Selected Papers of #AoIR2025:
The 26th Annual Conference of the
Association of Internet Researchers
Niterói, Brazil / 15 – 18 Oct 2025

META'S 3PFC SPEECH GOVERNANCE: AN INQUIRY INTO THE FACT-CHECKING CONTENT MODERATION INFRASTRUCTURE

Otávio Vinhas
Instituto Nacional de Ciência e Tecnologia em Disputas e Soberanias Informacionais (INCT-DSI)

Marco Bastos
University College Dublin

Introduction

Social media play a pivotal role in managing the environments where political debate and public deliberation occur, and therefore must contend with the targets set by official bodies and policymakers to ward off harmful speech and mis/disinformation in their platforms. Following concerns about the spread of problematic information online, several government and multistakeholder policy frameworks have endorsed fact-checking to moderate content on social platforms (European Commission, 2022; Wardle & Derakhshan, 2017). The establishment of fact-checking-as-content-moderation initiatives, particularly Meta's Third-Party Fact-Checking Program (3PFC), has sparked controversies, as they impinge on fact-checkers' public-oriented principles to partake in social platforms' opaque bureaucracy (Bélair-Gagnon et al., 2022). Despite the mixed, if mostly positive, impact of fact-checking on information online (Graham & Porter, 2024; Walter et al., 2019), little is known about the criteria applied by social platforms' content moderation infrastructure to address problematic information.

Recent work has highlighted the role of social platforms in regulating online speech and noted that the content moderation of social platform operate within broader speech governance infrastructures responsible for distributing, promoting, categorizing, and restricting user-generated content (Douek, 2022; Gillespie, 2022; Iliadis & Ford, 2023). Rather than merely responding to bad user behavior associated with mis/disinformation, these infrastructures rely on normative parameters that anticipate the characteristics of speech ultimately subjected to moderation (Finn & Ananny, 2024). Nonetheless, internet scholars widely agree on the lack of operational transparency and hindered access to the criteria driving the content moderation policies of social platforms (Walker et al., 2019). Given the multilayered infrastructure of both human and automated tools administering speech and determining the visibility of content (McNally & Bastos, 2025), incoming research must develop methods that can shed light onto social platform's opaque infrastructures (de Keulenaar et al., 2023).

Suggested Citation (APA): Lastname, Firstinitial. (2025, October). *Paper (or panel) title*. Paper (or panel) presented at AoIR2025: The 26th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

This paper examines the speech governance parameters applied by Meta's 3PFC to moderate problematic information. Leveraging digital methods and approaches developed in journalism studies, it implements a series of manual and computational techniques to curate a comprehensive dataset combining fact-checking content commissioned by Meta's 3PFC program—available through the Facebook URLs Dataset (Meta, n.d.)—and fact-checks produced independently from Meta's program. We probe this database to identify the criteria applied by Meta's speech governance through the 3PFC program across five countries: Argentina, Philippines, Portugal, United Kingdom, and South Africa, with content spanning three languages (English, Spanish, and Portuguese).

