation, deciding whether to address factual errors is not straightforward. As I have discussed earlier, some people may hesitate and even refuse to issue corrections or retract misleading information. Others may choose not to directly engage with people who share it. In the context of social media, Lee et al. (2022) identify several such passive strategies, including taking breaks from platforms or removing misleading content from one's own feed. Similarly, platform features such as blocking can be used to shield oneself from epistemically dubious claims (Kaiser et al., 2022).

Although these strategies may help individuals to manage their own exposure to misinformation, they conflict with the established view of corrections as a means of presenting accurate information to others in order to acknowledge, challenge or remedy factual errors. Following Bode and Vraga (2025), this requires publicly voicing and standing behind the correction. Furthermore, the existing literature agrees that corrections can and should be designed in a manner that persuades their audiences to reject misinformation. This does not necessarily imply that those correctors seek to 'cure' public opinion by forcing their version of the truth on others against their will (cf. Freiling et al., 2023). For example, research on correction effectiveness subscribes to a view that audiences are ultimately free to choose whether to accept or refuse corrective efforts. Similarly, research on journalistic or scientific corrections struggles to articulate how the corrective information would even reach and persuade misinformed parties, let alone manipulate them. In this sense, correction appears to share a similarity to educating people, and to 'applying more speech to dangerous speech' according to liberal democratic norms (see Howard, 2021).

Following Takahashi and Beebe (1993), we can use the tools of speech act theory (Austin, 1962; Searle, 1969) to define the simplest form of a factual confrontation as an interaction where one person (the corrector) informs another (the recipient) that their statement is incorrect. Such interactions resemble objections in that they too involve individuals verbally standing up to a perceived 'verbal offence' (see Zhao et al., 2024). Intuitively, such a response should contain information that the corrector believes to be true (Bode et al., 2024). However, there is no compelling reason why they would not be mistaken or lie intentionally, providing a refutation that is factually unfounded or misleading – a phenomenon tentatively recognised as a 'reverse correction' (Bode & Vraga, 2021a)[10]. Furthermore, there seems to be a consensus that performing corrections (*illocutions*) can take different forms (*locutions*) whose corrective effect (*perlocutions*) on the recipient may succeed or fail (be *felicitous* or *infelicitous*) depending on the circumstances (for example, corrections may be misunderstood or refused as spurious or incoherent). In summary, although corrections may encourage others to reject misinformation, this effect is far from guaranteed (corrections can fail), and persuasion is just one of many potential motivations behind a correction (alongside less noble goals).

Such confrontations become much more complex in mediated settings. For example, corrections made on semi-public social media typically involve numerous invisible observers, and those making corrections may perform for these secondary 'imagined audiences' (Marwick & boyd, 2011) as well as, or instead of, the person to whom the correction is addressed directly (Bode & Vraga, 2025). The process of broadcasting corrections also faces the challenge of reaching the subset of recipients who would benefit most from 'setting the record straight'. The structures of information environments also place corrective actors in unequal positions. Those who control the technologies used to issue corrections (including digital news outlets or social media platforms) can, to some extent, decide whether they also want to engage in heated debates with audiences who receive their corrections poorly. Conversely, such engagement may be unavoidable for individuals who correct their peers.[11] Accordingly, the possibility that recipients 'talk back' with hostile or dismissive feedback can make one-to-one corrections feel intimidating (Lee et al., 2022) or deeply frustrating (Bieliková et al., 2025), thus increasing the reluctance to engage in correction at all (Bode & Vraga, 2025, Chapter 6). At the same time, direct engagement with their recipients can facilitate learning. Repeated

[10] Note that Austin and Searle disagreed about recognising insincere speech acts. I agree with Searle's (1969, p. 62) argument that speech acts are valid even if the speaker only purports to possess the requisite intentions or beliefs to perform them. In the present context, an insincere correction would involve calling someone's assertion false while privately believing otherwise.

[11] This is not to suggest that fact-checkers could not also receive scathing, unfair or intimidating feedback on their work (see Graves, 2017). Even still, publishers, media companies or social media platforms that control the extent to which they allow direct audience feedback are positioned differently than publics who cannot make these kinds of decisions.

corrective encounters can lead individuals to refine their approaches (Heiss et al., 2023), seek support from other users (Lee et al., 2022) and develop their communication skills over time (Bieliková et al., 2025).

3.3 Who corrects misinformation on social media?

Bode and Vraga (2015) were among the first to suggest that corrections on social media platforms should naturally involve *sources* native to these environments (see Sundar & Nass, 2001). To this end they envisioned that factual disputes could arise through two distinct mechanisms: an end user might receive or witness corrective information made by another end user, or they might interact with a feature of a platform that provides corrective information algorithmically. Accordingly, they devised conditions mimicking two already existing technological affordances: corrections that were contained in Facebook's user comments (which the authors originally referred to as 'social correction', see also Margolin et al., 2018)[12] or provided by the platforms 'related news' feature. Both mechanisms were indeed found to significantly reduce false beliefs.

This typology highlights two complementary approaches to corrective agency on social media platforms. On the one hand, platforms are private businesses that exercise substantial control over their online services, including the analogue and digital protocols through which user-generated content is hosted and governed. They alone can add or remove features, establish or disband teams of employees, enter or exit voluntary partnerships and invest in or divest from technological assets (Bossetta, 2020; Caplan, 2018; Gorwa, 2019a). Most notably, platform companies have established a long-standing authority to intervene in whichever user content or behaviour they deem objectionable (Gillespie, 2018; Klonick, 2018).

The situation for platforms' end users is dramatically different. As recipients of the platforms' design and governance choices, their agency is usually defined in terms of affordances which are interactions with social media features that simultaneously enable and restrict their potential actions (Bucher & Helmond,

[12] Initially, the authors also included a category of 'expert corrections', consistent with a tradition of experiments in which the correction source was usually an expert source cited in a news article (Garrett et al., 2013), or a government agency website (Nyhan & Reifler, 2015). In the original framework, 'expert' and 'social' corrections differ in quality, with the latter representing lay views that lack the institutional credibility of the former. However, although experts do indeed engage in corrections on social media, I would argue that they fall within one of two previously established categories. For instance, if an expert organisation, such as the Finnish Institute for Health and Welfare (THL) (see Vraga & Bode, 2017), wishes to communicate their corrections on social media, it must do so as a platform user (e.g. via a Twitter account or Facebook page) who navigates the same affordances as 'non-professional' social media accounts (Hameleers, 2024). (In fact, THL cited this lack of agency as one of the reasons for leaving Twitter in 2023). Alternatively, experts may partner with platforms to co-create messages presented to their users via related news or warning labels (Vinhas & Bastos, 2025). Accordingly, the authors have since dropped the loaded terms such as *social* and *peer* in favour of the more inclusive category of *user correction* (Bode et al., 2024).

2018). It is therefore understood that most user corrections operate outside the scope of proper content moderation arrangements and tools. In addition to using formal reporting channels and complaint options provided by the platforms, users also improvise with practices like 'flagging' dubious information by commenting it (Achimescu & Chachev, 2021). This does not mean that users have no formal power, nor that some individuals could not be promoted to positions of greater authority (Jhaver et al., 2023). Nevertheless, the abilities of regular social media users to counteract misinformation on behalf of others are minimal compared to the volunteer moderators who have survived on a handful of online sites (Caplan, 2018; McDowell & Vetter, 2020). Users also appear to have no formal say in platforms' policymaking, even when they clearly express their preferences through collective action (Sybert, 2022). In fact, major social media companies have been slow to recognise user corrections, and only recently started to support them through 'community notes' -style arrangements (de Keulenaar, 2025).

Accordingly, these two actors raise very different empirical questions. When it comes to platforms, there is little doubt about their *ability* to address the misinformation they host. Monitoring user content and reviewing it for potential rule violations followed by enforcement is central to their day-to-day operations. A more pressing question is what might motivate them to verify claims for accuracy, how they would undertake such a task, and whether the result would resemble anything like the corrections outlined in the previous sections. For end users, these questions are compounded by even more fundamental questions about how their corrections are made in the first place, and whether users without vast resources and institutional frameworks nevertheless find opportunities to reject dubious claims and contest them in potentially persuasive confrontations.

3.3.1 Moderating misinformation

All platforms moderate the user-generated content they host (Gillespie, 2018). In its simplest form, this involves the detection and evaluation of speech or behaviour deemed objectionable at any given moment, with possible actions taken to sanction it (de Keulenaar et al., 2023). This process is supported by 'institutional mechanisms of adjudication, enforcement, and appeal' (Gillespie et al., 2020, p. 2). In practice, moderating very large platforms has evolved into industrial-scale operations that span in-house teams, outsourced labour, and machine-learning technologies (Caplan, 2018; Gorwa et al., 2020; Roberts, 2019). In a sense that these arrangements enable (or, rather, are legitimised as enabling) 'conditions for ordered rule and collective action' (Stoker 1998, p. 17 as cited in Gorwa, 2019b), they are increasingly discussed within the framework of *platform governance*.

Grimmelmann (2015) was among the first to note that a key characteristic of such governance is the range of enforcement options available to contemporary

platforms. Although many still remove content and suspend accounts (so-called 'hard' moderation), their 'content moderation remedies' (Goldman, 2021) increasingly include less severe punishments such as reducing visibility (Zeng & Kaye, 2022). The expansion and differentiation of enforcement options raise questions about how to justify and apply them in a disciplined manner, particularly given the worldwide impact of such measures on public communication (Flew, 2021; Klonick, 2018). This has led to a broader 'policy turn' (Flew et al., 2019) in the study of digital platforms, whether one interprets platforms' stated guidelines and rules as genuine policy-making instruments (e.g. Flew et al., 2019; Scharlach et al., 2023), or merely as performative gestures independent of actual moderation practices (e.g., Mündges & Park, 2024; Roberts, 2018). Regardless, these documents have intrinsic value in communicating platforms' aspirations, priorities and expectations to various stakeholders (Gillespie, 2018).

Major US-based social media companies have long avoided taking on the editorial responsibilities of traditional media publishers, supported by a particular legal mandate that allows them to simultaneously claim the right to enforce whatever rules and policies they see fit (Klonick, 2018; Napoli & Caplan, 2017). Nonetheless, platforms across the industry have steadily and considerably expanded their definitions of unacceptable speech since early 2010s (e.g. Dubois & Reepschlager, 2024)[13], a period that Bowers and Zittrain (2020) describe as the 'public health era' of platform governance. Against this backdrop, many observers have expressed disappointment at the apparent reluctance of platforms to moderate misleading and false information, despite research documenting its proliferation on these services long before the current misinformation controversies emerged (Kata, 2012). Indeed, companies such as Facebook (now Meta), Twitter/X and YouTube have repeatedly recognised the increase in falsehoods and inauthentic behaviours on their services and have continually revised their approach to this issue (Lien et al., 2022). Yet as recently as 2019, major platform executives were still openly reiterating their opposition to fact-checking their users, particularly about political matters. The COVID-19 pandemic marked a temporary departure from this position, prompting major platforms to adopt, for the first time, formal and global policies prohibiting certain forms of misinformation (Broniatowski et al., 2023; Krishnan et al., 2021). However, many of these policies have since been weakened or revoked. A milestone in this recent trend occurred in January 2025, when Meta

[13] The Platform Governance Archive (Katzenbach et al., 2023) provides simple tools to quantify this expansion across companies with great precision. For instance, between 2011 and 2020 the length of Facebook's community guidelines increased by a factor of 2.7, growing from 727 words covering 10 distinct areas, such as bullying, graphic violence and hate speech, to 2,028 words covering 27 areas, with entirely new tiered categories and sub-categories (e.g. 'Account Integrity and Authentic Identity') and added rationales and examples. During the same period, Twitter's rules underwent an even more drastic transformation, expanding 4.1-fold from 4,120 words in January 2011 to 16,925 words in December 2020.

ended its nine-year partnership with the International Fact-Checking Network (IFCN) under questionable pretences (IFCN, 2025).

One explanation for this ambivalence is a concept of *modulating moderation* proposed by de Keulenaar, Magalhães, and Ganesh (2023). Using Twitter/X as an example, they argue that platforms' seemingly contradictory moderation commitments nevertheless help maintain a pragmatic 'crisis-resistant speech architecture' (p. 281) that can withstand external shocks, criticism, and changing expectations about the limits of acceptable speech. Specifically, they observe that platforms approach to moderation differentiates over time along three distinct ideas of objectionability: 'ugly' content, such as violent threats and child exploitation, which are morally and legally indefensible and warrant severe penalties; which is more ambiguous in its objectionability but still grounded in strong public expectations and social norms; and finally, content such as misinformation, whose objectionability remains 'emergent', or subject of still unsettled normative judgements. Accordingly, policies against this final category are more tentative and potentially fragile as the other two, only emerging in response to new or rapidly developing crises.

Various other explanations as to why platforms have historically adopted or refused certain policy positions have also been proposed. Some argue that social media companies are ultimately motivated to discourage intervention by external regulators (Gorwa, 2019b) either by proactively submitting to voluntary state–firm arrangements (Gorwa, 2019a) or by asserting a commitment to quasi-democratic self-governance (Cowls et al., 2024). Others point to cultural factors, such as internalised values and principles (Hallinan et al., 2022; Scharlach et al., 2023), which appear strikingly similar among industry peers in different parts of the world (Mao et al., 2025). Some have suggested that certain policies are essentially won by creating social and political pressure via negative news coverage (Marchal et al., 2025) or public outcry (Ananny & Gillespie, 2016; Bossetta, 2020). Overall, there seems to be some consensus that these companies balance 'an underlying belief in free speech norms', 'a sense of corporate responsibility', and 'the necessity of meeting users' norms for economic viability' (Klonick, 2018, p. 1618).

According to the statements of US-based platforms, their willingness to intervene against misinformation is largely constrained by idiosyncratic free-speech ideals (see Maddox & Malson, 2020). This view is shared by many Americans, who consider platform moderation as bordering on censorship (Riedl et al., 2021). As far as there are legitimate exceptions to these protections, they tend to hinge on whether misinformation can be categorised as harmful or dangerous information (see Kozyreva et al., 2023). Unsurprisingly, platforms have historically acted most decisively against falsehoods that appear to pose a tangible risk to democracy or the health of its members. Following the widespread concern about 'fake news' during the 2016 US election, Facebook opted not to implement formal

policies prohibiting factually inaccurate content. Instead, it enlisted external fact-checkers to identify and review content on its behalf and submit their verdicts on disputed claims (Ananny, 2018). Similarly, when criticised for hosting anti-vaccine misinformation during the 2018–2019 measles outbreaks, YouTube made limited adjustments to its recommendation systems, yet continued to classify most health misinformation as 'borderline' content – material that could only be 'demoted' as it did not violate the platform's community guidelines (The YouTube Team, 2019).

The search for alternatives to traditional forms of content moderation has increasingly brought platform companies and academic researchers into closer alignment. In recent years, a growing body of behavioural research has specifically begun designing and testing individual-level interventions that could provide actionable 'real solutions' (Clayton et al., 2020, title) to social media's misinformation problem (e.g. Kozyreva et al., 2024; Roozenbeek et al., 2023). These interventions include related news prompts (Bode & Vraga, 2015), warning labels on disputed claims (Clayton et al., 2020), and accuracy nudges for users about to share questionable content (Pennycook et al., 2021). Notably, several such interventions have since been adopted by platforms. While this emerging collaboration between academia and industry is noteworthy, some have criticised it, arguing that it allows platforms to outsource and paper over their problems and avoid addressing the deeper structural conditions that enable problematic behaviours (Gillespie, 2023).

In recent years, users' exposure to such interventions has grown dramatically, particularly in the United States (Saltz et al., 2021). This shift is arguably tied to the outbreak of COVID-19, which led to many social media and messaging platforms introducing explicit policies targeting pandemic-related misinformation (Krishnan et al., 2021). At the same time, independent assessments of the impact of such measures are not entirely reassuring. For example, Broniatowski et al. (2023) found that Facebook's relatively assertive COVID-19 misinformation policies removed both pro- and anti-vaccine content, while inadvertently making the remaining anti-vaccine material more misinformative, polarised, and prominent in users' feeds. Smaller audits conducted by academics and independent investigators have also revealed that enforcement remains selective, with substantial amounts of misinformation identified by experts continuing to circulate on platforms (Brennen et al., 2020; Park & Culloty, 2022).

Conversely, there is also evidence to suggest that platforms do in fact exert some control over the dissemination of misinformation on their services. This influence seems to be driven by a combination of both actual and perceived shifts in content moderation priorities, as signals of policy change may encourage users to readjust to new norms voluntarily. For example, Allcott, Gentzkow, and Yu (2019) found that shares of false news sites on Facebook dropped sharply after the platform announced a ban on fake news in 2016. This was in contrast to Twitter, which never

adopted a similar policy. In another study, McCabe et al. (2024) identified a causal link between Twitter's deplatforming measures and a decrease in the amount of misinformation circulating on the platform in the aftermath of the 6 January attack on the US Capitol. They also suggest that account removals persuaded many other misinformation traffickers to leave Twitter voluntarily. Similarly, Özturan et al. (2025) examined Twitter/X after Elon Musk's public takeover of the platform concluding that the slight increase in low-quality content compared to factual content was due to groups emboldened by perceived or real moderation rollbacks, as well as others leaving the platform for the same reasons.

3.3.2 User corrections

A wide range of people report both witnessing and performing corrections to fellow online users (Bode & Vraga, 2021a, 2021b; Tang et al., 2024). These findings are corroborated by observational studies that have identified instances of user corrections across major platforms, including Facebook, Twitter/X and Reddit (Margolin et al., 2018; Johansen et al., 2022; Achimescu & Chachev, 2021). As these platforms host a wide range of users with varying levels of knowledge, it has become customary to describe these correctors in a way that downplays their epistemic authority and status. For example, they are described as 'peers' (Heiss et al., 2023), 'volunteers' (Kim & Walker, 2020), or a 'crowd' (Micallef et al., 2020) as opposed to 'experts' or 'professionals' working for fact-centred institutions such as health agencies (Vraga & Bode, 2017) or independent fact-checking organisations (Hameleers, 2024). This framing is similar to the concepts of bystander behaviour (Naab et al., 2018) and counter-speech (Fumagalli, 2021), which describe the ambivalent and precarious position of *ordinary* people who choose to police problematic content outside the bounds of professional norms and without explicit support from platforms that mediate their interactions.

Literature on corrections by non-professional users remains relatively scarce, and the conditions that enable them are still debated. While some dedicated users may intentionally monitor platforms for misinformation (Bieliková et al., 2025; Hatakka, 2020), it is equally plausible that others come across misinformation on sufficiently large and heterogeneous platforms comprising different minded people (Bode & Vraga, 2018). Indeed, recent evidence suggests that social media users are prompted to verify and confront information precisely because it challenges their preconceptions (Rossini et al., 2020). This would also help to explain why corrections occur not only among close personal contacts (Waruwu et al., 2020), but also between strangers (Margolin et al., 2018), whose social contexts seem far apart (Lee et al., 2022). In some instances, corrections have even been documented as penetrating otherwise well-insulated online communities (Zollo et al., 2017). Other studies have also found corrections crossing vastly different contexts.

Johansen et al. (2022), for instance, interpret the prevalence of derisive corrections in Twitter discussions about COVID-19 as a means of signalling group boundaries, with those who issued corrections displaying little interest in, or respect for, the perspectives behind the misleading claims they address.

