FACT-CHECKING AS A TOOL TO TACKLE HATE SPEECH: A SYSTEMATIC LITERATURE REVIEW

Juliana da Cunha Mota[1]

Cite as: Juliana da Cunha Mota. 'Fact checking as a tool to tackle hate speech: A systematic literature review' in Eliza Bechtold, Clara Burbano-Herrea, and Zoë Gross (eds), *Echoes of Hate. Digital communication, Populism, and the Regulation of Hate Speech,* (Human Rights in Context, Ghent University, 2025).

Abstract: Hate speech and misinformation are often connected, with misinformation often being used to support, justify, or fuel hateful content, hence creating populist hate narratives. In this context, if fact-checkers initiatives seek to curb the spread of misinformation, what is their role in curbing hate speech? This paper explores hate speech and fact-checkers and seeks to answer the following question: What does the literature (2015-2025) report about how collaborations between fact-checkers and platforms are associated with the prevalence of populist hate narratives? Through a systematic literature review, I argue that the existing evidence supports a normative case for sustaining FC–platform collaborations. While empirical proof of FC-platforms collaborations efficacy remains limited, their alignment with freedom of expression principles and their potential to curb misinformation-induced hate speech justify their continued promotion in human rights frameworks.

1. INTRODUCTION

Social media platforms were originally perceived as a tool for democratising access to knowledge and information. Especially from the 2010s, however, they have become a particularly fertile ground for propagation of mis- and disinformation and hate speech[2] (Matamoros-Fernández & Farkas, 2021; Schradie, 2019).

Hate speech and misinformation are two distinct phenomena that cause different harms (Heldt, 2019). Yet, they are often connected: during crises, misinformation can support, justify, or fuel hateful content – creating what we call 'populist hate narratives' (Erjavec & Kovačič, 2012; Mayagoitia-Soria et al., 2024; Poole et al., 2021).

Various tools and initiatives are employed to tackle hate speech and misinformation. AI and machine learning are increasingly used in the fight against hate speech (Jahan & Oussalah, 2023). To tackle mis- and disinformation, fact-checkers (FCs) play a crucial role (Adam, 2025; Graves, 2017;

[1] Digital Policy Postdoctoral Researcher at the Centre for Socio-Legal Studies of the University of Oxford.

[2] Admittedly, there is no universally accepted definition of hate speech and an investigation on its precise meaning falls outside of the scope of this paper. For an overview of the challenges and possibilities of defining hate speech, see (Hietanen & Eddebo, 2023). 'Hate speech' is used in this paper as any expressions which fall under one of these three categories: negative stereotyping, dehumanisation, and expressions of violence or hostility towards a defined group. This definition is broader, yet in line with that of the Council of Europe (CoE), which defines hate speech in specific reference to protected characteristics. As such, the expressions that '*incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as "race", colour, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation*'(CM/Rec(2022)16 - Recommendation of the Committee of Ministers to Member States on Combating Hate Speech, 2022)

Westlund et al., 2024). Given FCs' importance in reducing online disinformation, one could reasonably assume that they should also contribute to reducing the spread of hateful content. Nonetheless, the role of FCs in tackling hate speech through fact-checks of disinformation remains underexplored by the literature.

This paper seeks to assess the existing evidence on the role of FCs in tackling the spread of hate speech on platforms. Through a systematic literature review and critique of academic articles on hate speech and fact-checking published within the past decade, the paper aims to answer the following research question: What does the literature (2015-2025) report about how collaborations between fact-checkers and platforms are associated with the prevalence of populist hate narratives?

To answer this question, I performed a qualitative analysis of published material. I reviewed the literature to assess the existing evidence and scholars' perceptions of the following issues: (1) the relationship between hate speech and disinformation, (2) the role of FCs in tackling hate speech; (3) the role of platforms in facilitating or curbing the dissemination of hateful narratives.

By delving into the role of FCs for tackling hate speech, this paper makes three contributions. First, by identifying the gaps in the scholarship, it assesses areas that can benefit from further research. Second, by expanding the results of this study, it considers the consequences of scaling down on collaborations between platforms and FCs. This is a particularly timely contribution in light of Meta's decision to end the 3PFC in the US (Kaplan, 2025). While questions remain if Meta will replicate the measure elsewhere, the results of this analysis can provide useful insights into the impact of this measure. Finally, the results will guide broader legal discussions, contributing to the growing debate on States' obligations concerning electoral disinformation and/or positive obligations to tackle hate speech (Alkiviadou, 2025; Pentney & Shattock, 2025).

This paper proceeds as follows: Section 2 starts by expanding on the existing literature on the role of FCs and FC-platforms collaborations. Section 3 presents the method employed (systematic literature review) and its justification, further explaining the data collection. Section 4 contains my main findings. Finally, Section 5 expands the findings to propose broader discussions guided by the discussion question.

2. FACT-CHECKERS AND COLLABORATIONS BETWEEN FCS AND PLATFORMS

FC emerged as a global, yet fragmented, movement. FC firms emerged in the 1990s to hold politicians and other public figures accountable for false statements. Originally incubated in US newsrooms, the movement grew and changed as social media platforms challenged journalists' gate-keeping role and mis- and disinformation became more prevalent. As the literature indicates, 'In response to disinformation campaigns during the 2016 US presidential election, the field's focus shifted from verifying claims made by politicians to policing viral misinformation on digital platforms – the debunking turn'(Mahl et al., 2024, p. 3).

With the debunking turn, FC is now performed by different actors, including but not limited to professional journalists, political campaigns and party organisations, and third sector organisations. Each of these actors have different perceptions about the role of FCs, performing it by different standards and methods (Cavaliere, 2020, p. 158). Despite these differences, fact-checkers have somewhat similar goals: to verify information presented as a fact. They do not verify statements of opinion or statements which cannot be corroborated. Other similarities involve their areas of practice, which usually comprise: choosing claims to check, contacting the speaker, tracing false claims, dealing with experts, and showing your work (Graves, 2017). FCs' work supports both individual readers and content moderation on large-scale platforms through labelling mechanisms (Sehat et al., 2024).