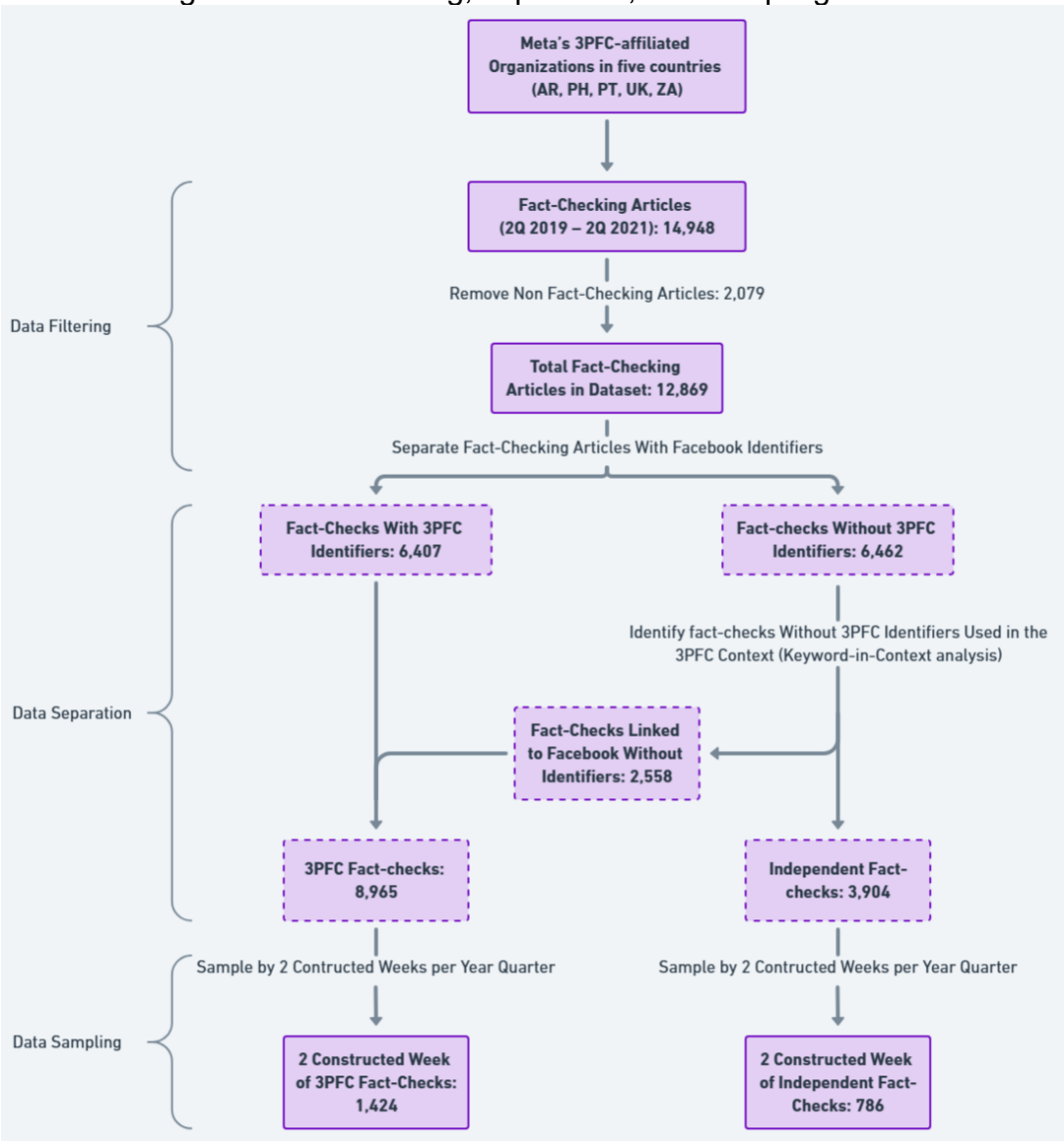
Considering the looming end of Meta's 3PFC system worldwide and the recently declared laissez-faire commitments of US-based tech companies to public speech (Silverman, 2025), this paper addresses the normative standards consolidated by Meta's content moderation infrastructure through outsourced fact-checking work. The findings contribute to developing platform governance policies in a context where countries like Australia, Brazil, India, and the European Union are designing or implementing regulatory frameworks to hold social platforms accountable (Anastácio, 2024; Liu, 2024; Ó Fathaigh et al., 2021).

Data & Method

The paper employs a quantitative content analysis on a sample of 2,210 fact-checking articles published by all fact-checking organizations affiliated with Meta's 3PFC in Argentina, Philippines, Portugal, United Kingdom, and South Africa between Q2-2019 and Q2-2021. This period covers the lion's share of fact-checked URLs that appear in the Facebook URLs Dataset (Evans & King, 2022). Data was collected by scraping all available fact-checking articles published in the websites of the organizations affiliated with the 3PFC program in the above countries. The selected organizations maintain a national editorial focus rather than an international one and therefore exclude organizations such as Reuters in the UK.