The idea of an independent crowd of potential correctors carries several important implications. Since such corrective potential could be present anywhere within social media networks, misinformation could be confronted much more quickly after its publication. This would enable debunking at a large-scale and virtually in real time (Allen et al., 2021; Bode & Vraga, 2018), as opposed to the slow, meticulous process of professional fact-checking, which has limited reach and reaction time, even when partnering with platforms themselves (Ananny, 2018; Vinhas & Bastos, 2025). User correctors might also benefit from inhabiting the same experience as their recipients. They may be familiar with the platform vernaculars (Gibbs et al., 2015), enabling more tailored interventions than the often confusing and blunt decisions provided by platform moderators (Saltz et al., 2021). User corrections may also gain credibility through personal relationships (Waruwu et al., 2020), which is an advantage that more detached 'asocial corrections' by unknown experts arguably lack (Margolin et al., 2018).

Although user corrections are arguably not supported by professional fact-checking codes of practice (Singer, 2021), this does not necessarily mean that these activities are entirely without principle. The idea that everyone has a responsibility to correct misinformation is widely supported by various online users (Bode & Vraga, 2021b). Similarly, correctors who actively engage with misinformation tend to invoke several normative ideals. They value comparing different information sources and only taking action when there is a solid, verified foundation of facts that permits it (Bieliková et al., 2025). In general, user corrections are also sensitive to prevailing speech norms, such as politeness, and tend to modulate their expression accordingly (Malhotra & Pearce, 2022). Individually, users have been observed to turning to fact-checkers and other authoritative sources in isolation to settle disputes and substantiate their claims with evidence (Margolin et al., 2018). Corrective efforts may also be organised collectively. For example, fan groups (Lee et al., 2022) and activist networks (Hatakka, 2020) have likewise been shown to develop shared strategies for responding to misinformation.

4 RESEARCH APPROACH

This section outlines the research approach and materials on which my dissertation is based. My approach is shaped by the aim of studying corrections observationally, drawing on established social media research designs as well as the limited number of studies that have examined corrections 'in the wild', outside experimental settings. Section 4.1 situates the dissertation within the methodological tradition of studying repurposed digital data and introducing the key epistemic insights that guide my analysis. Section 4.2. reviews existing observational approaches to corrections and highlights their strengths and limitations, which have informed my own approach to identifying corrective agency (Section 4.3).

Empirically, my sub-studies share a naturalistic approach (Lincoln & Guba, 1985) by relying primarily on qualitative methods applied to socially constructed, context-dependent cases. In particular, all three articles are set against the backdrop of intense contestation over the correct interpretation of politicised and tentative facts about the COVID-19 pandemic, spanning from early 2020 through mid-2022. However, they differ in their specific designs and approaches, which are summarised in Table 2.

4.1 Situating the research approach

This dissertation contributes to an emerging body of observational research on correction, drawing broadly on approaches that allow social scientific inquiry in 'real world' situations (Robson, 2009). Unlike experimental studies, which at best mimic correction scenarios, Articles II and III utilise repurposed digital materials – corporate blog posts and social media conversations, respectively – to examine how corrective agency unfolds in natural settings.

From an epistemological perspective, my research is based on the idea that in thoroughly digitised societies (Marres, 2017), the distinction between the 'virtual' and 'real' are becoming increasingly blurred (Rogers, 2013). This does not suggest that such distinctions have become entirely meaningless. For example, the ongoing debate about personhood on the internet indicates that online 'avatars' can extend one's 'offline' self as well as reinvent or disconnect from it (e.g. Markham & Buchanan, 2012; Moore, 2018). What is less contested is that, particularly in the everyday social interactions of the Global North, an 'increasing fraction of life is intrinsically digitally mediated' (Lazer & Radford, 2017, p. 9) by technologies that

Table 2 Research designs of the sub-studies.

Sub-study	Research questions	Materials	Methodology
Article I: Integrative review of criticisms of misinformation problem	How recent mis- and disinformation research has problematised the nature, scope and severity of misleading communication?	38 research articles problematizing the misinformation problem	Integrative literature review (Torraco, 2005) and thematic analysis
Article II: Platform-led misinformation moderation	What discursive strategies Meta, Twitter and YouTube use in their public statements when presenting their actions in countering health misinformation during the COVID-19 pandemic?	118 corporate blog posts by Meta, Twitter and YouTube	Discursive legitimation strategies (Van Leeuwen, 2007)
Article III: User corrections of misinformation in science	How Twitter users evaluate the unreliable COVID-19 studies they circulate? Do these evaluations differ before and after retraction? How often are tweets sharing retracted studies are corrected by others? Does the frequency of these corrections vary before and after retraction, and are corrective messages are constructed in line with best-practice recommendations?	33,730 tweets and 12,020 replies discussing retracted research articles pushing for hydroxycholoroquine and ivermectin as COVID-19 treatments	Mixed-method approach with content analysis, automated nonparametric content analysis (Jerzak et al., 2023) and qualitative content analysis (Elo & Kyngäs, 2008)

continuously capture partial imprints of people's very real communicative exchanges. Compared to past media inscriptions, such individual *digital trace data* (Jungherr, 2018; Latour et al., 2012; Venturini et al., 2018) has become a focal interest for social and political scientists. This is partly because such observations are generated unobtrusively without the direct involvement of researchers (see Given, 2008). Furthermore, these digital records bridge the methodological divide between the quantitative breadth and qualitative depth, enabling study designs that trace, allowing study designs that trace 'large collective dynamics down to single interactions' (Venturini & Rogers, 2019, p. 535).

Repurposing seemingly trivial social media analytics (see Anderson, 2021 for a well-deserved critique) as 'meaningful measures of human society' (Lazer et al., 2021, title) does, of course, present distinct challenges. Some of these relate to the contexts in which the data was originally 'found'. Social media content is famously generated within heavily curated environments, and therefore tend to produce patterns that are confounded by typically unknown algorithmic recommendation and moderation (Wagner et al., 2021). To take another example, even seemingly straightforward social media functions such as 'friending' or 'linking' can have varied and sometimes conflicting uses, making it exceedingly difficult to determine what they actually stand for without studying them in context (Freelon, 2014).

Moreover, even large self-selected populations of social media users rarely provide a good approximation of society as a whole (Hargittai, 2020). Similarly, platform-specific phenomena (e.g. *Twitter* activism) do not necessarily serve as helpful proxies for phenomena occurring outside of them (Venturini et al., 2018). Concerning the aims of this dissertation, it remains particularly unclear whether the online expressions from which political and social scientists frequently use to infer attitudes and opinions, necessarily mirror people's true beliefs about important issues (Jungherr, 2018). Accordingly, while this work broadly falls within what Craig (1999) describes as the sociopsychological tradition of communication research (communication as expressions, interactions, and social influences), I remain sceptical about the idea that online communication can reliably capture people's internal states or record whether someone has actually adjusted their beliefs following a correction.

As digital data can be easily organised into machine-readable structures, I have also been enabled to pursue study designs that use 'computers as an efficient research instrument' (Nelimarkka, 2022, p. 9). Consequently, all three sub-studies have benefited from tools and techniques commonly associated with computational social science (Lazer et al., 2009; Törnberg & Uitermark, 2021). In practice, this has involved automating parts of data collection and manipulation as well as analysis and visualisation using software scripts written in programming languages such as R (e.g. Li, 2019). Drawing on such techniques arguably places this work within a broader tradition of borrowing research practices across scientific domains (Marres, 2012), in no small part because of the epistemic authority and objectivity that 'hard sciences' seemingly lend to social scientific work (Päkkönen, 2025). However, I have exercised some caution in approaching the epistemological assumptions that underlie computational social science. As Törnberg and Uitermark (2021) observe, much of this field is inspired by a vision of 'social physics', which seeks to analyse societies as complex systems that 'can be understood through theories and methods that have been developed to understand other complex systems such as brains, ecosystems, or ant colonies' (pp. 1–2). Rather, I have chosen to adopt a more pragmatic and pluralistic understanding of computational methodology as a form of 'bricolage' (Lindgren, 2020, p. 19) where 'research method is not fully chosen beforehand, but rather emerges as a patchwork of solutions, old or new, to problems faced while carrying out the research'. This also means acknowledging that computational tools introduce their own limitations and artefacts, that need to be interpreted and critically assessed (van Es et al., 2018).

4.2 Identifying corrections ‘in the wild’

Most studies of real-world corrections rely on surveys (e.g. Tang et al., 2024), interviews (e.g. Lee et al., 2022) and focus groups (e.g. Waruwu et al., 2020) as a means of asking participants about their personal experiences of correcting and being corrected on social media. Such self-reports have proven useful for understanding why people choose to confront misinformation, including the motivations involved and the perceived social costs. At the same time, these approaches have limitations in addressing questions such as how corrections unfold, what they look like, and how prevalent they are. Misinformation researchers have also pointed to potential problems with the reliability of self-reported measures, such as social desirability bias (Pasquetto et al., 2020). Avoiding these issues does not, of course, guarantee unbiased research data. For example, both of the empirical datasets analysed in this dissertation are obtained from semi-public sources, where communicators are likely to modulate their expression in accordance with being observed by others (see Lazer et al., 2021).[14] For this reason, the approach taken here should be considered complementary to, rather than a replacement for, experimental and self-reported research approaches.

In attempting to address these gaps, observational studies face different challenges in locating signatures of web-native corrective interactions from digital datasets typically collected via web scraping or application programming interfaces (APIs). Web scraping involves extracting and combining content from the web by converting raw site layout features (e.g. HTML elements) into research data, typically with a help of automated tools (Trezza, 2023). Alternatively, researchers have historically exploited social media information that digital platforms already make available through their APIs (Freelon et al., 2024; Helmond, 2015), enabling, at least in principle, the on-demand harvesting of big data in an already structured format based on customised search queries. From a social scientific perspective, both data gathering techniques resemble what Marres and Gerlitz (2016) refer to as ‘interface methods’: methods that are not exactly developed by social researchers, yet which ‘resonate sufficiently with [their] interests and familiar approaches to offer a productive site of empirical engagement with wider research contexts, practices and apparatuses’ (p. 27).

Although researchers have typically used APIs to obtain complete, non-aggregated datasets that match specific queries over defined time periods

[14] This is especially true of the communications studied in the Article II, the resemblance of which to ‘blog posts’ in a broad sense is open to debate. For example, their stylistic choices are often indistinguishable from platforms’ official self-reporting reports (e.g. Mündges & Park, 2024). Nevertheless, I would argue that company blogs are distinct in that they are produced for a clearly defined audience or regulatory body. Rather than being self-evident policy documents, we have chosen to interpret and operationalise them as such, similarly to any other repurposed digital data. Consequently, these materials are less likely to exhibit biases arising from interaction with researchers.

(Jünger, 2022), recent academic audits have revealed that even privileged 'academic' APIs can suffer from significant data loss and systematic biases (Pearson et al., 2025; Rieder et al., 2025). In addition, social media platforms tightly control what data can be accessed and how by imposing their own rules and retaining the right to limit or revoke API access at any time (Halavais, 2019). Alternatives such as web scraping are not much better, since they tend to require more specialised skills and tools that must be tailored and updated to each website separately. Because scraping is usually conducted without the express consent of site administrators, it also raises ethical and legal concerns (Mancosu & Vegetti, 2020) and may prompt website owners to block data collection efforts altogether.

Several approaches have been developed to identify corrective communication in large-scale digital data, often by using techniques comparable to those employed in detecting misinformation or its purveyors. Some of the earliest procedures relied on approximating corrections as messages that linked to known fact-checking articles (Margolin et al., 2018) or otherwise validated 'debunking' sites (Zollo et al., 2017). The key limitation of this approach is the foundational assumption that people who share fact-check also invariably endorse their verdicts and perform corrective function themselves, which is difficult to defend without empirically validating how linking is used in practice (Freelon, 2014; Weigel & Gitomer, 2025). As with the limitations of link-based detection in misinformation research more broadly (Rogers, 2020), such methods also risk missing all corrective efforts that do not expressly substantiate their arguments with links to specific sources.

Another group of studies has focused on identifying corrections based on their shared linguistic properties, drawing on the methodological families of natural language processing and computational text analysis (Grimmer et al., 2022). Kim and Walker (2020), for instance, used sentence embeddings – a technique that approximates the latent semantic meaning of a text by its context in a numerical vector space – to identify corrections that most closely resembled official guidance from the World Health Organization. Although this approach enables the detection of corrections based solely on language, similar to link-tracing methods, it is severely limited to corrective messages that resemble the style and wording of a small set of authoritative sources. More flexible approaches, adopted by Micallef et al. (Micallef et al., 2020) and Achimescu and Chachev (2021), make use of supervised machine-learning classifiers that allow automated models to predict whether messages belong to predefined categories based on manually-coded training data from the same source. While such models can be optimised to identify corrections in large, previously unseen datasets with high sensitivity, they also come with considerable technical demands.

Finally, one of the simplest approaches has been to mostly substitute automated language processing techniques for content analysis. For example, Johansen et al. (2022) filtered a large corpus of social-media posts using issue-specific keywords,

subsequently manually annotating the material with a coding scheme that approximated corrections as expressions of detachment, irony, or counter-argumentation to misinformation.

4.3 Research materials and their operationalisations

Article I set to review how recent mis- and disinformation research has problematised the nature, scope and severity of misleading communication. Having found that conventional keyword-based search strategies performed poorly in identifying relevant scholarship, I developed an alternative approach based on multi-stage two-way snowball sampling to answer the research question. Firstly, I identified six 'seed' publications authored by key contributors to the debate on prevailing definitions of the misinformation problem: Johan Farkas, Felix Simon and Sacha Altay. Using Semantic Scholar's Open Research Corpus database, I then traced both the references cited in these publications ('backward' sampling) and the works that had cited them ('forward' sampling). Initially I screened only the abstracts before examining potentially relevant publications in full-text form. This first round yielded 21 additional articles, spanning peer-reviewed and non-peer-reviewed research articles published in scientific journals or conference proceedings. A second round of forward and backward snowball sampling based on these articles identified a further seven publications. Finally, I added four other critical papers I was already familiar with. The final dataset comprised 38 critical scholarly contributions published between 2018 and 2024.

In **Article II**, my co-authors and I investigated the strategies used by Meta, Twitter and YouTube in their public statements regarding their efforts to counter health misinformation during the COVID-19 pandemic. To answer this question, we gathered a corpus of corporate blog posts that were programmatically scraped using a Python script with the Selenium framework from the English-language websites of these companies. Following previous operationalisations (e.g. Cotter et al., 2022; Gillespie, 2010; Gillett et al., 2022) such blogs can be understood as a public performance stage where platform companies have a free range to articulate their self-image, values, and policy priorities. Unlike formal community guidelines, platforms' blogs provide a more fine-grained documentation of how policies are justified, revisited and purportedly carried out, for example in terms of the successes of their enforcement. It is worth noting that the value of this corpus did not depend directly on its accuracy in reflecting platform practices. Indeed, recent research suggests that social media platforms tend to 'bullshit' in their communications, offering empty, meaningless or even misleading talk that bears little resemblance to their organisational realities (Frig et al., 2026). Accordingly, we interpreted our research materials first and foremost as expressions of the platforms' aspirations, commitments, and self-presentations, which the companies

sought to communicate to their key stakeholders (Gillespie, 2018, Chapter 2), rather than as honest or complete accounts of how and why moderation decisions were actually made (see Mündges & Park, 2024).

To limit our focus to topics related to the moderation of misinformation about the Coronavirus pandemic (excluding, for example, recurring 'inauthentic behaviour' self-reports), the corpus was filtered by date (1 January 2020 to 31 August 2022) and by keywords commonly associated with Coronavirus, misinformation, and, following Krishnan et al. (2021), credible information. To maximise recall, the search keywords were built gradually around the platforms' preferred terminology using a computer-assisted set discovery process akin to the King-Lam-Roberts algorithm (2017). In practice, the corpus was first searched using unambiguous seed terms such as 'covid', 'misinformation' and 'reliable', after which the query was iteratively expanded with related terms identified in the surrounding text. This produced three lists: 10 keywords related to COVID-19 (e.g. 'emergency' and 'quarantine'), 22 keywords related to misinformation ('deceptive', 'incorrect'), and 19 keywords related to reliable information ('authentic', 'confidence'). Of the 1,351 blog posts, there were 498 posts that contained at least one COVID-19 keyword together with either a misinformation- or credible-information keyword. After we manually inspected and excluded 380 spurious matches, the final dataset comprised 118 posts.

Article III sought to determine how Twitter users evaluated retracted COVID-19 articles on Twitter, as well as the prevalence and sophistication of user corrections presented to these tweets. Materials were first gathered by searching the Retraction Watch database for retractions of articles touting hydroxychloroquine and ivermectin as COVID-19 treatments. Twitter discussions relating to the 12 identified articles were operationalised as tweets that had linked any of these articles before or after their retraction and replies to such tweets, where the message content constituted the primary unit of analysis. Due to the known biases and data loss associated with even privileged research API access (Pearson et al., 2025), I took advantage of Altmetric Explorer's archive of all tweet IDs associated with its indexed papers, only used the API to hydrate these identifiers (retrieval rate of 82.9%), and similarly queried all first-level replies associated with these tweets. After removing retweets, non-English messages, and duplicates, the final dataset comprised 33,730 tweets and 12,020 replies (excluding 1,723 self-replies from the original tweet author).

Corrections were identified by manually annotating a random sample of 1,500 replies, based on a coding scheme in which they were indicated by exchanges in which one social media user publicly refuted another user for making a statement that they considered to be false or misleading, specifically regarding the retracted studies. Notably, this did not require verifying the factual accuracy of the corrections, thereby considerably simplifying the process. A recruited doctoral

candidate and I performed the coding after first testing the codebook and agreement with three rounds of redrawn 150 replies, before reaching an adequate level of inter-rater reliability (Krippendorff's $\alpha = 0.800$, Cohen's $\kappa = 0.799$). These results were further cross-validated using the supervised machine learning algorithm 'ReadMe2' (Jerzak et al., 2023), which was trained using 800 randomly selected replies. Finally, the correction messaging was explored by using a qualitative content analysis procedure (Elo & Kyngäs, 2008), whereby all 539 hand-coded corrections were imported into the qualitative data analysis software ATLAS.ti and repeatedly reviewed to determine the presence or absence of 'best correction practices'.

4.4 Ethical considerations

All digital trace data used in Article III was publicly available on Twitter and obtained in accordance with the Twitter Academic API terms of service (see Halavais, 2019). The additional legal basis for storing and processing copyrighted materials (Articles II and III) is provided by Section 13b of the Finnish Copyright Act (263/2023) and point (e) of Article 6(1) of the EU General Data Protection Regulation (2016/679). As the data were obtained from public sources (namely social media services and their corporate websites), the study design was considered exempt from institutional review in accordance with the Research Ethics Committee guidelines for the Humanities and Social and Behavioural Sciences in Finland. This was the case even though research participants were not contacted to secure their informed consent.

While Article III conforms to basic ethical and legal requirements, its design raises additional considerations relating to participants' autonomy and safety that cannot simply be dismissed because the research does not involve direct interaction with participants (Metcalf & Crawford, 2016). The fact that people share aspects of their social lives in semi-public spaces does not imply that they are comfortable with their data being reused for scientific research (Beninger, 2016; Fiesler & Proferes, 2018; Williams et al., 2017). A key source of tension concerns visibility and privacy. Whereas research exploiting digital trace data often aims to capture passive, everyday online activity, not all users are aware that their interactions are being monitored, collected, and accessed via social media APIs (Zimmer, 2018). Another challenge relates to respect for autonomy. Although informed consent is considered the backbone of procedural research ethics, it is recognised that obtaining it from all participants in big data studies is often infeasible (Beninger, 2016; Markham & Buchanan, 2012).