Acknowledging the importance of FCs, platforms started collaborating with FC organisations in the last decade. For instance, Meta instituted its third-party fact-checking programme ('3PFC') in 2016 'to reduce the spread of misinformation and provide more reliable information to users' (Bengtsson et al., 2025, p. 249). Similarly, Google has partnered with FCs to develop data standards to surface fact-checks in search results. Some FC organisations choose to join these partnerships to amplify the reach of their content (Bélair-Gagnon et al., 2023).

The importance of FC-platforms collaborations cannot be overstated. Some FC organisations exist and operate thanks to platform partnerships. However, in January 2025, Meta announced it was scaling down on these partnerships in the US. The measure was taken under the guise of strengthening the protection to freedom of expression, allegedly threatened by FCs who were, in Mark Zuckerberg's words, '*too politically biased'*(Kaplan, 2025). In FCs' view, the information ecosystem of several countries could be affected should Meta decide to expand its decision to scale back FC collaborations globally. This is because Facebook and Instagram are the main sources of news in countries from the global south. Thus, the lack of fact-checking initiatives could increase the circulation of misleading information (Kahn, 2025).

Scholars in media studies have explored how FCs and platforms conceive making trade-offs in their platforms to build their fields as agents of knowledge. They found that platforms and FCs are marked by both asymmetric and mutual dependence. They also recognise the need to acknowledge the inter-relational and dynamic element of platform companies, and how they relate to a larger platform and information ecosystem (Bélair-Gagnon et al., 2023). While explored in other areas of knowledge, the relationship between FCs and platforms remains underexplored by legal and socio-legal scholars.

3. METHOD

To explore FC-platforms collaborations, I employed a systematic literature review of the existing works on the topic. The aims of this work justify choosing standalone systematic literature reviews (SSLR) as a method. Specifically, this paper aims at (1) answering the research question based on the existing scholarly evidence; and (2) identifying gaps in the literature. These are precisely the same aims as those of SSLR (Okoli, 2015; Petticrew & Roberts, 2008). Furthermore, unlike scoping reviews, systematic SSLR are somewhat targeted. For instance, they may aim at identifying the evidence on the efficacy of a specific intervention, instead of broadly exploring the characteristics of an area of knowledge (López-Borrull & Lopezosa, 2025). Thus, SSLR is in line with the aims of this paper.

Importantly, SSLR is not valuable in the early days of a research field, which further justifies its usage in this specific research. In this paper, I assess a decade of the body of literature on misinformation, fact-checking, and hate speech. The choice for setting 2015 as the cutoff date for this analysis is twofold. First, the Trump election and Brexit happened in 2016 and 2017, respectively. These two episodes triggered debates into the role of platforms in spreading mis- and disinformation, especially during electoral periods. Second, Meta established the 3PFC in 2016 and the programme grew bigger in the following years. Accordingly, analysing ten years of corpus literature into misinformation, hate speech, and fact-checking from 2015 to 2025 might reveal how scholars perceived these topics in light of societal and technological developments, if at all.

SSLR require a systematic methodological approach, explicit explanation of the procedures followed, and a comprehensive scope of the material (Okoli, 2015). Below, I provide an overview of the methodical approach adopted and the procedures followed.

3.1. Data collection

The first step into the data collection phase consists of choosing the appropriate database of academic works. In line with existing literature, I chose to extract works from Scopus, given its prestige and quality of papers (López-Borrull & Lopezosa, 2025). Admittedly, choosing one database limits the results of this paper. Nonetheless, due to time and scope constraints, performing an in-depth analysis of other databases was not feasible.

After selecting the database, I defined the search terms. I performed several searches with the following keywords:

Search	Keywords
1	'hate speech' and 'fact-checking'
2	'hate speech' and 'fact-checkers'
3	'hate speech' and 'fact-checkers' and 'social media platforms'
4	'hate speech' and 'fact-checkers' and 'platforms'
5	'hate' and 'fact-checking'
6	'hate' and 'fact-checkers'
7	'hate' and 'fact checkers' and 'platforms'
8	'hate speech' 'misinformation' and 'elections'[3]
9	'hate speech' and 'fake news' and 'elections'
10	'hate speech' and 'fake news' and 'platforms'
11	'hate speech' and 'populism' and 'platforms'
12	'hate speech' and 'populism'
13	'hate speech' and 'misinformation'

I considered solely papers published from 2015 to 2025, for the reasons mentioned in the section above. As illustrated by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram shown in Figure 1, our searches returned n=3,375 works in total (Gibson et al., 2025; Liberati et al., 2009).

I uploaded the full database of papers into Ryann.AI, a software for systematic literature review. I removed 1,637 papers which were duplicated, after which I conducted the first round of revision. In this round, I considered solely the title of the papers, the abstract, and the key words. I excluded (i) books and chapters; (ii) conference proceedings. Furthermore, I also did not consider papers that (i) focus on countries outside of Europe; (ii) in languages other than English; (iii) which had not been peer-reviewed; and (iv) focused on the role of traditional media or dark web forums in disseminating hate speech, without further consideration to social media platforms. This round resulted in the exclusion of 1,686 works. Subsequently, I performed a second round of revision, mainly to exclude book chapters and conference proceedings. After the second round of exclusion, I had a preliminary database of works constituted of 51 papers.

All these 51 papers were downloaded and assessed. Some papers were subsequently excluded as for similar reasons as those previously exposed (i.e. they focused on the role of traditional media, delved into content moderation issues through removal of content and not fact-checking, or concerned models for fact-checking, mis- or disinformation, without exploring hate speech issues or

[3] The choice for 'election' as keyword is justified as the research was originally designed to explore populist hate narratives during electoral periods. However, few papers were published solely on this subject, significantly restricting the conclusions and justifying a subsequent change in the research design and question. Instead of focusing on electoral periods, the paper considers populist hate narratives in both electoral and non-electoral periods.

the role of platforms). Admittedly, the exclusion of grey literature may have led to the omission of potentially relevant works. However, this limitation does not compromise the validity of the study, as the author believes theoretical saturation was reached with the papers included. Finally, the database of examined papers included 32 works. A full list of papers considered is provided in the Appendix A.