Next we used CrowdTangle and Facebook search queries to manually cross-check each URL marked as fact-checked in the Facebook URLs Dataset to identify the fact-checking article attached to posts labeled on the platform. We identified 778 unique fact-checking articles in the full period covered by the database, which served as a reference for identifying the formal attributes and common textual characteristics of fact-checks produced within the context of the 3PFC program. These parameters outline the fact-checking formats (fact-checks, debunking, explainers, etc.) produced by organizations in the context of the 3PFC program used to differentiate 3PFC fact-checks from independent fact-checks.

Figure 1. Flow diagram of data filtering, separation, and sampling



We employed a constructed week sampling (Monday to Friday) to extract a dataset conducive to manual quantitative content analysis. Constructed week sampling is considered more effective than random sampling as it accounts for topical variation of news-related content (Riffe et al., 2019), a feature that is particularly suitable to capture the seasonal nature of fact-checking trends (Singer, 2023). This method has been extensively used and validated, with six constructed weeks per year considered the best approach for news articles (Luke et al., 2011). Table 1 provides a breakdown of the distribution of fact-checking articles in our dataset.

Table 1. Distribution of Fact-Checking Articles in the Dataset per Organization

Organization	Country	Type	3PFC	Independent
AFP Argentina	Argentina	Global News Agency	68	2
Chequeado		Non-profit	177	212
AFP Philippines	Philippines	Global News Agency	38	1
Rappler		Digital News Media	112	42
VERA Files		Non-profit	89	72
Observador	Portugal	Digital News Media	112	27
Poligrafo		Independent Agency	206	150
Africa Check	South Africa	Non-profit	416	90
AFP South Africa		Global News Agency	21	2
Full Fact	UK	Non-profit	179	160
FactCheck.NI		Non-profit	6	28

The quantitative content analysis is effective at surfacing tangible differences between 3PFC and independent fact-checks. Three independent coders have undergone 35 hours of training and are in the final stages of achieving optimal reliability levels on 10% of the data. The analysis is conducted at both the article level, examining (a) theme, (b) author of the claim, (c) geographic scope of the claim, and (d) verification sources; and at the source level to determine the geographic location of sources. We build on previous research to develop a codebook (Cazzamatta, 2024; Leon et al., 2022; Mahl et al., 2024) that integrates a pre-analysis of the data conducted by the authors through inductive coding.

Preliminary Results

Descriptive findings from the data cataloging process highlight the disparity in the quantity of 3PFC fact-checks produced in comparison to independent fact-checks. With the exception of a few organizations (Chequeado, VERA Files, Poligrafo, and Full Fact), most organizations show a disproportionate reliance on the 3PFC program, particularly the national fact-checking units from AFP. This issue is more salient in South Africa, with all organizations working primarily on content sourced through Meta's fact-checking program. These results are consistent with previous findings from qualitative interviews that point to the development of two substantially distinct fact-checking regimes—one conducted independently in line with fact-checkers' public-oriented mission and the other aligned with Meta's speech governance through the 3PFC program (Graves et al., 2023; Vinhas & Bastos, 2023).

Our preliminary results also show that the vast majority of 3PFC fact-checks focus on viral posts authored by anonymous users on social platforms. Consistent with Meta's fact-checking framework (Meta, 2022), this approach prioritizes corporate guidelines that enforce a tiered content governance infrastructure (Caplan & Gillespie, 2020), a dimension that our dataset makes observable. It further elucidates that this speech governance regime assigns responsibility for problematic content to end-users by administering the visibility of their speech acts while exempting elite actors—such as politicians and influential accounts—from accountability (Cotter et al., 2022).

References

Anastácio, K. (2024). Framing disinformation through legislation: Evidence from policy proposals in Brazil. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-150

Bélair-Gagnon, V., Graves, L., Kalsnes, B., Steensen, S., & Westlund, O. (2022). Considering Interinstitutional Visibilities in Combating Misinformation. *Digital Journalism*, *10*(5), 669-678. https://doi.org/10.1080/21670811.2022.2072923

Caplan, R., & Gillespie, T. (2020). Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. *Social Media + Society*, *6*(2). https://doi.org/10.1177/2056305120936636

Cazzamatta, R. (2024). Global misinformation trends: Commonalities and differences in topics, sources of falsehoods, and deception strategies across eight countries. *New Media & Society*. https://doi.org/10.1177/14614448241268896

Commission, E. (2022). *The Strengthened Code of Practice on Disinformation*. Retrieved from https://ec.europa.eu/newsroom/dae/redirection/document/87585

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). https://doi.org/10.1177/20563051211069048

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

Douek, E. (2022). Content Moderation as Systems Thinking. *Harvard Law Review*, *136*(2).