To address these concerns, additional ethical precautions were taken in line with the Association of Internet Researchers' ethical guidelines (franzke et al., 2019). Data collection was deliberately kept minimal by limiting it to a narrowly defined

set of tweets linking to selected retracted papers. Most analyses were conducted on aggregated data, with only a small subset of materials examined manually (see Fiesler & Proferes, 2018). Messages and metadata (e.g. timestamps and engagement counts) were analysed separately from user profile information to reduce the handling of personal data. All quoted material was pseudonymised using randomly assigned identifiers. Additionally, reported excerpts were lightly edited in accordance with the principle of 'ethical fabrication', whereby representative features from multiple messages were combined to form new, potentially unsearchable composites (Markham, 2012).

5 KEY FINDINGS

5.1 Article I: Substantive critique of the misinformation problem as exaggerated, ambiguous, short-sighted and purposive presents new possibilities to reimagine its potential solutions

In recognising that the framing of a problem shapes its proposed solutions, Article I contributes to the aims of this project by reviewing the ongoing debate concerning the severity and nature of the problems that misinformation poses to individuals and societies (e.g. Altay, 2026a; Ecker et al., 2025; Pasquetto et al., 2024). Specifically, I adopted the theoretical framework of the constructionist perspective on social problems (Best, 2013; Kitsuse & Spector, 1973) to conceptualise misinformation problems as a publicly negotiated and contested collective definition, which is frequently influenced by, though not ontologically reducible to, any 'true' or objectively existing inherent harm that misinformation may or may not possess. The article builds on the observation that criticism of dominant problem framings, such as the assumptions underlying buzzwords like 'post-truth' and 'infodemics', has remained fragmented. At face value, such criticism also relies on somewhat analytically vague yet morally charged accusations of 'moral panic' or 'misinformation about misinformation' (Altay, Berriche, & Acerbi, 2023), which risk obscuring their substantive contributions. To address this issue, Article I analysed thematically 38 scholarly critiques that were identified to problematise the nature, scope and severity of misinformation as a societal threat between 2018 and 2025 (see Figure 2).

A common theme running through the materials was the contrast between what critics described as alarmist and often vague public discourse on the risks of misinformation and the ideals of more careful, transparent and ethically grounded research. While most contributions did not reject the idea that misinformation can pose 'real' problems, they challenged particular aspects of popular problem definitions, arguing that they were based on claims poorly aligned with empirical evidence and sound scientific principles. Furthermore, while critics acknowledged that researchers have played a part in legitimising and establishing such definitions (as a constructivist theory of social problems would also suggest, Blumer, 1971; Kitsuse & Spector, 1973; Schneider, 1985) they also contested the idea that the risks and harms of misinformation represent conditions borne out by systematic evidence and are more likely to be products of amorphous cultural anxieties

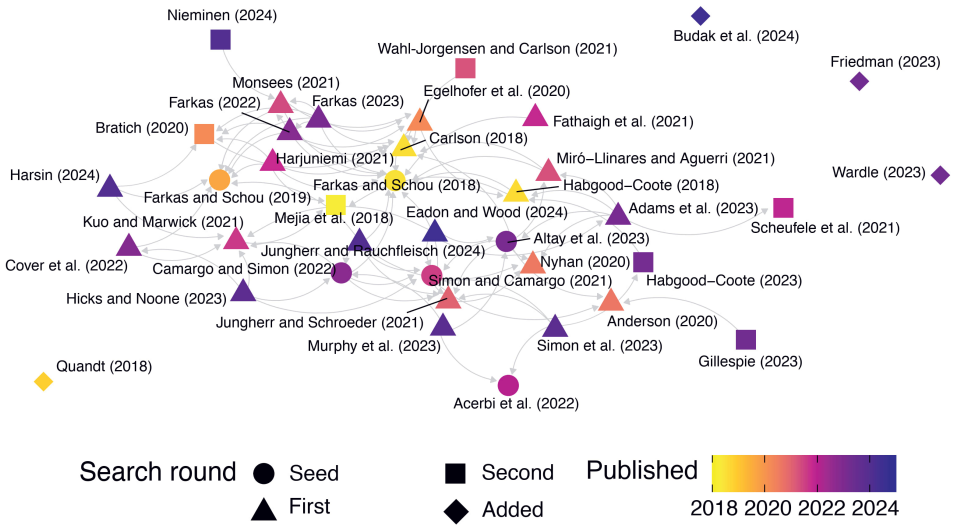


Figure 2 Citation network between the included sample of critical contributions (citation → reference).

surrounding events such as the electoral shocks of 2016, the crisis surrounding the Coronavirus pandemic, and Russia's escalating aggression in Ukraine.

More specifically, the main points of criticism were organised into four key themes. The most prominent concerns the apparent exaggeration and threat inflation in public and scholarly accounts of the supposedly 'rampant' misinformation problem. This critique raises two related issues: firstly, exposure to misleading communication is often assumed to be far more widespread than the available evidence can support, and secondly, the direct impact of misinformation on collective behaviour is also frequently exaggerated beyond current empirical research findings. Concerns were also raised about the conceptual and normative ambiguity of many problem definitions. Critics argue that concrete descriptions of how misinformation works are often substituted with overly dramatic and speculative accounts of what might happen, which often rely on worst-case scenarios and insufficiently motivated projections. Such accounts also tend to resort to metaphorical language that risks oversimplifying or misrepresenting the mechanisms of misinformation dissemination and adoption while also concealing implicit normative assumptions about what constitutes a 'healthy' information environment.

A third critique focuses on the limited scope of current problem definitions. According to these critics, misleading communication is often considered a new, technology-driven issue with little reference to earlier research or interdisciplinary perspectives. Furthermore, these accounts are said to conflate causes and effects, often overlooking the mechanisms through which misleading communication operates. The final theme highlights concern about the intentional or strategic

framing of problems. Critics suggest that some dominant framings of misinformation are uncritically adopted from parties whose interests extend beyond selflessly exploring misleading communication to better understand it. From this perspective, certain problem definitions align with the objectives of individuals or groups who benefit from allocating blame, promoting preferred solutions or attracting attention, rather than acting in good faith to mitigate its negative effects.

5.2 Article II: Moderating COVID-19 misinformation serves platform companies' strategic interests

Article II explores how corrective actors define and ultimately assume their role in addressing misinformation by focusing on digital platform companies – entities that arguably exercise the most control over the contemporary social media services they own and operate. Following the outbreak of the COVID-19 pandemic, we observed that many platforms that had previously resisted fact-checking their users announced relatively comprehensive measures to tackle misinformation (see also Cotter et al., 2022; Krishnan et al., 2021).

To analyse this shift, we conducted a qualitative analysis of corporate blog posts published during the first 20 months of the pandemic by three major US-based platforms – Meta, YouTube and Twitter – that had maintained a comparatively similar stance on content governance. While such public-facing communications offer only a partial insight into the platforms' actual decision-making (see Kreiss & McGregor, 2019; Mündges & Park, 2024), they demonstrate how the platforms sought to explain and justify their actions within the newly adopted policy area at the intersection of the pandemic and misinformation. Specifically, we used theories of discursive legitimation – the linguistic process through which actors aim to construct legitimacy and authority for their practices (Suchman, 1995; Van Leeuwen, 2007) – to examine how the companies strategically mobilised discourse to manage and repair their public standing and seek approval from their core stakeholders.

The findings show all three platforms rapidly converging around announcing similar approaches to 'combat' COVID-19-related misinformation on their services. They do this both by introducing new content moderation remedies highlighted in recent correction literature, such as content labelling and by falling back on more traditional and intrusive measures such as content removal, the appropriate application of which was ostensibly determined by enlisting the help of external public health experts. Similarly, the justifications for these countering methods involved a shared set of discursive framings that map into Van Leeuwen's (Van Leeuwen, 2007) four types of legitimation strategy. Through *moral evaluation*, the platforms portrayed themselves as responsible for protecting users from 'harmful

misinformation', while attempting to balance this role with their commitment to freedom of expression by claiming to apply sanctions that are proportionate to the severity of the offence. Through *rationalisation*, increasingly automated and discriminatory moderation decisions were justified as a necessary way of curbing misinformation with maximum efficiency, with platforms presented as mere administrators who connect their users to the most useful information for their wellbeing. By *invoking authority*, the platforms essentially externalised their responsibility for verifying the accuracy of misinformation to the enlisted medical experts, while presenting themselves as supportive partners to real-world crisis managers. Finally, through *mythopoetic legitimation*, the unusual policy changes were defended as being necessitated by an exceptional but temporary necessity to a crisis with a clear beginning and expected end.

Although the platforms' understanding of misinformation closely reflects the information disorder framework, the data also shows how this view depends on several questionable simplifications required to reduce the epistemological complexities of misinformation into easily scalable, actionable atomised content. At the same time, the emerging conceptualisation of misinformation is thoroughly intertwined with and often subordinate to various strategic goals, especially the platforms' need to defend their existing moderation practices and to signal to external regulators their capability to self-govern, even during particularly disruptive events. However, while the corpus indeed portrays platforms as actors willing to assume more responsibility for the misinformation they host, it remains conspicuously silent on how temporary emergency measures might inform future approaches. In other words, platforms frame misinformation primarily as a pandemic-specific problem, present their interventions as temporary responses and largely avoid engaging with longer-standing epistemic challenges, such as persistent anti-vaccine communities or broader commitments to truthfulness that extend beyond the COVID-19 context.

5.3 Article III: User corrections on Twitter are common and sophisticated

A large body of experimental research has identified and actively promoted a set of 'best correction practices' (Lewandowsky et al., 2020; van der Linden, 2022) , which typically include factual elaboration, simple language, alternative explanations, and a conciliatory tone (e.g. Chan et al., 2017; Ecker et al., 2022; Prike & Ecker, 2023; Walter & Murphy, 2018). To date, however, there has been little evidence that corrective actors follow these tactics in everyday online interactions.

The available evidence would suggest that user correction, often considered one of the more promising mechanisms for countering misinformation, appears particularly unable to fully capitalise on this potential (Bode et al., 2024), as social

media users may avoid engaging with misinformation altogether (Lee et al., 2022), or use corrections as opportunities to ridicule those perceived to be misinformed, instead (Hatakka, 2020; Johansen et al., 2022). These challenges are particularly pronounced in situations where 'enhanced corrections' (MacFarlane et al., 2021) could offset the credibility asymmetry between seemingly ordinary laypeople who try to challenge misinformation originating from high-credibility sources, such as misleading claims based on retracted scientific articles.

Against this backdrop, Article III draws on psychological research into misinformation to examine the actual correction messages posted by Twitter users in response to 'misinformation in science'. This is defined as unreliable research rejected by relevant experts for violating norms of scientific integrity (see also Swire-Thompson & Lazer, 2022; West & Bergstrom, 2021). Empirically, the study revisits Twitter discussions surrounding 12 retracted studies that falsely promoted hydroxychloroquine and ivermectin as viable treatments for COVID-19, given that prior work has shown that Twitter discussions of COVID-19 scholarship was driven overwhelmingly by non-academic users, but without clear expectations as to whether users referencing retracted literature would primarily criticise (Amiri et al., 2024) or endorse it (Abhari & Horvát, 2025).

The findings show that, although most users shared retracted research as genuine evidence (see Figure 2) – possibly due to good-faith misunderstandings of its scientific quality or as part of attempts to contest its retracted status – user corrections were nonetheless common, accounting for nearly half of all replies. A qualitative content analysis (Elo & Kyngäs, 2008) of the 539 identified correction messages further revealed that their design closely aligned with established best-practice recommendations.

Corrections typically employed multiple debunking strategies in combination, such as questioning the credibility of retracted studies, appealing to the consensus among higher-ranking scientists who rejected their findings, and offering alternative explanations, including financial conflicts of interest on the part of the authors. These messages were generally detailed yet accessible, supporting well-reasoned counterarguments with contextual information and references to authoritative sources. At times, they also translated complex issues of research quality into plain, non-technical language. In terms of tone, neutral and matter-of-fact corrections were decidedly more common than frustrated or overtly disrespectful responses, although some did ridicule recipients based on their political, religious or social identities unnecessarily. While such messages may have been less persuasive to their immediate targets, they may nonetheless have served additional functions, such as reinforcing norms surrounding acceptable standards for sharing scientific information among secondary audiences observing the interaction.

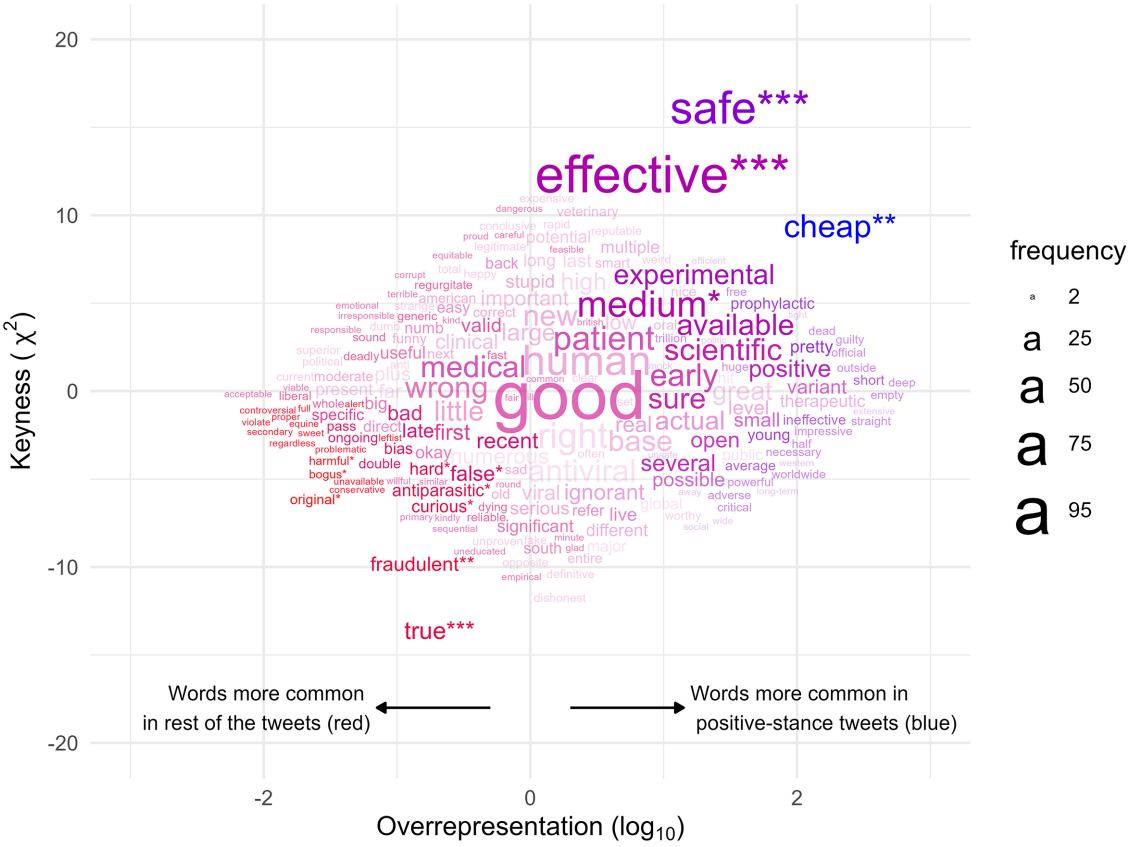


Figure 3 Vocabulary of positive-stance tweets vs. other categories.

Beyond these main findings, the study also shows that corrections were not concentrated among a small group of highly active individuals, but instead involved a broad-based effort, with most users posting only a single correction. This suggests that the ability to recognise misleading scientific claims – even before formal retraction – was relatively widespread, perhaps reflecting heightened public awareness of criticism directed at low-quality COVID19 studies and their unorthodox authors (Karim & Devnarain, 2022; Szalinski, 2021). Finally, the study identifies a statistically significant association between correction prevalence and retraction status: user corrections accounted for 52.1% of all post-retraction replies, compared to 40.4% before retractions were issued, $\chi^2(df = 1,\ n = 1241) = 10.20$, $p = .001$, $\varphi = .09$. Although the effect size is modest, the importance of retraction notices that are timely, clearly signalled, and well-communicated is further emphasised by the substantial proportion of corrections that referenced to articles' retraction notices as their primary argument for rejecting them as misinformation.

6 DISCUSSION

Misinformation can undoubtedly channel dark human impulses and expose alarming vulnerabilities in both our minds and media systems. In this dissertation, however, I have considered the real possibility that contested factual claims are here to stay, the nature of which, along with potential solutions, remain open to academic and political debate (Altay, 2026a; Ecker et al., 2025; Jack, 2019; Nieminen, 2024). As with other unpleasant social media practices (de Seta, 2018), this does not mean that individuals and organisations affected by misinformation do not have good reasons to improve the quality of their information environments. Accordingly, an impressive body of multidisciplinary research, detailing potential solutions to misinformation has emerged over the past few decades. However, this literature is dominated by invented scenarios and strategies that describe the world as it *should be*, including ideal designs for persuasive corrections (Ecker et al., 2022) and optimal 'toolboxes' containing the most well-tested technical interventions (Kozyreva et al., 2024). While I appreciate these contributions, they remain largely disconnected from the ways in which people understand and engage with attempts to mitigate misinformation in their everyday lives.

Of the many potential interventions against misinformation, I have chosen to focus on correction, as it is one of the simplest and most widely studied techniques. As I have argued, it may be precisely because 'correcting someone' seems such a common-sense way of resolving factual disputes that its conceptual dimensions have largely remained unexplored, particularly beyond simple, one-to-one interactions. This gap seems still surprising, given that corrections possess qualities that make them appealing from a democratic perspective and distinguish them from more intrusive techniques. According to the long-standing defence of free-speech protections in the United States, the appropriate response to harmful speech is more speech, not enforced silence (Howard, 2021). While this argument can be misused to justify complacency in the face of genuinely abusive online content, it also provides a principled justification for speaking out against statements that violate one's epistemic sensibilities. The aim is to better inform people while still respecting their right to refuse this information. Even when such corrective actions fail to persuade others to abandon their false beliefs, voicing disagreement publicly seems to have intrinsic value in reinforcing social norms of truthfulness, accountability, and integrity.

6.1 Conceptualising corrective agency ‘in the wild’

In light of the aforementioned aims and research gaps, I have come to view corrections as the actions of individuals and organisations performing corrective functions, namely communicating their disagreements about the facts rather than as a predefined menu of effective *messages* from which people choose how to compose and exchange corrective *information*. Conversely, I have chosen the term *agency* to differentiate this approach from mainstream correction research. This is partly because the term connotes the capacities, abilities, and freedoms of actors who willingly participate in socially meaningful projects (‘Do I have the agency to act?’) as well as the responsibility to exercise this power prudently (‘They had agency all along, but neglected to act on it!’). Conceptually, this perspective connects corrections to research on counter-speech and bystander behaviour, both of which examine how actors lacking formal authority can nevertheless influence social norms and public discourse. It also resonates with platform governance research, which explores how societal actors attempt to shape the behaviour of powerful technology companies and how those companies in turn respond to such social pressure in ways that preserve their legitimacy and autonomy.