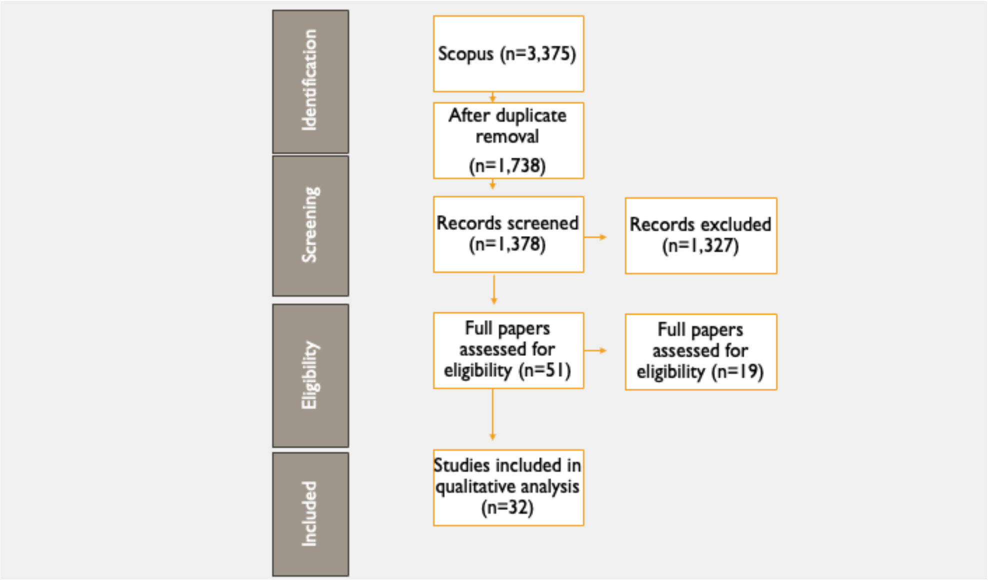

Identification
Screening
Eligibility
Included
Scopus (n=3,375)
After duplicate removal (n=1,738)
Records screened (n=1,378)
Records excluded (n=1,327)
Full papers assessed for eligibility (n=51)
Full papers assessed for eligibility (n=19)
Studies included in qualitative analysis (n=32)

Figure 1: PRISMA flowchart

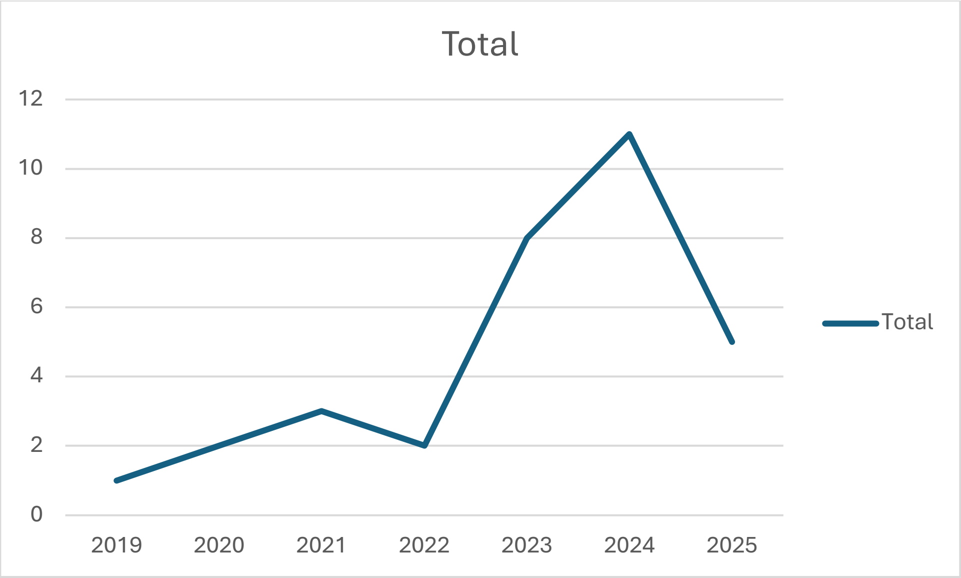
4. Findings

The analysis of the papers revealed insights about the type of analysis being conducted in the fields of hate speech, misinformation, FCs, and platforms. Below, I present some of these findings.

4.1. Growing interest in the topic

There appear to be growing interest for the topics of hate speech, misinformation, platforms, and FCs. No eligible studies appear prior to 2019, with outputs rising and peaking in 2024.

Some political and social factors might explain the surge in papers from 2022 onwards. First, as previously mentioned, platforms' actions and omissions to tackle misinformation came to the centre stage from 2016 onwards, with the Trump election, the Brexit Referendum, and the Cambridge Analytica scandal. Subsequently, the COVID-19 further highlighted the role of platforms in tackling hate speech and misinformation.



4.2. Empirical vs theoretical analyses

Most of the papers consisted of empirical analyses of the issues, often adopting quantitative and qualitative methods to assess harmful content. Theoretical works are scarce, fewer than 10 papers. Yet, these theoretical contributions provided invaluable insights which should be accounted for.

Most of the empirical papers focused on one specific type of hate speech and platform for analysis. For instance, Asardag (2025) employed critical feminist lens to explore the discourses around Meta's changes to their hate speech policies, announced in January 2025. She found that the changes would likely negatively impact LGBTQAI+ people, consistently targeted by online hate speech and disinformation.

Other papers empirically analysed disinformation and hate speech. These studies are crucial for understanding how hate speech spreads within different contexts and platforms. Conversely, as Liu et al. had previously noticed, most of these studies are limited to specific cases, and there is a lack of longitudinal and generalised study of hate speech (Liu et al., 2024). Understandably, hate speech is context-dependent, thus it is questionable whether an overarching empirical study would produce significant and adequate results. For instance, Carrasco-Farré (2022) conducted a computational linguistics assessment of misinformation and harmful content, Cinelli et al. (2021) considered hate comments on YouTube, and Dowining & Dron (2020) focused on discourse analyses of tweets.