Evans, G., & King, G. (2022). Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset. *Political Analysis*, *31*(1), 1-21. https://doi.org/10.1017/pan.2022.1

Finn, M., & Ananny, M. (2024). Making events: How anticipatory infrastructures produce shared temporalities. *New Media & Society*. https://doi.org/10.1177/14614448241236709

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, *8*(3). https://doi.org/10.1177/20563051221117552

Graham, M. H., & Porter, E. V. (2024). Increasing Demand for Fact-Checking. *Political Communication*, 1-24. https://doi.org/10.1080/10584609.2024.2395859

Graves, L., Bélair-Gagnon, V., & Larsen, R. (2023). From Public Reason to Public Health: Professional Implications of the “Debunking Turn” in the Global Fact-Checking Field. *Digital Journalism*, 1-20. https://doi.org/10.1080/21670811.2023.2218454

Iliadis, A., & Ford, H. (2023). Fast Facts: Platforms From Personalization to Centralization. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231195546

Leon, B., Martinez-Costa, M. P., Salaverria, R., & Lopez-Goni, I. (2022). Health and science-related disinformation on COVID-19: A content analysis of hoaxes identified by fact-checkers in Spain. *PLoS One*, *17*(4), e0265995. https://doi.org/10.1371/journal.pone.0265995

Liu, D. (2024). Borderline content and platformised speech governance: Mapping TikTok's moderation controversies in South and Southeast Asia. *Policy & Internet*, *16*(3), 543-566. https://doi.org/10.1002/poi3.388

Luke, D. A., Caburnay, C. A., & Cohen, E. L. (2011). How Much Is Enough? New Recommendations for Using Constructed Week Sampling in Newspaper Content Analysis of Health Stories. *Communication Methods and Measures*, *5*(1), 76-91. https://doi.org/10.1080/19312458.2010.547823

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*. https://doi.org/10.1177/19401612241270004

McNally, N., & Bastos, M. (2025). The News Feed is Not a Black Box: A Longitudinal Study of Facebook's Algorithmic Treatment of News. *Digital Journalism*, 1-20. https://doi.org/10.1080/21670811.2025.2450623

Meta. (2022). *Meta's third-party fact-checking program* https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking

Meta. (n.d.). *URL Shares dataset*. https://developers.facebook.com/docs/url-shares-dataset/

Ó Fathaigh, R., Helberger, N., & Appelman, N. (2021). The perils of legally defining disinformation. *Internet Policy Review*, *10*(4). https://doi.org/10.14763/2021.4.1584

Riffe, D., Lacy, S., Watson, B. R., & Fico, F. (2019). *Analyzing Media Messages* (4 ed.). Routledge.

Silverman, C. (2025). As Facebook Abandons Fact-Checking, It's Also Offering Bonuses for Viral Content. https://www.propublica.org/article/facebook-meta-abandons-fact-checking-boosts-viral-content

Singer, J. B. (2023). Closing the Barn Door? Fact-Checkers as Retroactive Gatekeepers of the COVID-19 "Infodemic". *Journalism & Mass Communication Quarterly*, *100*(2), 332-353. https://doi.org/10.1177/10776990231168599

Vinhas, O., & Bastos, M. (2023). The WEIRD governance of fact-checking and the politics of content moderation. *New Media & Society*. https://doi.org/10.1177/14614448231213942

Walker, S., Mercea, D., & Bastos, M. (2019). The disinformation landscape and the lockdown of social platforms. *Information, Communication & Society*, *22*(11), 1531-1543. https://doi.org/10.1080/1369118x.2019.1648536

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2019). 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

Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*.