The concept of agency also suggests that it is possible to mitigate misinformation outside of established professional cultures, even if efforts that remain voluntary, unplanned, underdeveloped, or temporary may not develop into routines, codes and practices. Social media platforms and their users may engage in activities *resembling* editorial duties or fact-checking. However, as my findings show, it is unfair to compare them to individuals working within organised, fact-centred disciplines that adhere to firmly established norms and protocols, even if they might aspire to similar forms of epistemic authority, either sincerely or strategically. In this sense, corrective agency refers to actions that individuals may feel compelled to take, but may also choose to avoid or pursue for strategic or symbolic reasons that do not necessarily align with an idealised ‘best practice’ for correction.

I propose that ‘corrective agency’ is a *capacity* to publicly perform a certain communicative act: *confronting people for making claims that the corrector rejects as misinformation*, presumably to convince others to adopt their own understanding of facts. This description builds on existing definitions of correction in several ways. Drawing on communication studies, it conceptualises corrections as involving both verification and confrontation. However, it retains the psychological insight that correcting individuals whose beliefs have been shaped by misinformation typically requires persuasion. It also emphasises that real-world corrections differ fundamentally from those in controlled experimental settings, as the former emerge through a process marked by several potential points of failure. For instance, not everyone exposed to misinformation rejects it, not all of those who

reject it choose to act, not all action takes the form of confrontation, and not every confrontation seeks to, or succeeds in, persuading others.

Adjudicating facts is a difficult and often thankless task. Correction involves making numerous judgements that even professional fact-checkers, newsrooms and academic editors routinely struggle to articulate into clear guidelines. This is all while being acutely aware that issuing factual verdicts invites both fair and unfair epistemic criticism (Amazeen, 2015; Graves, 2017; Vinhas & Bastos, 2022). The path of least resistance is often inaction, whether it involves downplaying misinformation as a trivial concern or yielding to fringe factual claims without questioning or thoroughly assessing their validity. However, my empirical evidence also indicates that – at least during a global public health emergency – platforms and their users demonstrated a genuine interest in defending specific facts and actively promoting the value of integrity in public digital communications. Even then, actors appeared to be aware that their factual judgements would not be universally accepted and that they might benefit from conforming to epistemic standards with broad appeal. In other words, the observed corrections did not consist of asserting facts without providing additional supporting evidence. While most of these verification activities were not directly observable, actors often sought to 'show their work' by presenting supporting standards, arguments and evidence, in the spirit of 'thinking like intuitive scientists' (Margolin et al., 2018).

Admittedly, this dissertation demonstrates that this verification process routinely involves referring to more knowledgeable experts, who were effectively delegated the responsibility for establishing the baseline truth. This included deferring to external authorities to determine whether unsettled claims about the nature and effects of COVID-19 were dangerously misleading (Article II) and whether emerging evidence on off-label treatments for the disease could be considered reliable (Article III). However, in many other situations, platforms and their users seemed to act based on a more subjective assessment of misinformation (see Uscinski et al., 2024), rather than trying to establish its veracity by citing epistemic experts or employing more rigorous methods. This was particularly evident in cases where individuals seemed bothered by the factual accuracy of certain statements that did not neatly fit the category of 'false or misleading' information or obvious examples of dubious sources. The user corrections examined in Article III, for instance, involved considerable effort to discredit otherwise credible-looking academic research by combining strategies that delegitimised the research's claim to be 'proper' science (for example, highlighting disreputable authors or absurd research designs) with arguments that revealed logical inconsistencies and conflicting evidence. This undermined their factual credibility.

A more straightforward way of resolving this challenge involves judging supposed misinformation based on moral standards, thereby leaning into the public

discourse of misinformation as an urgent societal problem, particularly during the COVID-19 'infodemic' (Simon & Camargo, 2021). The tendency to paint misinformation as 'dangerous information' was arguably most pronounced among digital platforms – the actors most accustomed to identifying, reviewing and enforcing 'ugly content' (de Keulenaar et al., 2023) whose moderation is least likely to upset their key stakeholders. While this approach was perhaps not unexpected – most academics would also define misinformation in pathological terms (Krause et al., 2022) – it raises interesting questions about whether corrective actors distinguish clearly enough between epistemic and moral categories compared to more established, fact-centred institutions. However, the idea that real-world corrections are genuinely motivated by more value-based epistemologies, as in the case of professional fact-checkers (Suomalainen et al., 2025), should be approached with caution, particularly in the case of platforms whose understanding of online harms has also proven to be flexible and opportunistic (Gillett et al., 2022).

Similar social considerations influence how platform users address misinformation. For example, user corrections were overwhelmingly found to resemble the most persuasive strategies identified in the psychological correction research. It is unlikely that these Twitter users were particularly astute students of human psychology or possessed a deep understanding of misperceptions and how to address them. Undoubtedly, users to some extent expected their responses to influence their audience, and they clearly cared enough about the misperceptions to craft intricate corrections to prove them wrong. However, I find it more plausible that people recognised how challenging it is to convince others to disregard peer-reviewed research findings based on what someone says on the internet. Accordingly, they were probably motivated to use persuasive techniques, such as paying attention to logical coherence, alternative explanations and credibility, to overcome this imbalance. At the same time, anticipating resistance from those they confronted may help to explain why few responses were overtly hostile or condescending. Furthermore, the detached, matter-of-fact tone adopted in most corrections might suggest that users were content to simply warn people about flawed reasoning or express their disagreement, rather than inviting extended debates about whose account was more accurate.

Finally, the proposed definition of 'corrective agency' may help to clear up persistent questions such as raised by Kozyreva et al. (2024): is content moderation a correction as much as experimentally proven 'misinformation interventions'? After all, the Article II revealed only a few examples of platform interventions (namely warning labels, friction, and accuracy prompts), which most correction researchers would readily recognise as 'corrective'. There was virtually no evidence that these tools were specifically chosen to correct misperceptions. In fact, the platform did not present any explicit theory linking its moderation to cognitive effects. Instead, the interventions were justified by goals such as ending the

pandemic, curbing misinformation, and connecting people with the information they need and want. On the other hand, the intuitive notion that some moderation remedies are inherently more 'correction-like' than others turns out to be difficult to defend. Even 'hard' moderation, such as deplatforming, hardly a paradigm for challenging someone for making a false statement, can nonetheless involve identifying and effectively retracting a factual error, be justified in terms of preventing further harm, and still leave affected users with an option to accept or reject the platform's version of reality. Whether such approaches are effective, proportionate or genuinely aimed at changing attitudes is a separate, and important, question. They nevertheless represent one way of addressing the practical challenges of mitigating misinformation, that is, a real-world correction.

6.2 Lessons from observationally grounded understanding of misinformation interventions

The approach outlined in this thesis foregrounds a more nuanced and pragmatic understanding of misinformation mitigation that does not readily translate into normative prescriptions for addressing it. This does not mean that it lacks broader implications at various levels.

At a micro level, the framework suggests that we should pay closer attention to the conditions that enable individuals and organisations to become corrective actors. This involves developing a more contextualised understanding of who engages in correction and why, as well as the *social logics, motivations and incentives* that shape this agency. This differs from approaches that prioritise identifying and prescribing tools that, based on my findings, both platforms and users seem to already employ. A more pressing issue is that these interventions may be adopted for various reasons, some of which could conflict with idealised notions of genuinely correcting misinformation.

Consider warning labels, which are arguably one of the most promising behavioural interventions that platforms could implement *to reduce misperceptions* (Clayton et al., 2020; Mena, 2020). As I demonstrate in Article II, even when platforms adopt such powerful tools into their moderation repertoires, it is unclear whether this has anything to do with their effectiveness. According to the platforms' own accounts, warning labels attached to COVID-19 information seem to serve an entirely different purpose: to justify penalties for ambiguous 'misleading content' through 'softer' interventions, while the companies continued to use aggressive and sometimes indiscriminate content removals to enforce their 'harmful misinformation' policies. Conversely, there was no discussion about the positive impact of warning labels on erroneous beliefs. While such evidence-based interventions may then be preferable to other alternatives, researchers should bear

in mind the risk that their recommendations simply serve what Sætra and Selinger (2024) call 'technofixes': engineering solutions intended to shape user behaviour without challenging the political or economic power of platform owners nor advocating changes in their governing principles.

An alternative approach would be to empower corrective actors to further develop their subjective epistemic sensibilities and lower the barriers to voicing them in factual disagreements. For individual platform users, this could involve discouraging antagonistic interactions where disputes over facts may become an excuse to bully or marginalise people with different opinions and create conditions in which challenging misinformation does not feel overwhelmingly intimidating, risky or arduous. In practice, this might entail symbolic actions, such as incorporating integrity alongside the professed values and principles of other platforms, as well as more tangible tools, such as requiring users to commit to 'pro-truth pledges' (Tsipursky et al., 2018) or providing online communities with templates for locally negotiated and enforced 'misinformation rules' (Chadwick et al., 2023). Realistically, for digital platforms to commit to any of these steps would most likely require continued dialogue with policymakers and civil society actors, and both informal and binding policy arrangements that bind these corporations to approaches beyond impromptu 'governance by shocks' (Ananny & Gillespie, 2016). This would require an honest and multidisciplinary assessment of the evolving factors hindering platforms from developing such commitments independently, including profit motives, ideological views, domestic political realities, and the appetite for quick and easy solutions.

The above observations already touch upon meso-level implications by looking beyond individual corrective interactions. While these interactions are meaningful in aggregate, they tell us little about what might sustain and foster such corrective agency over time. To understand these dynamics, we need to consider how corrective practices might stabilise and evolve when supported collectively rather than performed sporadically. From this perspective, it is evident that the growth of well-established communities of practice (Wenger-Trayner, 1998) around debunking misinformation could be encouraged by relatively modest forms of support. For example, opportunities could be created for these communities to articulate their own interests and co-create their own tools to support this voluntary labour. More broadly, fostering these efforts could involve granting online communities greater independence and autonomy to define their own local norms and responses to misinformation. If platforms were genuinely committed to supporting such developments, this could entail transitioning to more multi-level governance arrangements (Jhaver et al., 2023; Schoenebeck & Blackwell, 2020) that combine centrally mandated rules with at least some community-based, localised moderation (Caplan, 2018). However, given my scepticism that platforms would voluntarily cede this degree of control, it seems more likely that meaningful

resources and support would currently need to come from societal actors with a genuine interest in understanding the lived experiences of those involved in correcting misinformation and in supporting their development.

This vision contrasts sharply with recent plans to introduce 'community notes'-style initiatives across platforms. These changes appear to be less about strengthening platform governance – for example, by enforcing misinformation policies with dedicated moderation teams – and more about shifting the responsibility for contested factual judgements onto users. While such initiatives are still emerging and may offer some improvements over earlier approaches to intervening with misinformation, I fear that absorbing existing, organic forms of corrective agency into platform-controlled systems risks producing the worst of both worlds. There are well-founded concerns about replacing professional fact-checkers with crowdsourced volunteers (IFCN, 2025), particularly ensuring the quality of interventions and preventing them from being weaponised by online mobs. On the other hand, platforms' earlier attempts to outsource moderation while retaining decisive control over their goals, scope and implementation should make us cautious (Ananny, 2018; Crawford & Gillespie, 2016; Vinhas & Bastos, 2025). Given the increasingly explicit political positions articulated by major US-based platforms and their owners, it is not implausible that decisions over which community notes are permitted would be guided by ideology rather than epistemology.

At a *macro level*, a more nuanced understanding of how real-world actors operate provides society with a clearer idea of what to expect from them. It also reveals the tensions between the corrective agency and the broader societal goal of addressing misinformation systematically. While the corrections I have studied are arguably native to their online environments, they are not divorced from wider societal narratives, external actors, and exogenous events. My case studies on the controversies over appropriate COVID-19 protections clearly demonstrate that platforms and their users can be attentive to the growing epistemic consensus around what constitutes misinformation and who can determine it, as well as social pressures for engaging with it.

However, merely because real-world corrections seem to align with these collective efforts does not necessarily mean that they are also driven by a genuine intention to persuade people to abandon false beliefs. While digital platforms may publicly announce sweeping actions against misinformation to demonstrate their ability to manage sudden crises without external regulation, these communications should not be confused with complete or truthful accounts of long-term commitments to curb misinformation. This may be less of an issue for researchers seeking to learn how platforms assume or hedge responsibility. However, as mitigating misinformation is also a political concern, policymakers can justifiably demand that companies that host it align with democratic goals and principles, such

as the rule of law and due process (Suzor, 2018), as well as epistemic rights, such as access to information and freedom of expression (Horowitz et al., 2024). To ensure this, it is wise to remain cautious about how corrective actors portray their own actions and to continue to verify and audit interventions independently. Yet given the current downward trend in data access and unreliable research collaborations, researchers, civil society organisations, and public authorities may still need to consider observational approaches as a necessary, albeit imperfect, means of maintaining a level of oversight.

Finally, this dissertation also highlights to some *meta-level* implications for how misinformation research could benefit more broadly from reconsidering some of its motivations. I echo the call to temper exaggerated threats of misinformation as an usually widespread and potent cause of social harm. As the literature reviewed in Article I convincingly argues, current evidence does not support such sweeping conclusions, even as misinformation is found to be correlated with anti-social tendencies. These concerns highlight the need to reflect more carefully on the assumptions of existing problem definitions, the character of the social arrangements that proposed solutions are meant to protect, and whose interests they prioritise – particularly with respect to vulnerable groups. The findings also suggest that research on misinformation and correction would benefit from a more holistic and interdisciplinary perspective, rather than focusing almost exclusively on technical or behavioural 'fixes' (Camargo & Simon, 2022; Jack, 2019; Tromble & McGregor, 2019). At the same time, researchers are encouraged to critically examine the motivations of actors, such as social media platforms, who stand to gain from appearing to 'solve' misinformation. This includes academics who also play a role in reinforcing certain problem framings (Gillespie, 2023).

Importantly, these arguments are more cautious than some of the strongest critiques in the recent literature. While the current research agenda is still skewed towards certain perspectives and disciplines and has blind spots, this does not mean that existing interventions are inherently flawed or unethical. Some of these may, in fact, actively keep the spread and impact of misinformation in check. The deeper concern is what we may neglect when we promote misinformation as one of the fundamental problems of our time. At the very least, such framing diverts attention away from what are arguably much more urgent global challenges. Regrettably, it also seem to legitimise disproportionate measures that risk undermining civil liberties across the world.

6.3 Conclusion and ways forward

As is so often the case in misinformation research, the question I posed at the beginning seems to have no straightforward answers. Instead of providing a definitive one, I have chosen to present a worthwhile workaround that does not

assume that a definitive solution is necessary or perhaps even attainable: exploring one well-known, but ill-defined mitigation strategy, correction, in the hope that it can unlock new approaches not only to misinformation itself, but also to the broader forms of social action it inspires.

In studying this agency and sketching out its outlines, it has become abundantly clear that a single dissertation can only achieve so much. While the contours of corrective agency may have become clearer, more work is needed in defining and observing corrections online across contexts, platforms, affordances and actor groups. Accordingly, I submit that my work should be considered in light of several limitations. I will discuss these below in the hope of motivating future research.

The theoretical framework and definition of corrective agency adopted in this dissertation are grounded in several research traditions and extensive empirical evidence. Nevertheless, they remain open to further refinement. While my conceptualisation of corrections is intended to be primarily descriptive, it is clear that any conception of communication built around notions of 'mending', 'recalling' or 'retracting' beliefs inevitably carries normative dimensions. A systematic identification of such assumptions (see Freiling et al., 2023), potentially through dialogue with a wider range of disciplines and other salient examples of 'correcting' antisocial behaviour, would be a worthwhile area for investigation.

The two core elements of correction could also be revisited. For example, it could be argued that actors may question or doubt claims before being motivated to verify them. Similarly, previous research suggests that actors often plan or shape their responses before actually confronting misinformation (Bieliková et al., 2025; Lee et al., 2022). Both of these 'pre-phases' could plausibly be treated as distinct and measurable dimensions that influence subsequent acts of verification and confrontation. Finally, the growing awareness of actors who cynically deploy 'corrections' and 'fact-checks' to further their own strategic goals (EDMO, 2022) suggests that the sincerity of the intent to correct may also constitute a useful analytical dimension for distinguishing between different manifestations of real-world correction.

Limiting the scope of this dissertation to the two most well-known and widely discussed types of corrective actors on social media also creates the potential for additional studies. While a typology of agency that distinguishes platforms and their users is supported by existing research, there are good reasons to suspect that it fails to capture the full range of social media actors involved in countering online misinformation. As discussed elsewhere, the terms 'platform' and 'user' are already strategically vague categories (e.g. van Dijck et al., 2018). The latter masks a heterogeneous group of individuals, institutions, and corporations that, while sharing some of the same limitations in terms of their agency, nevertheless can vary greatly along other relevant aspects, such as their online and offline influence and epistemic authority. Correcting misinformation may differ substantially depending

on whether it is issued by anonymous accounts, named individuals, public health agencies, or social media influencers. Corrective agency by 'platforms' in not a monolith, either, as it tends to involve a variety of governance arrangements, including voluntary partnerships with journalistic and fact-checking organisations (Vinhas & Bastos, 2025) as well as ongoing negotiations on the implementation of new legal frameworks, such as the Digital Services Act (Fasel & Weerts, 2024). It is also worth noting that this dissertation focuses solely on a handful of mainstream social media platforms at specific points in their corporate histories. This choice leaves considerable gaps in our understanding of how corrective agency could evolve due to changes in ownership and policies, or how it could develop on smaller special-purpose or 'alternative' platforms that seek to differentiate their governance models from those of their industry peers (Otero & Scharlach, 2025).

Other choices associated with the observational research approach also introduce important limitations. Most notably, findings concerning corrective agency derived from case studies conducted during the global COVID-19 pandemic are necessarily shaped by that context. Medicine and public health are, broadly speaking, fact-rich epistemic fields based on scientific evidence and institutional expertise, and are therefore especially conducive to corrective interventions. This was true even though scientific findings, including those that later attracted broad consensus, could be interpreted in markedly different ways across polarised political and social divides (e.g. Abhari & Horvát, 2025).

From a methodological perspective, it is sufficient to reiterate that while digital trace data opens new opportunities for research design, it is neither inherently superior nor inferior to other techniques for observing corrections 'in the wild'. Moreover, an often-overlooked trade-off of leveraging the speed and scale enabled by computational methods is, somewhat paradoxically, the slow and labour-intensive work to develop the requisite 'technological imagination' (Nelimarkka, 2022, p. 10) needed to work with large digital datasets. These challenges are compounded by the fact that research infrastructures based on social media data continue to evolve and deteriorate. I have been fortunate to conduct most of this project during a relatively stable period of the 'academic API regime' (Freelon et al., 2024, p. 24), when researchers could still expect (or plausibly pretend) to gain easy access to detailed and reasonably comprehensive data. Future researchers already face a markedly different situation, as platforms become more indifferent or even hostile to granting such privileges. This development re-asks the ethical and legal questions about the right to bypass a platform's terms of service, while also increasing interest in alternative approaches, such as data donations, to meaningfully continue monitoring the diversity of agency on social media.