While in the minority, theoretical papers were also important for drawing connections between disinformation and hate speech. For instance, Cepollaro et al.'s work theoretically assessed efficacy and deontic questions of counterspeech for tackling harmful messages. Ultimately, they argued that philosophers should focus more on the efficacy question to advocate for better frameworks to conceptualise the efficacy of counter speech. A paper linking these philosophical views and empirical works was, however, missing from our database of papers (Cepollaro et al., 2023).

4.3. Platforms investigated

Most of the empirical papers focused solely on some platforms for analysis. The most popular platform was X (formerly Twitter), which accounted for 9 papers, followed by Facebook (with 7 papers), TikTok (3 papers), and YouTube (2 papers).

Twitter's API for research purposes might explain the platform's popularity among researchers. Until 2023, the platform was 'a playground for academic research', providing access to millions of tweets every day for free. After Twitter's acquisition by Elon Musk, the policies changed and access to the API is now behind a paywall (Sarah Grevy Gotfredsen, 2023). While there were four papers published in 2024 using Twitter's data, I found only 2 in 2025. This raises questions whether we will continue to see a decrease in works with Twitter's data because of Musk's new policies.

Another interesting point concerns papers analysing TikTok content. Scholarly attention to the platform did not appear to accompany its surge in users and importance (Pérez Rastrilla et al., 2023). It is unclear why scholars paid so little attention to TikTok and whether this platform will appear more prominently in future studies.

4.4. Exploring links between hate speech and disinformation

Not all hateful messages contain disinformation, and vice versa. Yet, numerous studies recognised the connections between disinformation, hate speech, radicalisation, and polarisation.

For instance, according to Carrasco-Farré's (2022) computational linguistics study, false information and hate speech share some commonalities, including the negativity of the content and their elevated appeal to moral values. Similarly, Mohsen et al. (2024) applied linear regression models with standardised coefficients to assess a database of more than 8 million tweets from N=6,832 Twitter users, revealing a link between misinformation and harmful language. Echoing this, Zhou et al. (2025) found that 28% of the hate speech instances against Asians contained misinformation.

Conversely, some studies found no relation between hate language and misinformation communities. For instance, Cinelli et al. (2021) did not find any correlation between the usage of hate language by YouTube users and their involvement in misinformation communities in the platform. Downing and & Dron (2020) echoed the finding that social media was not a conduit for hateful or deceitful messages. Their analysis focused on Tweets about the Grenfell fire and the Muslim community, and they found that the most influential Tweets pictured Muslims in a positive light. Poole et al. had similar findings when analysing tweets containing the hashtag #StopIslam. In their study, the most shared tweets contained counter-narrative and positive messages about Islam instead of hateful content(Poole et al., 2021).

In summary, with some exceptions, I found evidence in the existing literature that there are links between misinformation and hate speech online – the extent of these links is, however, disputed.

4.5. Exploring the role of FCs in tackling hate speech

I found few papers acknowledging the importance of FCs or fact-checking techniques in tackling misinformation and hateful content online. For instance, Miškolci et al. (2020) adopted a quasi-experimental research design to dissect the spread of anti-Roma discourses on Facebook. In the study, the authors adopted the role of FCs to assess the consequences of entering pro-Roma comments, often checking facts presented in posts targeting this minority. They found a positive correlation between their positive comments and new users' interventions. Put differently, when they made posts with pro-Roma comments, it motivated 'other followers of those particular Facebook profiles to join the discussion arguing in favour of the Roma as well'(Miškolci et al., 2020, p. 139). Conversely, they acknowledge that fact-checking as a counterspeech strategy showed limitations, as most of the users continued to post hateful comments within the thread, while others rarely acknowledged the argument from their intervention. Finally, the authors highlighted that emotions spread through Facebook might have also contributed to users' comments and reactions. As such, the authors concluded that platforms' structural design may contribute to creating communities of hate and a revision of this business model is needed to tackle hate speech (Miškolci et al., 2020).

While some papers acknowledged the importance of fact-checking in tackling hate narratives, Munn (2024) highlighted some limitations to this approach. According to him, the existing approaches to tackle misinformation (including FCs) are based an idealised rational version of humans, which fails to consider that humans are primarily, rational, factional, and bigoted. Accordingly, they argued that more information will not solve the misinformation issue. Instead, the author proposed linking fact-checking approaches to psychological research to robustly connect human insights into real-world interventions.

4.6. Exploring the role of platforms in disseminating or curbing hate narratives

Finally, some papers addressed the role of platforms in curbing or disseminating disinformation and populist hate narratives. For instance, Munn (2024) highlighted that humans are emotional beings, often acting driven by personal and/or moral values rather than by objective evidence. Online platforms, in turn, capture and amplify this emotion through algorithms that privilege controversial or attention-seeking content – including populist and hate narratives.

Silva & Parker (2025) argued that platforms' structural power in speech dissemination impedes their regulation. According to them, platforms use discursive power to maintain control over the conditions for their own regulation. In other words, platforms discursively claim that they constitute the ones, or most legitimate ones, who should serve as the arbiters of the speech they host. Consequently, they are central in disseminating or curbing hate speech. Heldt (2019) echoes this.

Echoing the argument that platforms play an active role in choosing which content to disseminate, Abdul Reda and Alkhonin (2025) found that 'algorithms could be adjusted (…). This strategic shift could reduce the space available for misinformation to spread, aligning public focus with verified informative content continuously' (Abdul Reda & Alkhonin, 2025, p. 18). Similarly, Silva & Parker (2025) claimed that platform power can directly contribute to the proliferation of hate speech when they fail to remove content. In doing so, they proposed a typology, building on Fuchs (2013) for platforms' actions in contributing to hate speech: amplification, accommodation, or discursive authorisation of hate speech.