Lastly, the rapid diffusion of LLM-powered chatbots opens up entirely new avenues in correction research. Generative AI not only enables the cheaper and faster production of more realistic misinformation, but people are also increasingly

using chatbots like ChatGPT and Grok to seek factual information in private sessions that lack the kind of corrective agency outlined in this dissertation. Currently, these technologies are agnostic to social facts, yet can convincingly produce 'bullshit' responses (Flyvbjerg, 2025), while also being vulnerable to being misused for promoting propaganda or fringe views that align with their designers (Kerr, 2025; Makhortykh et al., 2024). On the other hand, these same applications are also known to correct their users, and dialogues with AI chatbots have been shown to persuade some individuals to even abandon conspiratorial beliefs in the long term (Costello et al., 2024). Clearly, further attention is needed to understand how people navigate these characteristics individually and collectively.

REFERENCES

Abhari, R., & Horvát, E.-Á. (2025). “They Only Silence the Truth”: COVID-19 retractions and the politicization of science. *Public Understanding of Science*, *34*(3), 291–306. https://doi.org/10.1177/09636625241290142

Acerbi, A., Altay, S., & Mercier, H. (2022). Research note: Fighting misinformation or fighting for information? *Harvard Kennedy School Misinformation Review*, *3*(1), 1–15. https://doi.org/10.37016/mr-2020-87

Achimescu, V., & Chachev, P. D. (2021). Raising the Flag: Monitoring User Perceived Disinformation on Reddit. *Information*, *12*(1), Article 1. https://doi.org/10.3390/info12010004

Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023). (Why) Is Misinformation a Problem? *Perspectives on Psychological Science*, *18*(6), 1436–1463. https://doi.org/10.1177/17456916221141344

Aikin, K. J., Betts, K. R., O’Donoghue, A. C., Rupert, D. J., Lee, P. K., Amoozegar, J. B., & Southwell, B. G. (2015). Correction of Overstatement and Omission in Direct-to-Consumer Prescription Drug Advertising. *Journal of Communication*, *65*(4), 596–618. https://doi.org/https://doi.org/10.1111/jcom.12167

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2), 2053168019848554. https://doi.org/10.1177/2053168019848554

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), eabf4393. https://doi.org/10.1126/sciadv.abf4393

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), eaay3539. https://doi.org/10.1126/sciadv.aay3539

Allport, F. H., & Lepkin, M. (1945). Wartime rumors of waste and special privilege: why some people believe them. *The Journal of Abnormal and Social Psychology*, *40*(1), 3–36. https://doi.org/10.1037/h0058110

Altay, S. (2026a). Rethinking the problem of misinformation and its solutions. *New Media & Society*, 14614448261428635. https://doi.org/10.1177/14614448261428635

Altay, S. (2026b). Who is concerned about online false news and why? Evidence from 46 countries between 2018 and 2023. *Information, Communication & Society*, *29*(7), 2141–2157. https://doi.org/10.1080/1369118X.2025.2551615

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on Misinformation: Conceptual and Methodological Challenges. *Social Media + Society*, *9*(1), 20563051221150412. https://doi.org/10.1177/20563051221150412

Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, *4*(4), 1–13. https://doi.org/10.37016/mr-2020-119

Altay, S., Hacquin, A.-S., & Mercier, H. (2020). Why do so few people share fake news? It hurts their reputation. *New Media & Society*, 1461444820969893. https://doi.org/10.1177/1461444820969893

Amazeen, M. A. (2015). Revisiting the Epistemology of Fact-Checking. *Critical Review*, *27*(1), 1–22. https://doi.org/10.1080/08913811.2014.993890

Amazeen, M. A., & Krishna, A. (2024). Refuting misinformation: Examining theoretical underpinnings of refutational interventions. *Current Opinion in Psychology*, *56*, 101774. https://doi.org/10.1016/j.copsyc.2023.101774

Amazeen, M. A., Vargo, C. J., & Hopp, T. (2019). Reinforcing attitudes in a gatewatching news era: Individual-level antecedents to sharing fact-checks on social media. *Communication Monographs*, *86*(1), 112–132. https://doi.org/10.1080/03637751.2018.1521984

Amiri, M., Yaghtin, M., & Sotudeh, H. (2024). How do tweeters feel about scientific misinformation: an infoveillance sentiment analysis of tweets on retraction notices and retracted papers. *Scientometrics*, *129*(1), 261–287. https://doi.org/10.1007/s11192-023-04871-7

Ananny, M. (2018). *The Partnership Press: Lessons for Platform-Publisher Collaborations as Facebook and News Outlets Team to Fight Misinformation* [A Tow/Knight Report]. Tow Center for Digital Journalism, Columbia University. https://doi.org/10.7916/D85B1JG9

Ananny, M., & Gillespie, T. (2016, September 22). Public platforms: Beyond the cycle of shocks and exceptions. *The Platform Society, The Internet, Policy and Politics Conference (PP2016)*. https://blogs.oii.ox.ac.uk/ipp-conference/sites/ipp/files/documents/anannyGillespie-publicPlatforms-oii-submittedSept8.pdf

Andersen, J., & Søe, S. O. (2020). Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news – the case of Facebook. *European Journal of Communication*, *35*(2), 126–139. https://doi.org/10.1177/0267323119894489

Anderson, C. W. (2021). Fake News is Not a Virus: On Platforms and Their Effects. *Communication Theory*, *31*(1), 42–61. https://doi.org/10.1093/ct/qtaa008

Appelman, A., & Hettinga, K. (2015). Do news corrections affect credibility? Not necessarily. *Newspaper Research Journal*, *36*(4), 415–425. https://doi.org/10.1177/0739532915618403

Appelman, A., & Hettinga, K. E. (2021). The Ethics of Transparency: A Review of Corrections Language in International Journalistic Codes of Ethics. *Journal of Media Ethics*, *36*(2), 97–110. https://doi.org/10.1080/23736992.2021.1899826

Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M. N., Zhang, Y., Pennycook, G., & Rand, D. G. (2023). Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, *7*(9), 1502–1513. https://doi.org/10.1038/s41562-023-01641-6

Arendt, H. (1967, February 25). Truth and Politics. *The New Yorker*, 49–88.

Austin, J. L. (1962). *How To Do Things With Words*. Oxford University Press. https://academic.oup.com/book/5162

Bar-Ilan, J., & Halevi, G. (2017). Post retraction citations in context: a case study. *Scientometrics*, *113*(1), 547–565. https://doi.org/10.1007/s11192-017-2242-0

Barkin, S. M., & Levy, M. R. (1983). All the News That's Fit to Correct: Corrections in the Times and the Post. *Journalism Quarterly*, *60*(2), 218–225. https://doi.org/10.1177/107769908306000202

Beninger, K. (2016). Social Media Users' Views on the Ethics of Social Media Research. In L. Sloan & A. Quan-Haase, *The SAGE Handbook of Social Media Research Methods* (pp. 57–73). SAGE Publications Ltd. https://doi.org/10.4135/9781473983847.n5

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Bennett, W. L., & Pfetsch, B. (2018). Rethinking Political Communication in a Time of Disrupted Public Spheres. *Journal of Communication*, *68*(2), 243–253. https://doi.org/10.1093/joc/jqx017

Berinsky, A. J. (2023). *Political rumors: why we accept misinformation and how to fight it*. Princeton University Press.

Besançon, L., Bik, E., Heathers, J., & Meyerowitz-Katz, G. (2022). Correction of scientific literature: Too little, too late! *PLOS Biology*, *20*(3), e3001572. https://doi.org/10.1371/journal.pbio.3001572

Best, J. (2013). Constructionist Social Problems Theory. *Annals of the International Communication Association*, *36*(1), 237–269. https://doi.org/10.1080/23808985.2013.11679134

Betsch, C., & Sachse, K. (2012). Dr. Jekyll or Mr. Hyde? (How) the Internet influences vaccination decisions: Recent evidence and tentative guidelines for online vaccine communication. *Vaccine*, *30*(25), 3723–3726. https://doi.org/10.1016/j.vaccine.2012.03.078

Bieliková, K., Macková, A. P., & Novotná, M. (2025). Doing What Is Right: Role of Social Media Users in Resilience to Disinformation. *Social Media + Society*, *11*(2), 20563051251342223. https://doi.org/10.1177/20563051251342223

Bimber, B., & Gil de Zúñiga, H. (2020). The unedited public sphere. *New Media & Society*, *22*(4), 700–715. https://doi.org/10.1177/1461444819893980

Blake-Turner, C. (2025). Fake news, relevant alternatives, and the degradation of our epistemic environment. *Inquiry*, *68*(10), 3148–3168. https://doi.org/10.1080/0020174X.2020.1725623

Blumer, H. (1971). Social Problems as Collective Behavior. *Social Problems*, *18*(3), 298–306. https://doi.org/10.2307/799797

Bode, L., & Vraga, E. K. (2015). In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication*, *65*(4), 619–638. https://doi.org/https://doi.org/10.1111/jcom.12166

Bode, L., & Vraga, E. K. (2018). See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication*, *33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

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

Bode, L., & Vraga, E. K. (2021b). Value for Correction: Documenting Perceptions about Peer Correction of Misinformation on Social Media in the Context of COVID-19. *Journal of Quantitative Description: Digital Media*, *1*, 1–22. https://doi.org/10.51685/jqd.2021.016

Bode, L., & Vraga, E. K. (2025). *Observed correction: how we can all respond to misinformation on social media*. Oxford University Press.

Bode, L., Vraga, E. K., & Tang, R. (2024). User correction. *Current Opinion in Psychology*, *56*, 101786. https://doi.org/10.1016/j.copsyc.2023.101786

Bossetta, M. (2020). Scandalous Design: How Social Media Platforms' Responses to Scandal Impacts Campaigns and Elections. *Social Media + Society*, *6*(2), 2056305120924777. https://doi.org/10.1177/2056305120924777

Boulianne, S. (2019). Revolution in the making? Social media effects across the globe. *Information, Communication & Society*, *22*(1), 39–54. https://doi.org/10.1080/1369118X.2017.1353641

Bowers, J., & Zittrain, J. (2020). Answering impossible questions: Content governance in an age of disinformation. *Harvard Kennedy School Misinformation Review*, *1*(1), 1–8. https://doi.org/10.37016/mr-2020-005

Brainard, J. (2018). Rethinking retractions. *Science*, *362*(6413), 390–393. https://doi.org/10.1126/science.362.6413.390

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*, 499–515. https://doi.org/10.1146/annurev-psych-010419-050807

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5). https://doi.org/10.1073/pnas.2020043118

Bratich, J. (2020). Civil Society Must Be Defended: Misinformation, Moral Panics, and Wars of Restoration. *Communication, Culture and Critique*, *13*(3), 311–332. https://doi.org/10.1093/ccc/tcz041

Brennen, J. S., Simon, F. M., Howard, P. N., & Nielsen, R. K. (2020). *Types, Sources, and Claims of COVID-19 Misinformation* [Fact-sheet]. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation

Briand, S. C., Cinelli, M., Nguyen, T., Lewis, R., Prybylski, D., Valensise, C. M., Colizza, V., Tozzi, A. E., Perra, N., Baronchelli, A., Tizzoni, M., Zollo, F., Scala, A., Purnat, T., Czerniak, C., Kucharski, A. J., Tshangela, A., Zhou, L., & Quattrociocchi, W. (2021). Infodemics: A new challenge for public health. *Cell*, *184*(25), 6010–6014. https://doi.org/10.1016/j.cell.2021.10.031

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, *48*(2), 139–166. https://doi.org/10.1080/23808985.2024.2323736

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health*, *108*(10), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Broniatowski, D. A., Simons, J. R., Gu, J., Jamison, A. M., & Abroms, L. C. (2023). The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. *Science Advances*, *9*(37), eadh2132. https://doi.org/10.1126/sciadv.adh2132

Bruns, A. (2019). Filter bubble. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1426

Bucher, T., & Helmond, A. (2018). The Affordances of Social Media Platforms. In J. Burgess, A. Marwick, & T. Poell (Eds), *The SAGE Handbook of Social Media* (pp. 259–280). SAGE.

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature*, *630*(8015), 45–53. https://doi.org/10.1038/s41586-024-07417-w

Bugeja, M. (2007). Making Whole: The Ethics of Correction. *Journal of Mass Media Ethics*, *22*(1), 49–65. https://doi.org/10.1080/08900520701315285

Camargo, C. Q., & Simon, F. M. (2022). Mis- and disinformation studies are too big to fail: Six suggestions for the field's future. *Harvard Kennedy School Misinformation Review*, *3*(5), 1–9. https://doi.org/10.37016/mr-2020-106

Caplan, R. (2018). *Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches*. Data & Society Research Institute. https://datasociety.net/library/content-or-context-moderation/

Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B., & Reifler, J. (2022). The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature Human Behaviour*, *6*, 236–243. https://doi.org/10.1038/s41562-021-01278-3

Carlson, M. (2020). Fake news as an informational moral panic: the symbolic deviancy of social media during the 2016 US presidential election. *Information, Communication & Society*, *23*(3), 374–388. https://doi.org/10.1080/1369118X.2018.1505934

Cazzamatta, R. (2025). Decoding Correction Strategies: How Fact-Checkers Uncover Falsehoods Across Countries. *Journalism Studies*, *26*(7), 777–799. https://doi.org/10.1080/1461670X.2025.2470177

Chadwick, A., Hall, N.-A., & Vaccari, C. (2023). Misinformation rules!? Could "group rules" reduce misinformation in online personal messaging? *New Media & Society*, 14614448231172964. https://doi.org/10.1177/14614448231172964

Chadwick, A., & Stanyer, J. (2022). Deception as a Bridging Concept in the Study of Disinformation, Misinformation, and Misperceptions: Toward a Holistic Framework. *Communication Theory*, *32*(1), 1–24. https://doi.org/10.1093/ct/qtab019

Chadwick, A., Vaccari, C., & O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New Media & Society*, *20*(11), 4255–4274. https://doi.org/10.1177/1461444818769689

Chan, M. S., & Albarracín, D. (2023). A meta-analysis of correction effects in science-relevant misinformation. *Nature Human Behaviour*, *7*, 1514–1525. https://doi.org/10.1038/s41562-023-01623-8

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 16598. https://doi.org/10.1038/s41598-020-73510-5

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real Solutions for Fake News? Measuring the Effectiveness of

General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior*, *42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

Cotter, K., DeCook, J. R., & Kanthawala, S. (2022). Fact-Checking the Crisis: COVID-19, Infodemics, and the Platformization of Truth. *Social Media + Society*, *8*(1), 20563051211069048. https://doi.org/10.1177/20563051211069048

Cover, R., Haw, A., & Thompson, J. D. (2023). Remedying disinformation and fake news? The cultural frameworks of fake news crisis responses and solution-seeking. *International Journal of Cultural Studies*, *26*(2), 216–233. https://doi.org/10.1177/13678779221136881

Cowls, J., Darius, P., Santistevan, D., & Schramm, M. (2024). Constitutional metaphors: Facebook's "supreme court" and the legitimation of platform governance. *New Media & Society*, *26*(5), 2448–2472. https://doi.org/10.1177/14614448221085559

Craig, M., & Vijaykumar, S. (2023). One Dose Is Not Enough: The Beneficial Effect of Corrective COVID-19 Information Is Diminished If Followed by Misinformation. *Social Media + Society*, *9*(2), 20563051231161298. https://doi.org/10.1177/20563051231161298

Craig, R. T. (1999). Communication Theory as a Field. *Communication Theory*, *9*(2), 119–161. https://doi.org/10.1111/j.1468-2885.1999.tb00355.x

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Dan, V., Paris, B., Donovan, J., Hameleers, M., Roozenbeek, J., van der Linden, S., & von Sikorski, C. (2021). Visual Mis- and Disinformation, Social Media, and Democracy. *Journalism & Mass Communication Quarterly*, *98*(3), 641–664. https://doi.org/10.1177/10776990211035395

de Keulenaar, E. (2025). *From Twitter to X: demotion, community notes and the apparent shift from adjudication to consensus-building* (SSRN Scholarly Paper No. 5165083). Social Science Research Network. https://doi.org/10.2139/ssrn.5165083

de Keulenaar, E., Magalhães, J. C., & Ganesh, B. (2023). Modulating moderation: a history of objectionability in Twitter moderation practices. *Journal of Communication*, *73*(3), 273–287. https://doi.org/10.1093/joc/jqad015

de Seta, G. (2018). Trolling, and Other Problematic Social Media Practices. In Jean Burgess, Alice Marwick, & Thomas Poell (Eds), *The SAGE Handbook of Social Media* (pp. 390–411). SAGE Publications Ltd.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, *113*(3), 554–559. https://doi.org/10.1073/pnas.1517441113

DiFonzo, N., & Bordia, P. (2007). Rumor, Gossip and Urban Legends. *Diogenes*, *54*(1), 19–35. https://doi.org/10.1177/0392192107073433

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding Conspiracy Theories. *Political Psychology*, *40*(S1), 3–35. https://doi.org/10.1111/pops.12568

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society*, *21*(5), 729–745. https://doi.org/10.1080/1369118X.2018.1428656

Dubois, E., & Reepschlager, A. (2024). How harassment and hate speech policies have changed over time: Comparing Facebook, Twitter and Reddit (2005–2020). *Policy & Internet*, *16*(3), 523–542. https://doi.org/10.1002/poi3.387

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, *14*, 62. https://doi.org/10.1038/s41467-022-35576-9

Ecker, U. K. H., Hogan, J. L., & Lewandowsky, S. (2017). Reminders and Repetition of Misinformation: Helping or Hindering Its Retraction? *Journal of Applied Research in Memory and Cognition*, *6*(2), 185–192. https://doi.org/10.1016/j.jarmac.2017.01.014

Ecker, U. K. H., Lewandowsky, S., & Chadwick, M. (2020). Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, *5*(1), 41. https://doi.org/10.1186/s41235-020-00241-6

Ecker, U. K. H., Tay, L. Q., Roozenbeek, J., van der Linden, S., Cook, J., Oreskes, N., & Lewandowsky, S. (2025). Why misinformation must not be ignored. *American Psychologist*, *80*(6), 867–878. https://doi.org/10.1037/amp0001448

Ecker, U., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Edgerly, S., Mourão, R. R., Thorson, E., & Tham, S. M. (2020). When Do Audiences Verify? How Perceptions About Message and Source Influence Audience Verification of News Headlines. *Journalism & Mass Communication Quarterly*, *97*(1), 52–71. https://doi.org/10.1177/1077699019864680

EDMO. (2022, March 17). Russian propaganda disguising as fact-checking: a statement from the EDMO Taskforce. *EDMO*. https://edmo.eu/publications/russian-propaganda-disguising-as-fact-checking-a-statement-from-the-edmo-taskforce/

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Emirbayer, M., & Mische, A. (1998). What Is Agency? *American Journal of Sociology*, *103*(4), 962–1023. https://doi.org/10.1086/231294

Evans, J. St. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Fallis, D. (2004). On Verifying the Accuracy of Information: Philosophical Perspectives. *Library Trends*, *52*(3), 463–487.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033. https://doi.org/10.1073/pnas.1212247109

Farkas, J. (2023). News on fake news: Logics of media discourses on disinformation. *Journal of Language and Politics*, *22*(1), 1–21. https://doi.org/10.1075/jlp.22020.far

Farkas, J., & Schou, J. (2018). Fake News as a Floating Signifier: Hegemony, Antagonism and the Politics of Falsehood. *Javnost - The Public*, *25*(3), 298–314. https://doi.org/10.1080/13183222.2018.1463047

Farkas, J., & Schousboe, S. (2024). Facts, values, and the epistemic authority of journalism: How journalists use and define the terms fake news, junk news, misinformation, and disinformation. *Nordicom Review*, *45*(1), 137–157. https://doi.org/10.2478/nor-2024-0016

Fasel, M., & Weerts, S. (2024). Can Facebook's community standards keep up with legal certainty? Content moderation governance under the pressure of the Digital Services Act. *Policy & Internet*, *16*(3), 588–606. https://doi.org/10.1002/poi3.391

Fiesler, C., & Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, *4*(1), 2056305118763366. https://doi.org/10.1177/2056305118763366

Flew, T. (2021). *The Challenge of Media Platform Regulation for Small and Medium-Sized Nations* (SSRN Scholarly Paper No. 3951610). Social Science Research Network. https://doi.org/10.2139/ssrn.3951610

Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, *10*(1), 33–50. https://doi.org/10.1386/jdmp.10.1.33_1

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics. *Political Psychology*, *38*(S1), 127–150. https://doi.org/https://doi.org/10.1111/pops.12394

Flyvbjerg, B. (2025). AI as artificial ignorance. *Project Leadership and Society*, *6*, 100208. https://doi.org/10.1016/j.plas.2025.100208

Fox, C. J. (1983). *Information and Misinformation: An Investigation of the Notions of Information, Misinformation, Informing, and Misinforming*. Greenwood Publishing Group.