The analysis of platforms' terms and services further highlighted their active role in curbing hate narratives (Arora et al., 2024). Despite this, several studies found hateful content in the platforms analysed. For instance, González-Aguilar et al. (2023) found few hateful messages on TikTok in a study assessing populist speech on the platform. The authors highlighted that hate speech played no particular influence on the videos' views (González-Aguilar et al., 2023). While somewhat surprising, the authors acknowledged one important caveat to their study: TikTok might have played an important role in downgrading videos with hateful content, thus affecting the videos' views. Therefore, the platform might have become a tool and strategy in downplaying harmful content.

Poole et al. (2021) echoed the argument that platforms play a crucial role in disseminating or curbing populist hate narratives. They argued that Twitter played an active agent role in setting the narrative surrounding the use of the hashtag #StopIslam. This is because Twitter (now X) had removed a significant number of tweets containing this contested hashtag. The authors also found that Twitter reinforced the dominance of the 'elite', by using algorithms that amplified the visibility and reach of content posted by influencers. Thus, they claimed that social media platforms' dynamics afforded more agency to well-organised groups with stronger ties. In their study, these groups were aligned with the right, fuelled by populist discourse. Accordingly, they concluded that 'What is particularly concerning about #stopIslam, therefore, is that it illustrates how the strategies of social media platforms can create conditions that lend themselves not just to the actions, but ideological commitments of right-wing populist groups' (Poole et al., 2021, p. 1438).

While I did not find studies generally disputing platforms' role in spreading hate speech, I found at least one study challenging pre-established conceptions about the dissemination of harmful speech. Among others, Budak et al. (2024) challenged three media claims. First, that disinformation is growing in the age of social media and that more people are exposed to it. Second, that exposure to misinformation and harmful content is primarily driven by the algorithms of platforms. Finally, that there are correlations between exposure to harmful content online and undesirable psychological or behavioural effects. According to them, the existing studies on online harmful speech do not support any of these conclusions. In fact, they argue that the literature demonstrates that the exposure to misinformation is low, often concentrated among a small minority. Similarly, they contend that users are not fed harmful content through 'filter bubbles' – instead, users who are already attentive to this kind of content harmful content seek it out across mediums.

In light of the above, the existing evidence on the scholarly literature points to a link between the platforms and the dissemination of hateful messages. If platforms' architecture contributes to the dissemination of populist discourses and hate speech, the question is how FCs impact this ecosystem, if at all. Sehat et al. (2024)'s study found that platforms indicated which content they wanted prioritised for fact-checking. (Sehat et al., 2024). Thus, platforms could arguably indicate the need to have populist hate narratives checked.

In sum, it is evident that platforms actively try to curtail hate speech, which often serves as a basis for populist narratives. However, their approach is arguably different when it comes to mis- and disinformation, with platforms historically taking a stance in favour of a free marketplace of ideas. Analysing the T&Cs of 42 platforms, Arora et al. (2024) found that at least 10 of them did not address misinformation. Furthermore, the sum of explicit mentions to misinformation in all T&Cs analysed taken together was significantly narrower (N=231) in comparison to hate speech (N=513) or violent content (N=620). As the authors themselves acknowledge, the number of mentions to a particular issue in a T&C does not perfectly encapsulate how important a topic is to a platform. However, it is a useful proxy to demonstrate how much attention platforms pay to a particular topic and the importance of detecting infringing content.

5. Discussion

In short, the literature showed both theoretical and empirical evidence that platforms play a central role in enabling or restraining populist hate narratives through their algorithms and terms and conditions. Regrettably, I found no papers examining FC-platforms collaborations, especially on the effects of such collaborations in curbing or enabling harmful content. This is a notable gap in the literature, which should readily be addressed.

Considering this gap, I found no compelling and clear evidence in the literature on how FC-collaborations are associated with the prevalence of populist hate narratives. Notwithstanding this, having broadly explored the existing evidence in the literature on the links between misinformation, hate speech, the role of FCs and that of platforms, I make a second attempt at answering this question by extrapolating the arguments and evidence.

5.1. Do platforms restrain populist hate narratives through collaborations with FCs?

As mentioned, I did not find any works focusing on the relationships between FCs and platforms, and the impact of these collaborations to the broader hate speech context. This absence of direct evidence necessitates an inferential approach, drawing from adjacent findings on misinformation and hate speech dynamics.

First, the evidence points to platforms' economic incentives to develop algorithms which prioritise divisive content, as this type of content tends to catch users' attention. The types of divisive content available online is not, however, unfettered. Notably, the prevalence of hate speech in a given platform can lead to user avoidance. Consequently, the largest platforms in the market explicitly prohibit the dissemination of hate speech through their terms and conditions.

At first, therefore, the evidence points to a reduced importance of collaborations between FCs and platforms; if platforms are strict in their policies when it comes to hate speech, then FC collaborations play virtually no part in enabling or restraining hate narratives - because the content will be removed anyway. Conversely, if the content is more nuanced and does not constitute prima facie hate speech, FCs can play a crucial role in providing further context.

Further evidence also contributes to the argument that FCs and platforms should collaborate to reduce the spread of hate speech online. First, even if platforms enforce their terms and conditions and filter hateful content, Budak et al. (2024) demonstrated they still show harmful content to users who seek it out. Drawing from this, fact-checking initiatives should contribute to play correcting false or misleading statements and avoiding further radicalisation of these users actively looking for harmful content.

Conversely, not all scholars are optimistic about the corrective potential of fact-checking. According to Munn (2024), collaborations between platforms and FCs would not necessarily restrain the spread of hateful content, as the theoretical target of these interventions is a rational human being as opposed to the actual target of these interventions, which are irrational humans. Notwithstanding, he acknowledged that FCs could contribute to reducing misinformation, if linked to psychological research on how to present facts to irrational beings. As such, he admitted that connecting human insights into real-world interventions potentially contributes to reducing misinformation.

Extrapolating the existing evidence, collaborations between FCs and online platforms arguably have the potential to mitigate populist hate narratives. Although, the degree of efficacy of such collaborations might depend on the engagement with users' emotional and cognitive biases

This argument leads to a logical follow up conclusion: By striking down collaborations with FCs, platforms potentially enable hate narratives. In other words, if Meta expands its decision to crack down on collaborations with FCs, or if other platforms decide to follow suit, they are arguably contributing to enabling hateful messages.

This conclusion merits, however, one important caveat. To restrict hateful messages, platforms should not only maintain or expand their collaborations with FCs, but remove barriers to checking certain content, such as political content. Currently, platforms such as Meta impose restrictions on the verification of political speech. As a fact-checker pointed out '*The bar to be a political figure [under Meta's fact-checking policies] is very low – just run for a local county council...and you can lie your heart out!*'(Gutierrez et al., 2025). Consequently, keeping these limitations in place, platforms enable the spread of hateful populist messages by political leaders.

5.2. Broader legal and regulatory consequences of the findings

The legal and regulatory consequences of the findings above are twofold. First, from a state perspective, hate speech laws should account for FCs' role in promoting a heterogeneous information environment. As such, States normative frameworks and policies should incentivise FC-platforms collaborations. Second, collaborations between FCs and platforms ensure an approach which is in line with the ECHR. I will assess each issue in turn.

First, as argued, a significant number of papers found links between misinformation and hate speech. From an empirical perspective, Liu et al. (2024) demonstrated that, while there are links between legal regulation and reduced hate speech, there is a non-significant conditional effect of online legal regulation. In other words, the authors argued that legal regulation can constrain hate speech, while also limiting other kinds of lawful speech. Put differently, they claimed that hate speech laws might build an invisible information barrier that operates against the regulation's goals. They further expand on the argument that, 'to eliminate hate speech, people need to be exposed to adequate and heterogeneous information to nurture a tolerant attitude and an open mind' (Liu et al., 2024, p. 542).

Expanding from their argument, hate speech laws must account for FCs' role in ensuring that people are exposed to heterogeneous and reliable information, fostering a tolerant attitude that can, indirectly, help curb hate speech. Instead of enacting hate speech laws that seek to restrict or eliminate certain forms of speech, States can benefit from enacting laws and policies that promote collaborations between FCs and platforms and promoting the role of FCs more broadly speaking.

Second, by collaborating with FCs, platforms ensure an approach that not only observes the DSA and their obligations under Union law, but also the ECHR. Notably, the Article 10 ECHR protects speech that offends, shocks or disturbs, and any measures restricting these kinds of speech must be prescribed by law, in pursuance of a legitimate aim, and be necessary and proportionate. A legal approach which promotes counterspeech instead of restricting speech is preferable as it would not trigger an interference with Article 10 Rights. In other words, by collaborating with FCs, platforms do not interfere or restrict freedom of expression: FCs promote counterspeech instead of censoring users.

In short, overall, the literature supports a normative case for sustaining FC–platform collaborations. While empirical proof of their efficacy remains limited, their alignment with freedom of expression principles and their potential to curb misinformation-induced hate speech justify their continued promotion in human rights frameworks.

6. Conclusion

In conclusion, this paper considered the implications of collaborations between platforms and FCs to curb online hate and populist narratives. Through a systematic literature review, this study contributed to three scholarly debates. First, I considered how misinformation and hate speech intersect online. Second, I assessed how fact-checking practices may indirectly address hate speech. Finally, I consider the existing evidence on how collaborations between platforms and FCs can reshape responsibility for harmful content under European human rights law.

The paper revealed an important gap in the literature: I found no theoretical or empirical works assessing the impact of collaborations between FCs and platforms in tackling hate speech and populist narratives. However, I found important evidence on broader issues. For instance, there is extensive literature on how misinformation and hate speech intersect online, as well as on the importance of fact-checking initiatives to tackle misinformation and platforms' terms of conduct to restrain hate speech.

Admittedly, some authors emphasise the limitations inherently associated with fact-checking initiatives (such as primarily reaching a different target audience, not necessarily the hate speech spreaders, as argued by Roozenbeek et al. (2023)). Yet, by extrapolating the existing evidence, I argue that FCs can still play a crucial role as a tool to tackle populist hate narratives. This is because, if there is as solid relationship between misinformation and hate speech as portrayed by the literature, these initiatives can still be valuable to constrain hate speech without censorship. Consequently, I argue that FC-platform collaborations advance the fight against hate speech in line with Article 10 ECHR. Future work should, however, consider in greater depth empirical evidence of the efficacy of these collaborations, as well as the consequences of scaling down or withdrawing from such collaborations.

List of papers considered

	Paper
1	Abdul Reda, A., & Alkhonin, A. (2025). More pressing matters: Can priority reorientation beat online misinformation? *Journal of Computational Social Science*, *8*(2).
2	Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., & Augenstein, I. (2024). Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go. *ACM Computing Surveys*, *56*(3), 1–17.
3	Asardag, D. (2025). Feminist exploratory interpretive study of the content policy changes of Meta and the corresponding news coverage. *Frontiers in Communication*, *10*.
4	Baptista, J. P., Gradim, A., & Fonseca, D. (2024). Populist Leaders as Gatekeepers: André Ventura Uses News to Legitimize the Discourse. *Journalism and Media*, *5*(3), 1329–1347.
5	Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature (London)*, *630*(8015), 45–53.
6	Caldevilla-Domínguez, D., Barrientos-Báez, A., & Padilla-Castillo, G. (2023). Dilemmas Between Freedom of Speech and Hate Speech: Russophobia on Facebook and Instagram in the Spanish Media. *Politics and Governance*, *11*(2), 147–159.
7	Carrasco-Farré, C. (2022). The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities & Social Sciences Communications*, *9*(1), 1–18.
8	Cepollaro, B., Lepoutre, M., & Simpson, R. M. (2023). Counterspeech. *Philosophy Compass*, *18*.
9	Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P. K., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific Reports*, *11*(1), 22083–12.
10	Doncel-Martín, I., Catalan-Matamoros, D., & Elías, C. (2023). Corporate social responsibility and public diplomacy as formulas to reduce hate speech on social media in the fake news era. *Corporate Communications*, *28*(2), 340–352.
11	Downing, J., & Dron, R. (2020). Tweeting Grenfell: Discourse and networks in critical constructions of British Muslim social boundaries on social media. *New Media & Society*, *22*(3), 449–469.
12	Fraser, R. (2023). *How to talk back: Hate speech, misinformation, and the limits of salience*.
13	Garg, V., Xu, G., & Singh, M. P. (2025). Understanding Inciting Speech as New Malice. *IEEE Transactions on Computational Social Systems*, *12*(3), 947–956.
14	González-Aguilar, J. M., Segado-Boj, F., & Makhortykh, M. (2023). Populist Right Parties on TikTok: Spectacularization, Personalization, and Hate Speech. *Media and Communication (Lisboa)*, *11*(2), 232–240.
15	Heldt, A. (2019). Let's Meet Halfway: Sharing New Responsibilities in a Digital Age. *Journal of Information Policy (University Park, Pa.)*, *9*, 336–369.
16	Komendantova, N., Erokhin, D., & Albano, T. (2023). Misinformation and Its Impact on Contested Policy Issues: The Example of Migration Discourses. *Societies (Basel, Switzerland)*, *13*(7), 1–16.
17	Kumar, A., & Maurya, M. K. (2024). Online Public Sphere and Threats of Disinformation, Extremism and Hate Speech: Reflections on Threat-Mitigation. *The Journal of Communication Inquiry*.