Frampton, G., Woods, L., & Scott, D. A. (2021). Inconsistent and incomplete retraction of published research: A cross-sectional study on Covid-19 retractions and recommendations to mitigate risks for research, policy and practice. *PLOS ONE*, *16*(10), e0258935. https://doi.org/10.1371/journal.pone.0258935

Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.

franzke, aline shakti, Bechmann, A., Zimmer, M., Ess, C. M., & Association of Internet Researchers. (2019, October 6). *Internet Research: Ethical Guidelines 3.0*. Association of Internet Researchers. https://aoir.org/reports/ethics3.pdf

Freelon, D. (2014). On the Interpretation of Digital Trace Data in Communication and Social Computing Research. *Journal of Broadcasting & Electronic Media*, *58*(1), 59–75. https://doi.org/10.1080/08838151.2013.875018

Freelon, D., Monzer, C., Jeon, G., Moy, C., & Williams, N. (2024). The Post-API Age of Social Media Data Access: Past, Present, and Future. *The ANNALS of the American Academy of Political and Social Science*, *715*(1), 16–37. https://doi.org/10.1177/00027162251372557

Freiling, I., Krause, N. M., & Scheufele, D. A. (2023). Science and Ethics of "Curing" Misinformation. *AMA Journal of Ethics*, *25*(3), 228–237. https://doi.org/10.1001/amajethics.2023.228

Friedman, J. (2023). Post-Truth and the Epistemological Crisis. *Critical Review*, *35*(1–2), 1–21. https://doi.org/10.1080/08913811.2023.2221502

Frig, M., Dovbysh, O., & Laaksonen, S.-M. (2026). Greening the Platforms or Organisational Bullshit? A Typology of Empty and Misleading Environmental Communication by Global Platform Companies. In S.-M. Laaksonen, M. Pantti, & O. Dovbysh (Eds), *Platforms and the Planet: Big Tech, Digital Platforms and Environmental Responsibility* (pp. 167–182). Emerald Publishing Limited.

Fumagalli, C. (2021). Counterspeech and Ordinary Citizens: How? When? *Political Theory*, *49*(6), 1021–1047. https://doi.org/10.1177/0090591720984724

Gagliardone, I., Diepeveen, S., Findlay, K., Olaniran, S., Pohjonen, M., & Tallam, E. (2021). Demystifying the COVID-19 Infodemic: Conspiracies, Context, and the Agency of Users. *Social Media + Society*, *7*(3), 20563051211044233. https://doi.org/10.1177/20563051211044233

Garisto, D. (2025). Science sleuths raise concerns about scores of bioengineering papers. *Nature*. https://doi.org/10.1038/d41586-025-03870-3

Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the Corrective Effects of Media-Based Political Fact Checking? The Role of Contextual Cues and Naïve Theory. *Journal of Communication*, *63*(4), 617–637. https://doi.org/10.1111/jcom.12038

Gibbs, M., Meese, J., Arnold, M., Nansen, B., & Carter, M. (2015). #Funeral and Instagram: death, social media, and platform vernacular. *Information, Communication & Society*, *18*(3), 255–268. https://doi.org/10.1080/1369118X.2014.987152

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347–364. https://doi.org/10.1177/1461444809342738

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Illustrated edition). Yale University Press.

Gillespie, T. (2023). The Fact of Content Moderation; Or, Let's Not Solve the Platforms' Problems for Them. *Media and Communication*, *11*(2), 406–409. https://doi.org/10.17645/mac.v11i2.6610

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., Roberts, S. T., Sinnreich, A., & West, S. M. (2020). Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4). https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy

Gillett, R., Stardust, Z., & Burgess, J. (2022). Safety for Whom? Investigating How Platforms Frame and Perform Safety and Harm Interventions. *Social Media + Society*, *8*(4), 20563051221144315. https://doi.org/10.1177/20563051221144315

Goldman, E. (2021). Content Moderation Remedies. *Michigan Technology Law Review*, *28*(1), 1–60. https://doi.org/https://doi.org/10.36645/mtlr.28.1.content

Gorwa, R. (2019a). The platform governance triangle: conceptualising the informal regulation of online content. *Internet Policy Review*, *8*(2), 1–22.

Gorwa, R. (2019b). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. https://doi.org/10.1080/1369118X.2019.1573914

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 2053951719897945. https://doi.org/10.1177/2053951719897945

Graves, L. (2016). *Deciding What's True: The Rise of Political Fact-Checking in American Journalism*. Columbia University Press.

Graves, L. (2017). Anatomy of a Fact Check: Objective Practice and the Contested Epistemology of Fact Checking. *Communication, Culture and Critique*, *10*(3), 518–537. https://doi.org/10.1111/cccr.12163

Graves, L., & Amazeen, M. A. (2019). Fact-Checking as Idea and Practice in Journalism. In *Oxford Research Encyclopedia of Communication*. https://doi.org/10.1093/acrefore/9780190228613.013.808

Graves, L., Nyhan, B., & Reifler, J. (2016). Understanding Innovations in Journalistic Practice: A Field Experiment Examining Motivations for Fact-Checking. *Journal of Communication*, *66*(1), 102–138. https://doi.org/10.1111/jcom.12198

Greitemeyer, T. (2014). Article retracted, but the message lives on. *Psychonomic Bulletin & Review*, *21*(2), 557–561. https://doi.org/10.3758/s13423-013-0500-6

Grice, H. P. (1975). Logic and Conversation. In D. Davidson (Ed.), *The logic of grammar* (pp. 64–75). Dickenson Pub. Co.

Grieneisen, M. L., & Zhang, M. (2012). A Comprehensive Survey of Retracted Articles from the Scholarly Literature. *PLOS ONE*, *7*(10), e44118. https://doi.org/10.1371/journal.pone.0044118

Grimmelmann, J. (2015). The Virtues of Moderation. *Yale Journal of Law and Technology*, *17*, 42–109.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Guess, A. M., & Lyons, B. A. (2020). Misinformation, Disinformation, and Online Propaganda. In N. Persily & J. A. Tucker (Eds), *Social Media and Democracy* (1st edn, pp. 10–33). Cambridge University Press. https://doi.org/10.1017/9781108890960.003

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), Article 5. https://doi.org/10.1038/s41562-020-0833-x

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Habermas, J. (1984). *The theory of communicative action, volume one* (T. McCarthy, Trans.). Beacon Press.

Habgood-Coote, J. (2019). Stop talking about fake news! *Inquiry*, *62*(9–10), 1033–1065. https://doi.org/10.1080/0020174X.2018.1508363

Habgood-Coote, J. (2023). Deepfakes and the Epistemic Apocalypse. *Synthese*, *201*(3), 1–23. https://doi.org/10.1007/s11229-023-04097-3

Halavais, A. (2019). Overcoming terms of service: a proposal for ethical distributed research. *Information, Communication & Society*, *22*(11), 1567–1581. https://doi.org/10.1080/1369118X.2019.1627386

Hall, S. (1973). *Encoding and decoding in the television discourse* (CCCS Selected Working Papers. Vol. 2). CCCS Stencilled Occasional Papers. http://epapers.bham.ac.uk/2962/

Hallinan, B., Scharlach, R., & Shifman, L. (2022). Beyond Neutrality: Conceptualizing Platform Values. *Communication Theory*, *32*(2), 201–222. https://doi.org/10.1093/ct/qtab008

Hameleers, M. (2023a). Disinformation as a context-bound phenomenon: toward a conceptual clarification integrating actors, intentions and techniques of creation and dissemination. *Communication Theory*, *33*(1), 1–10. https://doi.org/10.1093/ct/qtac021

Hameleers, M. (2023b). The (Un)Intended Consequences of Emphasizing the Threats of Mis- and Disinformation. *Media and Communication*, *11*(2), 5–14. https://doi.org/10.17645/mac.v11i2.6301

Hameleers, M. (2024). Why Do Social Media Users Accept, Doubt or Resist Corrective Information? A Qualitative Analysis of Comments in Response to Corrective Information on Social Media. *Journalism Studies*, *25*(7), 776–793. https://doi.org/10.1080/1461670X.2024.2340591

Hameleers, M. (2025). Reconciling discrepancies between low estimates of misinformation exposure versus high perceived threats: a theoretical overview and a future research agenda. *Communication Theory*, qtaf021. https://doi.org/10.1093/ct/qtaf021

Hameleers, M., Brosius, A., Marquart, F., Goldberg, A. C., van Elsas, E., & de Vreese, C. H. (2022). Mistake or Manipulation? Conceptualizing Perceived Mis- and Disinformation among News Consumers in 10 European Countries. *Communication Research*, *49*(7), 919–941. https://doi.org/10.1177/0093650221997719

Hameleers, M., Meer, T. G. L. A. van der, & Brosius, A. (2020). Feeling "disinformed" lowers compliance with COVID-19 guidelines: Evidence from the US, UK, Netherlands, and Germany. *Harvard Kennedy School Misinformation Review*, *1*(3), 1–16. https://doi.org/10.37016/mr-2020-023

Hameleers, M., & Minihold, S. (2022). Constructing Discourses on (Un)truthfulness: Attributions of Reality, Misinformation, and Disinformation by Politicians in a Comparative Social Media Setting. *Communication Research*, *49*(8), 1176–1199. https://doi.org/10.1177/0093650220982762

Hameleers, M., Tulin, M., De Vreese, C., Aalberg, T., Van Aelst, P., Cardenal, A. S., Corbu, N., Van Erkel, P., Esser, F., Gehle, L., Halagiera, D., Hopmann, D., Koc-Michalska, K., Matthes, J., Meltzer, C., Mihelj, S., Schemer, C., Sheafer, T., Splendore, S., Stanyr, J., Stepinska, A., Stetka, V., Strömbäck, J., Terren, L., Theocharis, Y., & Zoizner, A. (2024). Mistakenly misinformed or intentionally deceived? Mis- and Disinformation perceptions on the Russian War in Ukraine among citizens in 19 countries. *European Journal of Political Research*, *63*(4), 1642–1654. https://doi.org/10.1111/1475-6765.12646

Hameleers, M., & van der Meer, T. G. L. A. (2020). Misinformation and Polarization in a High-Choice Media Environment: How Effective Are Political Fact-Checkers? *Communication Research*, *47*(2), 227–250. https://doi.org/10.1177/0093650218819671

Hargittai, E. (2020). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322

Harsin, J. (2024). + Three Critiques of Disinformation (For-Hire) Scholarship: Definitional Vortexes, Disciplinary Unneighborliness, and Cryptonormativity. *Social Media + Society*, *10*(1), 20563051231224732. https://doi.org/10.1177/20563051231224732

Hatakka, N. (2020). Expose, debunk, ridicule, resist! Networked civic monitoring of populist radical right online action in Finland. *Information, Communication & Society*, *23*(9), 1311–1326. https://doi.org/10.1080/1369118X.2019.1566392

Heiss, R., Nanz, A., Knupfer, H., Engel, E., & Matthes, J. (2023). Peer correction of misinformation on social media: (In)civility, success experience and relationship consequences. *New Media & Society*, *27*(4), 2293–2312. https://doi.org/10.1177/14614448231209946

Helmond, A. (2015). The Platformization of the Web: Making Web Data Platform Ready. *Social Media + Society*, *1*(2), 2056305115603080. https://doi.org/10.1177/2056305115603080

Hettinga, K. E. (2013). Experiencing Error: How Journalists Describe What It's Like When the Press Fails. *Journal of Mass Media Ethics*, *28*(1), 30–41. https://doi.org/10.1080/08900523.2012.746529

Hettinga, K. E., & Appelman, A. (2014). Corrections of Newspaper Errors Have Little Impact. *Newspaper Research Journal*, *35*(1), 51–63. https://doi.org/10.1177/073953291403500105

Hochschild, J. L., & Einstein, K. L. (2015). Do Facts Matter? Information and Misinformation in American Politics. *Political Science Quarterly*, *130*(4), 585–624. https://doi.org/https://doi.org/10.1002/polq.12398

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, 1–9. https://doi.org/10.1038/s41562-024-01884-x

Horowitz, M., Nieminen, H., Lehtisaari, K., & D'Arma, A. (Eds). (2024). *Epistemic Rights in the Era of Digital Disruption*. Springer Nature.

Howard, J. W. (2021). Terror, Hate and the Demands of Counter-Speech. *British Journal of Political Science*, *51*(3), 924–939. https://doi.org/10.1017/S000712341900053X

IFCN. (2025, January 9). An open letter to Mark Zuckerberg from the world's fact-checkers, nine years later. *Poynter*. https://www.poynter.org/ifcn/2025/an-open-letter-to-mark-zuckerberg-from-the-worlds-fact-checkers-nine-years-later/

Jack, C. (2017). *Lexicon of Lies* [Report]. Data & Society Research Institute. https://datasociety.net/library/lexicon-of-lies/

Jack, C. (2019). Wicked Content. *Communication, Culture and Critique*, *12*(4), 435–454. https://doi.org/10.1093/ccc/tcz043

James, W. (1907). Pragmatism's Conception of Truth. *The Journal of Philosophy, Psychology and Scientific Methods*, *4*(6), 141–155. https://doi.org/10.2307/2012189

Janmohamed, K., Walter, N., Nyhan, K., Khoshnood, K., Tucker, J. D., Sangngam, N., Altice, F. L., Ding, Q., Wong, A., Schwitzky, Z. M., Bauch, C. T., De Choudhury, M., Papakyriakopoulos, O., & Kumar, N. (2021). Interventions to Mitigate COVID-19 Misinformation: A Systematic Review and Meta-Analysis. *Journal of Health Communication*, *26*(12), 846–857. https://doi.org/10.1080/10810730.2021.2021460

Jerzak, C. T., King, G., & Strezhnev, A. (2023). An Improved Method of Automated Nonparametric Content Analysis for Social Science. *Political Analysis*, *31*(1), 42–58. https://doi.org/10.1017/pan.2021.36

Jhaver, S., Frey, S., & Zhang, A. X. (2023). Decentralizing Platform Power: A Design Space of Multi-Level Governance in Online Social Platforms. *Social Media + Society*, *9*(4), 20563051231207857. https://doi.org/10.1177/20563051231207857

Johansen, N., Marjanovic, S. V., Kjaer, C. V., Baglini, R. B., & Adler-Nissen, R. (2022). Ridiculing the "tinfoil hats:" Citizen responses to COVID-19 misinformation in the Danish facemask debate on Twitter. *Harvard Kennedy School Misinformation Review*, *3*(2), 1–23. https://doi.org/10.37016/mr-2020-93

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Jones-Jang, S. M., Kim, D. H., & Kenski, K. (2021). Perceptions of mis- or disinformation exposure predict political cynicism: Evidence from a two-wave survey during the 2018 US midterm elections. *New Media & Society*, *23*(10), 3105–3125. https://doi.org/10.1177/1461444820943878

Jun, Y., Meng, R., & Johar, G. V. (2017). Perceived social presence reduces fact-checking. *Proceedings of the National Academy of Sciences*, *114*(23), 5976–5981. https://doi.org/10.1073/pnas.1700175114

Jünger, J. (2022). A brief history of APIs: Limitations and opportunities for online research. In U. Engel, U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds), *Handbook of Computational Social Science, Volume 2: Data Science, Statistical Modelling, and Machine Learning Methods* (pp. 17–32). Routledge.

Jungherr, A. (2018). Normalizing Digital Trace Data. In N. Stroud & S. McGregor (Eds), *Digital Discussions* (pp. 9–35). Routledge.

Jungherr, A., & Rauchfleisch, A. (2024). Negative Downstream Effects of Alarmist Disinformation Discourse: Evidence from the United States. *Political Behavior*, *46*, 2123–2143. https://doi.org/10.1007/s11109-024-09911-3

Kaiser, J., Vaccari, C., & Chadwick, A. (2022). Partisan Blocking: Biased Responses to Shared Misinformation Contribute to Network Polarization on Social Media. *Journal of Communication*, *72*(2), 214–240. https://doi.org/10.1093/joc/jqac002

Kampf, Z., & Daskal, E. (2014). Communicating Imperfection: The Ethical Principles of News Corrections. *Communication Theory*, *24*(2), 165–185. https://doi.org/10.1111/comt.12035

Karim, S. S. A., & Devnarain, N. (2022). Time to Stop Using Ineffective Covid-19 Drugs. *New England Journal of Medicine*, *387*(7), 654–655. https://doi.org/10.1056/NEJMe2209017

Karlova, N. A., & Fisher, K. E. (2013). A social diffusion model of misinformation and disinformation for understanding human information behaviour. *Information Research*, *18*(1), A573.

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine*, *28*(7), 1709–1716. https://doi.org/10.1016/j.vaccine.2009.12.022

Kata, A. (2012). Anti-vaccine activists, Web 2.0, and the postmodern paradigm – An overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine, Special Issue: The Role of Internet Use in Vaccination Decisions*, *30*(25), 3778–3789. https://doi.org/10.1016/j.vaccine.2011.11.112

Katzenbach, C., Dergachava, D., Fischer, A., Kopps, A., Kolesnikov, S., Redeker, D., & Viejo Otero, P. (2023). *Platform Governance Archive (PGA) v2. [data set].* [Data set]. OSF. https://doi.org/10.17605/OSF.IO/XSBPT

Kerr, D. (2025, May 15). Musk's AI Grok bot rants about 'white genocide' in South Africa in unrelated chats. *The Guardian*. https://www.theguardian.com/technology/2025/may/14/elon-musk-grok-white-genocide

Kim, H., & Walker, D. (2020). Leveraging volunteer fact checking to identify misinformation about COVID-19 in social media. *Harvard Kennedy School Misinformation Review*, *1*(3), 1–10. https://doi.org/10.37016/mr-2020-021

King, G., Lam, P., & Roberts, M. E. (2017). Computer-Assisted Keyword and Document Set Discovery from Unstructured Text. *American Journal of Political Science*, *61*(4), 971–988. https://doi.org/10.1111/ajps.12291

Kitsuse, J. I., & Spector, M. (1973). Toward a Sociology of Social Problems: Social Conditions, Value-Judgments, And Social Problems. *Social Problems*, *20*(4), 407–419. https://doi.org/10.2307/799704

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harward Law Review*, *131*(6), 1598–1670.