18	Liu, Z., Luo, C., & Lu, J. (2024). Hate speech in the Internet context: Unpacking the roles of Internet penetration, online legal regulation, and online opinion polarization from a transnational perspective. *Information Development*, *40*(4), 533–549.
19	Mayagoitia-Soria, A., González-Aguilar, J. M., Gómez-García, S., & Paz-Rebollo, M. A. (2024). "Drop a Bomb on Them… and Problem Solved!" An Analysis of Poverty Discourse on TikTok. *International Journal of Communication*, *18*, 1135–1156.
20	Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review*, *38*(2), 128–146.
21	Mosleh, M., Cole, R., & Rand, D. G. (2024). Misinformation and harmful language are interconnected, rather than distinct, challenges. *PNAS Nexus*, *3*(3), 1–4.
22	Munn, L. (2024). Misinformation's missing human. *Media, Culture & Society*, *46*(6), 1287–1298.
23	Poole, E., Giraud, E. H., & de Quincey, E. (2021). Tactical interventions in online hate speech: The case of #stopIslam. *New Media & Society*, *23*(6), 1415–1442.
24	Roberts-Ingleson, E. M., & McCann, W. S. (2023). The Link between Misinformation and Radicalisation: Current Knowledge and Areas for Future Inquiry. *Perspectives on Terrorism (Lowell)*, *17*(1), 36–49.
25	Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions. *European Psychologist*, *28*(3), 189–205.
26	Sehat, C. M., Li, R., Nie, P., Prabhakar, T., & Zhang, A. X. (2024). Misinformation as a Harm: Structured Approaches for Fact-Checking Prioritization. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–36.
27	Silva, A. de, & Parker, C. (2025). Platformed hate speech against women: Beyond self-regulation. *UNSW Law Journal*, *48*(2), 637–678.
28	Šori, I., & Vehovar, V. (2022). Reported User-Generated Online Hate Speech: The 'Ecosystem', Frames, and Ideologies. *Social Sciences (Basel)*, *11*(8), 375-.
29	Vasist, P. N., Chatterjee, D., & Krishnan, S. (2024). The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative. *Information Systems Frontiers*, *26*(2), 663–688.
30	Vicari, R., Elroy, O., Komendantova, N., & Yosipof, A. (2024). Persistence of misinformation and hate speech over the years: The Manchester Arena bombing. *International Journal of Disaster Risk Reduction*, *110*, 1–15.
31	Vitullo, A. (2021). The Online Intersection among Islamophobia, Populism, and Hate Speech: An Italian Perspective. *Journal of Religion, Media and Digital Culture*, *10*(1), 95–114.
32	Zhou, J., Verma, G., Zhang, L., Chang, N., & De Choudhury, M. (2025). Harm in Layers: Compositions of Misinformative Hate in Anti-Asian Speech and Their Impacts on Perceived Harmfulness. *Proceedings of the ACM on Human-Computer Interaction*, *9*(2), 1–22.

Bibliography

Abdul Reda, A., & Alkhonin, A. (2025). More pressing matters: Can priority reorientation beat online misinformation? *Journal of Computational Social Science*, *8*(2).

Adam, D. (2025). Does fact-checking work? What the science says. *Nature (London)*.

Alkiviadou, N. (2025, March 7). Hate Speech, Positive Obligations And Free Speech: The ECtHR's Expanding Framework In Minasyan And Others V. Armenia (2025). *Strasbourg Observers*. https://strasbourgobservers.com/2025/03/07/hate-speech-positive-obligations-and-free-speech-the-ecthrs-expanding-framework-in-minasyan-and-others-v-armenia-2025/

Arora, A., Nakov, P., Hardalov, M., Sarwar, S. M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., & Augenstein, I. (2024). Detecting Harmful Content on Online Platforms: What Platforms Need vs. Where Research Efforts Go. *ACM Computing Surveys*, *56*(3), 1–17.

Asardag, D. (2025). Feminist exploratory interpretive study of the content policy changes of Meta and the corresponding news coverage. *Frontiers in Communication*, *10*.

Bélair-Gagnon, V., Larsen, R., Graves, L., & Westlund, O. (2023). Knowledge Work in Platform Fact-Checking Partnerships. *International Journal of Communication*, *17*, 1169–1189. https://doi.org/1932–8036/20230005

Bengtsson, M., Schousboe, S., Farkas, J., Schjøtt, A., Kjeldsen, J. E., & Hess, A. (2025). Fact-Checkers, Tech-Giants, and Algorithmic Systems: Between Autonomy and Automation in the Relational and Dispersed Construction of Ethos. In *Ethos, Technology, and AI in Contemporary Society* (1st ed., pp. 249–274). Routledge.