Kluknavská, A., & Eisele, O. (2023). Trump and circumstance: introducing the post-truth claim as an instrument for investigating truth contestation in public discourse. *Information, Communication & Society*, *26*(8), 1583–1600. https://doi.org/10.1080/1369118X.2021.2020322

Knuutila, A., Neudert, L.-M., & Howard, P. N. (2022). Who is afraid of fake news? Modeling risk perceptions of misinformation in 142 countries. *Harvard Kennedy School Misinformation Review*, *3*(3), 1–12. https://doi.org/10.37016/mr-2020-97

Korpela, K. M. (2010). How long does it take for the scientific literature to purge itself of fraudulent material?: the Breuning case revisited. *Current Medical Research and Opinion*, *26*(4), 843–847. https://doi.org/10.1185/03007991003603804

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, *120*(7), e2210666120. https://doi.org/10.1073/pnas.2210666120

Kozyreva, A., Lorenz-Spreen, P., Herzog, S. M., Ecker, U. K. H., Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Barzilai, S., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., Panizza, F., Pennycook, G., Rand, D. G., Rathje, S., Reifler, J., Schmid, P., Smith, M., Swire-Thompson, B., Szewach, P., van der Linden, S., & Wineburg, S. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

Krause, N. M., Freiling, I., & Scheufele, D. A. (2022). The "Infodemic" Infodemic: Toward a More Nuanced Understanding of Truth-Claims and the Need for (Not) Combatting Misinformation. *The ANNALS of the American Academy of Political and Social Science*, *700*(1), 112–123. https://doi.org/10.1177/00027162221086263

Kreiss, D., & McGregor, S. C. (2019). The "Arbiters of What Our Voters See": Facebook and Google's Struggle with Policy, Process, and Enforcement around Political Advertising. *Political Communication*, *36*(4), 499–522. https://doi.org/10.1080/10584609.2019.1619639

Krishnan, N., Gu, J., Tromble, R., & Abroms, L. C. (2021). Research note: Examining how various social media platforms have responded to COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*, *2*(6), 1–25. https://doi.org/10.37016/mr-2020-85

Kühberger, A., Streit, D., & Scherndl, T. (2022). Self-correction in science: The effect of retraction on the frequency of citations. *PLOS ONE*, *17*(12), e0277814. https://doi.org/10.1371/journal.pone.0277814

Kuklinski, J. H., Quirk, P. J., Jerit, J., Schwieder, D., & Rich, R. F. (2000). Misinformation and the Currency of Democratic Citizenship. *The Journal of Politics*, *62*(3), 790–816. https://doi.org/10.1111/0022-3816.00033

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kuo, R., & Marwick, A. (2021). Critical disinformation studies: History, power, and politics. *Harvard Kennedy School Misinformation Review*, *2*(4), 1–11. https://doi.org/10.37016/mr-2020-76

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D. (2012). 'The whole is always smaller than its parts' – a digital test of Gabriel Tardes' monads. *The British Journal of Sociology*, *63*(4), 590–615. https://doi.org/10.1111/j.1468-4446.2012.01428.x

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

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, *595*(7866), 189–196. https://doi.org/10.1038/s41586-021-03660-7

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Alstyne, M. V. (2009). Computational Social Science. *Science*, *323*(5915), 721–723. https://doi.org/10.1126/science.1167742

Lazer, D., & Radford, J. (2017). Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*, *43*(1), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457

Lee, H., & Albarracín, D. (2026). Changing beliefs or changing behavior? Understanding the belief-to-behavior process and intervening to curb the impact of misinformation. *Journal of Consumer Psychology*, *36*(1), 114–132. https://doi.org/10.1002/jcpy.70014

Lee, J. H., Santero, N., Bhattacharya, A., May, E., & Spiro, E. S. (2022). Community-based strategies for combating misinformation: Learning from a popular culture fandom. *Harvard Kennedy School Misinformation Review*, *3*(5), 1–12. https://doi.org/10.37016/mr-2020-105

Lee, S., Xiong, A., Seo, H., & Lee, D. (2023). "Fact-checking" fact checkers: A data-driven approach. *Harvard Kennedy School Misinformation Review*, *4*(5), 1–22. https://doi.org/10.37016/mr-2020-126

Levine, T. R. (2014). Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, *33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Lewandowsky, S., Cook, J., & Lombardi, D. (2020). *Debunking Handbook 2020* [Data set]. Databrary. https://doi.org/10.17910/B7.1182

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

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

Li, J., & Yang, X. (2025). Does exposure necessarily lead to misbelief? A meta-analysis of susceptibility to health misinformation. *Public Understanding of Science*, *34*(2), 222–242. https://doi.org/10.1177/09636625241266150

Li, Q. (2019). *Using R for data analysis in social sciences: a research project-oriented approach*. Oxford University Press.

Lien, C. H., Lee, J., & Tandoc, E. C. (2022). Facing Fakes: Understanding Tech Platforms' Responses to Online Falsehoods. *Digital Journalism*, *10*(5), 1–20. https://doi.org/10.1080/21670811.2021.1982398

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.

Lindgren, S. (2020). *Data theory: interpretive sociology and computational methods*. Polity.

Littrell, S., Klofstad, C., Diekman, A., Funchion, J., Murthi, M., Premaratne, K., Seelig, M., Verdear, D., Wuchty, S., & Uscinski, J. E. (2023). Who knowingly shares false political information online? *Harvard Kennedy School Misinformation Review*, *4*(4), 1–32. https://doi.org/10.37016/mr-2020-121

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*, *7*(1), 74–101. https://doi.org/10.1038/s41562-022-01460-1

MacFarlane, D., Tay, L. Q., Hurlstone, M. J., & Ecker, U. K. H. (2021). Refuting Spurious COVID-19 Treatment Claims Reduces Demand and Misinformation Sharing. *Journal of Applied Research in Memory and Cognition*, *10*(2), 248–258. https://doi.org/10.1016/j.jarmac.2020.12.005

Maddox, J., & Malson, J. (2020). Guidelines Without Lines, Communities Without Borders: The Marketplace of Ideas and Digital Manifest Destiny in Social Media Platform Policies. *Social Media + Society*, *6*(2), 2056305120926622. https://doi.org/10.1177/2056305120926622

Makhortykh, M., Sydorova, M., Baghumyan, A., Vziatysheva, V., & Kuznetsova, E. (2024). Stochastic lies: How LLM-powered chatbots deal with Russian disinformation about the war in Ukraine. *Harvard Kennedy School Misinformation Review*, *5*(4), 1–21. https://doi.org/10.37016/mr-2020-154

Malhotra, P., & Pearce, K. (2022). Facing Falsehoods: Strategies for Polite Misinformation Correction. *International Journal of Communication*, *16*, 22.

Mancosu, M., & Vegetti, F. (2020). What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media + Society*, *6*(3), 2056305120940703. https://doi.org/10.1177/2056305120940703

Mandavilli, A. (2011). Peer review: Trial by Twitter. *Nature*, *469*(7330), 286–287. https://doi.org/10.1038/469286a

Mao, T., Su, C. C., Chan, N. K., & Guo, L. (2025). Communicating About and Through Platform Values: Legitimizing Public Expression Regulation on X and Weibo 2007–2024. *Social Media + Society*, *11*(4), 20563051251395398. https://doi.org/10.1177/20563051251395398

Marchal, N., Hoes ,Emma, Klüser ,K. Jonathan, Hamborg ,Felix, Alizadeh ,Meysam, Kubli ,Mael, & and Katzenbach, C. (2025). How Negative Media Coverage Impacts Platform Governance: Evidence from Facebook, Twitter, and YouTube. *Political Communication*, *42*(2), 215–233. https://doi.org/10.1080/10584609.2024.2377992

Marcus, A., & Oransky, I. (2014). What Studies of Retractions Tell Us. *Journal of Microbiology & Biology Education*, *15*(2), 151–154. https://doi.org/10.1128/jmbe.v15i2.855

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political Fact-Checking on Twitter: When Do Corrections Have an Effect? *Political Communication*, *35*(2), 196–219. https://doi.org/10.1080/10584609.2017.1334018

Markham, A. (2012). Fabrication as Ethical Practice. *Information, Communication & Society*, *15*(3), 334–353. https://doi.org/10.1080/1369118X.2011.641993

Markham, A., & Buchanan, E. (2012). *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. http://aoir.org/reports/ethics2.pdf

Marres, N. (2012). The Redistribution of Methods: On Intervention in Digital Social Research, Broadly Conceived. *The Sociological Review*, *60*(1_suppl), 139–165. https://doi.org/10.1111/j.1467-954X.2012.02121.x

Marres, N. (2017). *Digital Sociology: The Reinvention of Social Research*. Polity Press.

Marres, N., & Gerlitz, C. (2016). Interface Methods: Renegotiating Relations between Digital Social Research, STS and Sociology. *The Sociological Review*, *64*(1), 21–46. https://doi.org/10.1111/1467-954X.12314

Marwick, A. E., & boyd, danah. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*(1), 114–133. https://doi.org/10.1177/1461444810365313

McCabe, S. D., Ferrari, D., Green, J., Lazer, D. M. J., & Esterling, K. M. (2024). Post-January 6th deplatforming reduced the reach of misinformation on Twitter. *Nature*, *630*(8015), 132–140. https://doi.org/10.1038/s41586-024-07524-8

McDowell, Z. J., & Vetter, M. A. (2020). It Takes a Village to Combat a Fake News Army: Wikipedia's Community and Policies for Information Literacy. *Social Media + Society*, *6*(3), 2056305120937309. https://doi.org/10.1177/2056305120937309

McKay, S., & Tenove, C. (2021). Disinformation as a Threat to Deliberative Democracy. *Political Research Quarterly*, *74*(3), 703–717. https://doi.org/10.1177/1065912920938143

Mejia, R., Beckermann, K., & Sullivan, C. (2018). White lies: a racial history of the (post)truth. *Communication and Critical/Cultural Studies*, *15*(2), 109–126. https://doi.org/10.1080/14791420.2018.1456668

Memon, S. A., Makovi, K., & AlShebli, B. (2025). Characterizing the effect of retractions on publishing careers. *Nature Human Behaviour*, *9*(6), 1134–1146. https://doi.org/10.1038/s41562-025-02154-0

Mena, P. (2020). Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet*, *12*(2), 165–183. https://doi.org/10.1002/poi3.214

Messing, S., & Westwood, S. J. (2014). Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Communication Research*, *41*(8), 1042–1063. https://doi.org/10.1177/0093650212466406

Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, *3*(1), 2053951716650211. https://doi.org/10.1177/2053951716650211

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics, Biases and Constraints in Communication: Argumentation, Persuasion and Manipulation*, *59*, 210–220. https://doi.org/10.1016/j.pragma.2013.07.012

Micallef, N., He, B., Kumar, S., Ahamad, M., & Memon, N. (2020). The Role of the Crowd in Countering Misinformation: A Case Study of the COVID-19 Infodemic. *2020 IEEE International Conference on Big Data (Big Data)*, 748–757. https://doi.org/10.1109/BigData50022.2020.9377956

Molina, M. D. (2025). Do people believe in misleading information disseminated via memes? The role of identity and anger. *New Media & Society*, *27*(2), 847–870. https://doi.org/10.1177/14614448231186061

Molina, M. D., Wang, J., Sundar, S. S., Le, T., & DiRusso, C. (2022). Reading, Commenting and Sharing of Fake News: How Online Bandwagons and Bots Dictate User Engagement. *Communication Research*, 00936502211073398. https://doi.org/10.1177/00936502211073398

Monsees, L. (2020). 'A war against truth' - understanding the fake news controversy. *Critical Studies on Security*. (world). https://www.tandfonline.com/doi/abs/10.1080/21624887.2020.1763708

Moore, A. (2018). Anonymity, Pseudonymity, and Deliberation: Why Not Everything Should Be Connected. *Journal of Political Philosophy*, *26*(2), 169–192. https://doi.org/10.1111/jopp.12149

Mündges, S., & Park, K. (2024). But did they really? Platforms' compliance with the Code of Practice on Disinformation in review. *Internet Policy Review*, *13*(3). https://policyreview.info/articles/analysis/platforms-compliance-code-of-practice-on-disinformation-review

Munn, L. (2024). Misinformation's missing human. *Media, Culture & Society*, *46*(6), 01634437241249164. https://doi.org/10.1177/01634437241249164

Murphy, G., Laurent, C. de S., Reynolds, M., Aftab, O., Hegarty, K., Sun, Y., & Greene, C. M. (2023). What do we study when we study misinformation? A scoping review of experimental research (2016-2022). *Harvard Kennedy School Misinformation Review*, *4*(6), 1–57. https://doi.org/10.37016/mr-2020-130

Naab, T. K., Kalch, A., & Meitz, T. G. (2018). Flagging uncivil user comments: Effects of intervention information, type of victim, and response comments on bystander behavior. *New Media & Society*, *20*(2), 777–795. https://doi.org/10.1177/1461444816670923

Nan, X., Thier, K., & Wang, Y. (2023). Health misinformation: what it is, why people believe it, how to counter it. *Annals of the International Communication Association*, *47*(4), 381–410. https://doi.org/10.1080/23808985.2023.2225489

Napoli, P. M., & Caplan, R. (2017). Why media companies insist they're not media companies, why they're wrong, and why it matters. *First Monday*, *22*(5). https://doi.org/10.5210/fm.v22i5.7051

Nelimarkka, M. (2022). *Computational thinking and social science: combining programming, methodologies and fundamental concepts* (First). SAGE Publications.

Nieminen, H. (2024). Why Does Disinformation Spread in Liberal Democracies? The Relationship between Disinformation, Inequality, and the Media. *Javnost - The Public*, *31*(1), 123–140. https://doi.org/10.1080/13183222.2024.2311019

Nieminen, S., & Rapeli, L. (2019). Fighting Misperceptions and Doubting Journalists' Objectivity: A Review of Fact-checking Literature. *Political Studies Review*, *17*(3), 296–309. https://doi.org/10.1177/1478929918786852

Nieminen, S., & Sankari, V. (2021). Checking PolitiFact's Fact-Checks. *Journalism Studies*, *22*(3), 358–378. https://doi.org/10.1080/1461670X.2021.1873818

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences*, *118*(15). https://doi.org/10.1073/pnas.1912440117

Nyhan, B., & Reifler, J. (2010). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine*, *33*(3), 459–464. https://doi.org/10.1016/j.vaccine.2014.11.017

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Guess, A. M., Kennedy, E., Kim, Y. M., Lazer, D., Malhotra, N., Moehler, D., Pan, J., Thomas, D. R., Tromble, R., Rivera, C. V., Wilkins, A., Xiong, B., De Jonge, C. K., Franco, A., Mason, W., Stroud, N. J., & Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, *620*(7972), 137–144. https://doi.org/10.1038/s41586-023-06297-w

Oreskes, N., & Conway, E. M. (2011). *Merchants of doubt: how a handful of scientists obscured the truth on issues from tobacco smoke to climate change*. Bloomsbury Press.

Osman, M. (2025). Evidencing the Impact of Misinformed and Disinformed Beliefs on Individual and Group Behaviors. *Psychological Inquiry*, *36*(1), 49–56. https://doi.org/10.1080/1047840X.2025.2482355

Otero, P. V., & Scharlach, R. (2025). 'We do not marshal your feed': How Alt Tech platforms (re) conceptualise safety. *Platforms & Society*, *2*, 29768624251327703. https://doi.org/10.1177/29768624251327703

Özturan, B., Quintana-Mathé, A., Grinberg, N., Ognyanova, K., & Lazer, D. (2025). Declining information quality under new platform governance. *Harvard Kennedy School Misinformation Review*, *6*(3), 1–29. https://doi.org/10.37016/mr-2020-176

Päkkönen, J. (2025). *Credibility-building for social science big data* [Dissertation]. University of Helsinki.

Paquin, R. S., Boudewyns, V., Betts, K. R., Johnson, M., O'Donoghue, A. C., & Southwell, B. G. (2022). An Empirical Procedure to Evaluate Misinformation Rejection and Deception in Mediated Communication Contexts. *Communication Theory*, *32*(1), 25–47. https://doi.org/10.1093/ct/qtab011

Park, K., & Culloty, E. (2022). Beyond Performative Transparency: Lessons Learned from the EU Code Of Practice on Disinformation. *AoIR Selected Papers of Internet Research*. https://doi.org/10.5210/spir.v2022i0.13067

Pasquetto, I., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond, R. M., Bozarth, L. C., Budak, C., Ecker, U. K. H., Fazio, L. K., Ferrara, E., Flanagin, A. J., Flammini, A., Freelon, D., Grinberg, N., Hertwig, R., Jamieson, K. H., Joseph, K., Jones, J. J., Wood, T. J., & Yang, K.-C. (2020). Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School Misinformation Review*, *1*(8), 1–14. https://doi.org/10.37016/mr-2020-49

Pasquetto, I. V., Lim, G., & Bradshaw, S. (2024). Misinformed about misinformation: On the polarizing discourse on misinformation and its consequences for the field. *Harvard Kennedy School Misinformation Review*, *5*(5), 1–8. https://doi.org/10.37016/mr-2020-159

Pearson, G. D. H., Silver, N. A., Robinson, J. Y., Azadi, M., Schillo, B. A., & Kreslake, J. M. (2025). Beyond the margin of error: a systematic and replicable audit of the TikTok research API. *Information, Communication & Society*, *28*(3), 452–470. https://doi.org/10.1080/1369118X.2024.2420032

Peng, H., Romero, D. M., & Horvát, E.-Á. (2022). Dynamics of cross-platform attention to retracted papers. *Proceedings of the National Academy of Sciences*, *119*(25), e2119086119. https://doi.org/10.1073/pnas.2119086119

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), Article 7855. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, The Cognitive Science of Political Thought*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, *20*(1), 293–310. https://doi.org/10.1177/1461444816661553

Popper, K. (1965). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Harper & Row.

Porter, E., Velez, Y., & Wood, T. J. (2022). Factual Corrections Eliminate False Beliefs About COVID-19 Vaccines. *Public Opinion Quarterly*, *86*(3), 762–773. https://doi.org/10.1093/poq/nfac034

Porter, E., Velez, Y., & Wood, T. J. (2023). Correcting COVID-19 vaccine misinformation in 10 countries. *Royal Society Open Science*, *10*(3), 221097. https://doi.org/10.1098/rsos.221097

Prike, T., & Ecker, U. K. H. (2023). Effective correction of misinformation. *Current Opinion in Psychology*, *54*, 101712. https://doi.org/10.1016/j.copsyc.2023.101712

Prior, M. (2007). *Post-broadcast democracy: how media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Quandt, T. (2018). Dark Participation. *Media and Communication*, *6*(4), Article 4. https://doi.org/10.17645/mac.v6i4.1519

Resnik, D. B., Wager, E., & Kissling, G. E. (2015). Retraction policies of top scientific journals ranked by impact factor. *Journal of the Medical Library Association : JMLA*, *103*(3), 136–139. https://doi.org/10.3163/1536-5050.103.3.006

Richardson, R. A. K., Hong, S. S., Byrne, J. A., Stoeger, T., & Amaral, L. A. N. (2025). The entities enabling scientific fraud at scale are large, resilient, and growing rapidly. *Proceedings of the National Academy of Sciences*, *122*(32), e2420092122. https://doi.org/10.1073/pnas.2420092122

Rieder, B., Padilla, A., & Coromina, Ò. (2025). Forgetful by design? A critical audit of YouTube's search API for academic research. *Information, Communication & Society*, 1–20. https://doi.org/10.1080/1369118X.2025.2591767

Riedl, M. J., Naab, T. K., Masullo, G. M., Jost, P., & Ziegele, M. (2021). Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. *Policy & Internet*, *13*(3), 433–451. https://doi.org/10.1002/poi3.257

Roberts, S. T. (2018). Digital detritus: 'Error' and the logic of opacity in social media content moderation. *First Monday*. https://doi.org/10.5210/fm.v23i3.8283

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Robson, C. (2009). *Real world research: a resource for social scientists and practitioner-researchers* (2. ed.). Blackwell.