Budak, C., Nyhan, B., Rothschild, D. M., Thorson, E., & Watts, D. J. (2024). Misunderstanding the harms of online misinformation. *Nature (London)*, *630*(8015), 45–53.

Cavaliere, P. (2020). *From journalistic ethics to fact-checking practices: Defining the standards of content governance in the fight against disinformation*. Routledge, Taylor & Francis Group.

Cepollaro, B., Lepoutre, M., & Simpson, R. M. (2023). Counterspeech. *Philosophy Compass*, *18*.

CM/Rec(2022)16 - Recommendation of the Committee of Ministers to Member States on Combating Hate Speech (2022).

Erjavec, K., & Kovačič, M. P. (2012). "You Don't Understand, This is a New War!" Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society*, *15*(6), 899–920. https://doi.org/10.1080/15205436.2011.619679

Gibson, R. C., Meiklem, R., Moncur, W., & Ruthven, I. (2025). Online Information Disclosure and Information Privacy Practices During Significant Life Transitions: A Scoping Review. *CHIIR '25: Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval*, 42–56.

González-Aguilar, J. M., Segado-Boj, F., & Makhortykh, M. (2023). Populist Right Parties on TikTok: Spectacularization, Personalization, and Hate Speech. *Media and Communication (Lisboa)*, *11*(2), 232–240.

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

Gutierrez, N. S. G., Mota, J. da C., & Stremlau, N. (2025, November). *ReMeD Policy Brief n1: Building a Stronger Information Ecosystem through Content Moderation: Perspectives from European fact-checkers.* ReMeD - Resilient Media for Democracy in the Digital Age. https://resilientmedia.eu/?p=1899

Heldt, A. (2019). Let's Meet Halfway: Sharing New Responsibilities in a Digital Age. *Journal of Information Policy (University Park, Pa.)*, *9*, 336–369.

Hietanen, M., & Eddebo, J. (2023). Towards a Definition of Hate Speech—With a Focus on Online Contexts. *The Journal of Communication Inquiry*, *47*(4), 440–458.

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, *546*, 126232. https://doi.org/10.1016/j.neucom.2023.126232

Kahn, G. (2025, January 17). *Amid war, vicious attacks and political turmoil, global fact-checkers fear the impact of the end of Meta's programme*. Reuters Institute. https://reutersinstitute.politics.ox.ac.uk/news/amid-war-vicious-attacks-and-political-turmoil-global-fact-checkers-fear-impact-end-metas

Kaplan, J. (2025, January 7). More Speech and Fewer Mistakes. *Meta Newsroom*. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ (Clinical Research Ed.)*, *339*, b2700. https://doi.org/10.1136/bmj.b2700

Liu, Z., Luo, C., & Lu, J. (2024). Hate speech in the Internet context: Unpacking the roles of Internet penetration, online legal regulation, and online opinion polarization from a transnational perspective. *Information Development*, *40*(4), 533–549.

López-Borrull, A., & Lopezosa, C. (2025). Mapping the Impact of Generative AI on Disinformation: Insights from a Scoping Review. *Publications (Basel)*, *13*(3), 33-.

Mahl, D., Zeng, J., Schäfer, M. S., Egert, F. A., & Oliveira, T. (2024). "We Follow the Disinformation": Conceptualizing and Analyzing Fact-Checking Cultures Across Countries. *The International Journal of Press/Politics*.

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, Hate Speech, and Social Media: A Systematic Review and Critique. *Television & New Media*, *22*(2), 205–224.

Mayagoitia-Soria, A., González-Aguilar, J. M., Gómez-García, S., & Paz-Rebollo, M. A. (2024). "Drop a Bomb on Them… and Problem Solved!" An Analysis of Poverty Discourse on TikTok. *International Journal of Communication*, *18*, 1135–1156.

Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering Hate Speech on Facebook: The Case of the Roma Minority in Slovakia. *Social Science Computer Review*, *38*(2), 128–146.

Munn, L. (2024). Misinformation's missing human. *Media, Culture & Society*, *46*(6), 1287–1298.

Okoli, C. (2015). A Guide to Conducting a Standalone Systematic Literature Review. *Communications of the Association for Information Systems*, *37*, 43-.

Pentney, K., & Shattock, E. (2025). Disinformation and Democracy on the Docket: Reformulating the Approach to Electoral Disinformation under the ECHR. *Oxford Journal of Legal Studies*. https://doi.org/10.1093/ojls/gqaf026

Pérez Rastrilla, L., Sapag M., P., & Recio García, A. (2023). Fast Politics: Propaganda in the Age of TikTok. In *Fast Politics: Propaganda in the Age of TikTok* (1st ed. 2023.). Springer Nature Singapore.

Petticrew, M., & Roberts, H. (2008). *Systematic reviews in the social sciences: A practical guide* (1st ed.). Wiley.

Poole, E., Giraud, E. H., & de Quincey, E. (2021). Tactical interventions in online hate speech: The case of #stopIslam. *New Media & Society*, *23*(6), 1415–1442.

Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering Misinformation: Evidence, Knowledge Gaps, and Implications of Current Interventions. *European Psychologist*, *28*(3), 189–205.

Sarah Grevy Gotfredsen. (2023, December 6). Q&A: What happened to academic research on Twitter? *Columbia Journalism Review*. https://www.cjr.org/tow_center/qa-what-happened-to-academic-research-on-twitter.php

Schradie, J. (2019). The revolution that wasn't: How digital activism favors conservatives. In *The revolution that wasn't: How digital activism favors conservatives*. Harvard University Press.

Sehat, C. M., Li, R., Nie, P., Prabhakar, T., & Zhang, A. X. (2024). Misinformation as a Harm: Structured Approaches for Fact-Checking Prioritization. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–36.

Silva, A. de, & Parker, C. (2025). Platformed hate speech against women: Beyond self-regulation. *UNSW Law Journal*, *48*(2), 637–678.

Westlund, O., Belair-Gagnon, V., Graves, L., Larsen, R., & Steensen, S. (2024). *What Is the Problem with Misinformation? Fact-checking as a Sociotechnical and Problem-Solving Practice*.