Rogers, R. (2013). *Digital methods*. The MIT Press.

Rogers, R. (2020). Research note: The scale of Facebook's problem depends upon how 'fake news' is classified. *Harvard Kennedy School Misinformation Review*, *1*(6), 1–14. https://doi.org/10.37016/mr-2020-43

Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering Misinformation. *European Psychologist*, *28*(3), 189–205. https://doi.org/10.1027/1016-9040/a000492

Rossini, P., Stromer-Galley, J., Baptista, E. A., & Veiga de Oliveira, V. (2020). Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections. *New Media & Society*, 1461444820928059. https://doi.org/10.1177/1461444820928059

Sætra, H. S., & Selinger, E. (2024). Technological Remedies for Social Problems: Defining and Demarcating Techno-Fixes and Techno-Solutionism. *Science and Engineering Ethics*, *30*(6), 60. https://doi.org/10.1007/s11948-024-00524-x

Salganik, M. J. (2019). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press. https://press.princeton.edu/books/paperback/9780691196107/bit-by-bit

Saltz, E., Barari, S., Leibowicz, C., & Wardle, C. (2021). Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School Misinformation Review*, *2*(5), 1–25. https://doi.org/10.37016/mr-2020-81

Scharlach, R., Hallinan, B., & Shifman, L. (2023). Governing principles: Articulating values in social media platform policies. *New Media & Society*, 14614448231156580. https://doi.org/10.1177/14614448231156580

Schneider, J. W. (1985). Social Problems Theory: The Constructionist View. *Annual Review of Sociology*, *11*, 209–229.

Schoenebeck, S., & Blackwell, L. (2020). Reimagining Social Media Governance: Harm, Accountability, and Repair In a New Light: Social Media Governance Reconsidered. *Yale Journal of Law and Technology*, *23*(Special 1), 113–152.

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). Metacognitive Experiences and the Intricacies of Setting People Straight: Implications for Debiasing and Public Information Campaigns. In *Advances in Experimental Social Psychology* (Vol. 39, pp. 127–161). Academic Press. https://doi.org/10.1016/S0065-2601(06)39003-X

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173438

Shin, J. (2020). How Do Partisans Consume News on Social Media? A Comparison of Self-Reports With Digital Trace Measures Among Twitter Users. *Social Media + Society*, *6*(4), 2056305120981039. https://doi.org/10.1177/2056305120981039

Shin, J., & Thorson, K. (2017). Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media. *Journal of Communication*, *67*(2), 233–255. https://doi.org/https://doi.org/10.1111/jcom.12284

Sihvonen, T. (2025). Arkista informaatiohäirintää vai luovaa ilmaisua? Tehostetut videosisällöt ja syvävääreennökset TikTokissa. *Media & viestintä*, *48*(1), 59–86. https://doi.org/10.23983/mv.149489

Silverman, C. (2011). The Case for the Corrections Page. *Columbia Journalism Review*. https://www.cjr.org/behind_the_news/the_case_for_the_corrections_page.php

Simon, F. M., & Camargo, C. Q. (2021). Autopsy of a metaphor: The origins, use and blind spots of the 'infodemic'. *New Media & Society*, 14614448211031908. https://doi.org/10.1177/14614448211031908

Singer, J. B. (2021). Border patrol: The rise and role of fact-checkers and their challenge to journalists' normative boundaries. *Journalism*, *22*(8), 1929–1946. https://doi.org/10.1177/1464884920933137

Smith, P., Bansal-Travers, M., O'Connor, R., Brown, A., Banthin, C., Guardino-Colket, S., & Cummings, K. M. (2011). Correcting Over 50 Years of Tobacco Industry Misinformation. *American Journal of Preventive Medicine*, *40*(6), 690–698. https://doi.org/10.1016/j.amepre.2011.01.020

Southwell, B. G., Brennen, J. S. B., Paquin, R., Boudewyns, V., & Zeng, J. (2022). Defining and Measuring Scientific Misinformation. *The ANNALS of the American Academy of Political and Social Science*, *700*(1), 98–111. https://doi.org/10.1177/00027162221084709

Southwell, B. G., & Thorson, E. A. (2015). The Prevalence, Consequence, and Remedy of Misinformation in Mass Media Systems. *Journal of Communication*, *65*(4), 589–595. https://doi.org/10.1111/jcom.12168

Stahl, B. C. (2006). On the Difference or Equality of Information, Misinformation, and Disinformation: A Critical Research Perspective. *Informing Science*, *9*, 83–96. https://doi.org/10.28945/473

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 127:1-127:26. https://doi.org/10.1145/3359229

Steenbuch Traberg, C. (2022). Misinformation: broaden definition to curb its societal influence. *Nature*, *606*(7915), 653–653. https://doi.org/10.1038/d41586-022-01700-4

Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *Academy of Management Review*, *20*(3), 571–610. https://doi.org/10.5465/amr.1995.9508080331

Sun, Y. (2022). Verification Upon Exposure to COVID-19 Misinformation: Predictors, Outcomes, and the Mediating Role of Verification. *Science Communication*, *44*(3), 261–291. https://doi.org/10.1177/10755470221088927

Sundar, S. S., & Nass, C. (2001). Conceptualizing Sources in Online News. *Journal of Communication*, *51*(1), 52–72. https://doi.org/10.1111/j.1460-2466.2001.tb02872.x

Sunstein, C. R. (2017). *#Republic: divided democracy in the age of social media*. Princeton university press.

Suomalainen, K., Nykänen, N., Seeck, H., Kim, Y., & McPherson, E. (2025). Fact-Checking in Journalism: An Epistemological Framework. *Journalism Studies*, *26*(10), 1129–1149. https://doi.org/10.1080/1461670X.2025.2492729

Suzor, N. (2018). Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms. *Social Media + Society*, *4*(3), 2056305118787812. https://doi.org/10.1177/2056305118787812

Swire, B., & Ecker, U. (2018). Misinformation and its Correction: Cognitive Mechanisms and Recommendations for Mass Communication. In B. G. Southwell, E. A. Thorson, & L. Sheble (Eds), *Misinformation and Mass Audiences*. University of Texas Press.

Swire-Thompson, B., & Lazer, D. (2022). Reducing Health Misinformation in Science: A Call to Arms. *The ANNALS of the American Academy of Political and Social Science*, *700*(1), 124–135. https://doi.org/10.1177/00027162221087686

Sybert, J. (2022). The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *New Media & Society*, *24*(10), 2311–2331. https://doi.org/10.1177/1461444821996715

Szalinski, C. (2021, September 29). Fringe Doctors' Groups Promote Ivermectin for COVID despite a Lack of Evidence. *Scientific American*. https://www.scientificamerican.com/article/fringe-doctors-groups-promote-ivermectin-for-covid-despite-a-lack-of-evidence/

Taber, C. S., & Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, *50*(3), 755–769. https://doi.org/https://doi.org/10.1111/j.1540-5907.2006.00214.x

Takahashi, T., & Beebe, L. M. (1993). Cross-Linguistic Influence in the Speech Act of Correction. In G. Kasper & S. Blum-Kulka (Eds), *Interlanguage Pragmatics* (pp. 138–158). Oxford University Press.

Tan, A. S. L., Lee, C., & Chae, J. (2015). Exposure to Health (Mis)Information: Lagged Effects on Young Adults' Health Behaviors and Potential Pathways. *Journal of Communication*, *65*(4), 674–698. https://doi.org/https://doi.org/10.1111/jcom.12163

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "Fake News": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Tandoc, E. C., Ling, R., Westlund, O., Duffy, A., Goh, D., & Zheng Wei, L. (2018). Audiences' acts of authentication in the age of fake news: A conceptual framework. *New Media & Society*, *20*(8), 2745–2763. https://doi.org/10.1177/1461444817731756

Tang, R., Vraga, E. K., Bode, L., & Boulianne, S. (2024). Who reports witnessing and performing corrections on social media in the United States, United Kingdom, Canada, and France? *Harvard Kennedy School Misinformation Review*, *5*(3), 1–27. https://doi.org/10.37016/mr-2020-145

Tay, L. Q., Lewandowsky, S., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2024). Thinking clearly about misinformation. *Communications Psychology*, *2*(1), 1–5. https://doi.org/10.1038/s44271-023-00054-5

The YouTube Team. (2019, December 3). The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation [Blog]. *YouTube Official Blog*. https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/

Thelwall, M., Lehtisaari, M., Katsirea, I., Holmberg, K., & Zheng, E.-T. (2025). Does ChatGPT Ignore Article Retractions and Other Reliability Concerns? *Learned Publishing*, *38*(4), e2018. https://doi.org/10.1002/leap.2018

Thorson, K., & Wells, C. (2016). Curated Flows: A Framework for Mapping Media Exposure in the Digital Age. *Communication Theory*, *26*(3), 309–328. https://doi.org/10.1111/comt.12087

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, *119*(42), e2207159119. https://doi.org/10.1073/pnas.2207159119

Törnberg, P., & Uitermark, J. (2021). For a heterodox computational social science. *Big Data & Society*, *8*(2), 20539517211047725. https://doi.org/10.1177/20539517211047725

Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review*, *4*(3), 356–367. https://doi.org/10.1177/1534484305278283

Torres, R., Gerhart, N., & Negahban, A. (2018). Epistemology in the Era of Fake News: An Exploration of Information Verification Behaviors among Social Networking Site Users. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, *49*(3), 78–97. https://doi.org/10.1145/3242734.3242740

Trezza, D. (2023). To scrape or not to scrape, this is dilemma. The post-API scenario and implications on digital research. *Frontiers in Sociology*, *8*. https://doi.org/10.3389/fsoc.2023.1145038

Trikalinos, N. A., Evangelou, E., & Ioannidis, J. P. A. (2008). Falsified papers in high-impact journals were slow to retract and indistinguishable from nonfraudulent papers. *Journal of Clinical Epidemiology*, *61*(5), 464–470. https://doi.org/10.1016/j.jclinepi.2007.11.019

Tromble, R., & McGregor, S. C. (2019). You Break It, You Buy It: The Naiveté of Social Engineering in Tech – And How to Fix It. *Political Communication*, *36*(2), 324–332. https://doi.org/10.1080/10584609.2019.1609860

Tsipursky, G., Votta, F., & Roose, K. M. (2018). Fighting Fake News and Post-Truth Politics with Behavioral Science: The Pro-Truth Pledge. *Behavior and Social Issues*, *27*(1), 47–70. https://doi.org/10.5210/bsi.v27i0.9127

Uscinski, J., & Butler, R. (2013). The Epistemology of Fact Checking. *Critical Review*, *25*(2), 162–180. https://doi.org/10.1080/08913811.2013.843872

Uscinski, J., & Enders, A. M. (2023). What Is a Conspiracy Theory and Why Does It Matter? *Critical Review*, *35*(1–2), 148–169. https://doi.org/10.1080/08913811.2022.2115668

Uscinski, J., Enders, A. M., Klofstad, C., & Stoler, J. (2022). Cause and effect: On the antecedents and consequences of conspiracy theory beliefs. *Current Opinion in Psychology*, *47*, 101364. https://doi.org/10.1016/j.copsyc.2022.101364

Uscinski, J., Littrell, S., & Klofstad, C. (2024). The importance of epistemology for the study of misinformation. *Current Opinion in Psychology*, *57*, 101789. https://doi.org/10.1016/j.copsyc.2024.101789

Valinciute, A., & Halffman, W. (2025). Do Journalists Update Retracted Science News? *Journalism Practice*, (0), 1–23. https://doi.org/10.1080/17512786.2025.2540460

van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 1–8. https://doi.org/10.1038/s41591-022-01713-6

Van Der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. (2023). *Using Psychological Science To Understand And Fight Health Misinformation: An APA Consensus Statement* [Data set]. https://doi.org/10.1037/e506432023-001

van Dijck, J., Poell, T., & Waal, M. de. (2018). *The Platform Society: Public Values in a Connective World*. Oxford University Press.

van Doorn, M. (2023). Advancing the debate on the consequences of misinformation: clarifying why it's not (just) about false beliefs. *Inquiry*, 1–27. https://doi.org/10.1080/0020174X.2023.2289137

van Es, K., Wieringa, M., & Schäfer, M. T. (2018). Tool Criticism: From Digital Methods to Digital Methodology. *Proceedings of the 2nd International Conference on Web Studies*, 24–27. https://doi.org/10.1145/3240431.3240436

Van Leeuwen, T. (2007). Legitimation in discourse and communication. *Discourse & Communication*, *1*(1), 91–112. https://doi.org/10.1177/1750481307071986

Venturini, T., Bounegru, L., Gray, J., & Rogers, R. (2018). A reality check(list) for digital methods. *New Media & Society*, *20*(11), 4195–4217. https://doi.org/10.1177/1461444818769236

Venturini, T., & Rogers, R. (2019). "API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. *Digital Journalism*, *7*(4), 532–540. https://doi.org/10.1080/21670811.2019.1591927

Vinhas, O., & Bastos, M. (2022). Fact-Checking Misinformation: Eight Notes on Consensus Reality. *Journalism Studies*, *23*(4), 448–468. https://doi.org/10.1080/1461670X.2022.2031259

Vinhas, O., & Bastos, M. (2025). The WEIRD governance of fact-checking and the politics of content moderation. *New Media & Society*, *27*(5), 2768–2787. https://doi.org/10.1177/14614448231213942

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

Vraga, E. K., & Bode, L. (2017). Using Expert Sources to Correct Health Misinformation in Social Media. *Science Communication*, *39*(5), 621–645. https://doi.org/10.1177/1075547017731776

Vraga, E. K., & Bode, L. (2020a). Correction as a Solution for Health Misinformation on Social Media. *American Journal of Public Health*, *110*(S3), S278–S280. https://doi.org/10.2105/AJPH.2020.305916

Vraga, E. K., & Bode, L. (2020b). Defining Misinformation and Understanding its Bounded Nature: Using Expertise and Evidence for Describing Misinformation. *Political Communication*, *37*(1), 136–144. https://doi.org/10.1080/10584609.2020.1716500

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, *595*(7866), 197–204. https://doi.org/10.1038/s41586-021-03666-1

Wahl-Jorgensen, K., & Carlson, M. (2021). Conjecturing Fearful Futures: Journalistic Discourses on Deepfakes. *Journalism Practice*, *15*(6), 803–820. https://doi.org/10.1080/17512786.2021.1908838

Walter, N., Brooks, J. J., Saucier, C. J., & Suresh, S. (2021). Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. *Health Communication*, *36*(13), 1776–1784. https://doi.org/10.1080/10410236.2020.1794553

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Walter, N., Edgerly, S., & Saucier, C. J. (2021). "Trust, Then Verify": When and Why People Fact-Check Partisan Information. *International Journal of Communication*, *15*(0), Article 0.

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564

Walter, N., & Salovich, N. A. (2021). Unchecked vs. Uncheckable: How Opinion-Based Claims Can Impede Corrections of Misinformation. *Mass Communication and Society*, *0*(0), 1–28. https://doi.org/10.1080/15205436.2020.1864406

Walter, N., & Tukachinsky, R. (2020). A Meta-Analytic Examination of the Continued Influence of Misinformation in the Face of Correction: How Powerful Is It, Why Does It Happen, and How to Stop It? *Communication Research*, *47*(2), 155–177. https://doi.org/10.1177/0093650219854600

Wardle, C. (2023). Misunderstanding Misinformation. *Issues in Science and Technology*, *39*(3), 38–40. https://doi.org/10.58875/ZAUD1691

Wardle, C., & Derakshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (Council of Europe Report DGI(2017)09). Council of Europe. https://www.coe.int/en/web/freedom-expression/information-disorder

Waruwu, B. K., Tandoc, E. C., Duffy, A., Kim, N., & Ling, R. (2020). Telling lies together? Sharing news as a form of social authentication. *New Media & Society*, 1461444820931017. https://doi.org/10.1177/1461444820931017

Weeks, B. E., & Gil de Zúñiga, H. (2021). What's Next? Six Observations for the Future of Political Misinformation Research. *American Behavioral Scientist*, *65*(2), 277–289. https://doi.org/10.1177/0002764219878236

Weigel, M., & Gitomer, A. (2025). Hate-sharing: A case study of its prevalence and impact on Gab. *New Media & Society*, *27*(9), 5022–5044. https://doi.org/10.1177/14614448241245349

Wenger-Trayner, É. (1998). *Communities of practice: learning, meaning, and identity*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803932

West, J. D., & Bergstrom, C. T. (2021). Misinformation in and about science. *Proceedings of the National Academy of Sciences*, *118*(15). https://doi.org/10.1073/pnas.1912444117

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. *Sociology*, *51*(6), 1149–1168. https://doi.org/10.1177/0038038517708140

Williams, P., & Wager, E. (2013). Exploring Why and How Journal Editors Retract Articles: Findings From a Qualitative Study. *Science and Engineering Ethics*, *19*(1), 1–11. https://doi.org/10.1007/s11948-011-9292-0

Wittenberg, C., & Berinsky, A. J. (2020). Misinformation and Its Correction. In N. Persily & J. A. Tucker (Eds), *Social Media and Democracy: The State of the Field, Prospects for Reform* (pp. 163–198). Cambridge University Press. https://doi.org/10.1017/9781108890960

Wood, T., & Porter, E. (2019). The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Woolley, S. C., & Howard, P. N. (Eds). (2018). *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press. https://doi.org/10.1093/oso/9780190931407.001.0001

Wu-Ouyang, B. (2026). Curation as a communicative act: conceptualizing personal curation within curated flows on social media. *Communication Theory*, qtag002. https://doi.org/10.1093/ct/qtag002

Yu, W., & Shen, F. (2023). Mapping verification behaviors in the post-truth era: A systematic review. *New Media & Society*, 14614448231191138. https://doi.org/10.1177/14614448231191138

Zeng, J., & Kaye, D. B. V. (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287

Zhao, P., Bazarova, N. N., DiFranzo, D., Hui, W., Kizilcec, R. F., & Margolin, D. (2024). Standing up to problematic content on social media: which objection strategies draw the audience's approval? *Journal of Computer-Mediated Communication*, *29*(1), zmad046. https://doi.org/10.1093/jcmc/zmad046

Zimmer, M. (2018). Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity. *Social Media + Society*, *4*(2), 2056305118768300. https://doi.org/10.1177/2056305118768300

Zollo, F., Bessi, A., Vicario, M. D., Scala, A., Caldarelli, G., Shekhtman, L., Havlin, S., & Quattrociocchi, W. (2017). Debunking in a world of tribes. *PLOS ONE*, *12*(7), e0181821. https://doi.org/10.1371/journal.pone.0181821