Selected Papers of #AoIR2022:
The 23rd Annual Conference of the
Association of Internet Researchers
Dublin, Ireland / 2-5 Nov 2022

THE TOXIC TURN? CONCEPTUAL AND METHODOLOGICAL ADVANCES ON PROBLEMATIC/TOXIC CONTENTS ON SOCIAL MEDIA

Marco Bastos
University College Dublin

Eugenia Siapera,
University College Dublin

Marc Tuters
University of Amsterdam

Shawn Walker
Arizona State University

Sanaz Rasti
University College Dublin

Padraig Cunningham
University College Dublin

Cliona Curley
University College Dublin

Brendan Scally
University College Dublin

Mansura Khan
University College Dublin

The ‘toxic turn’ in social media platforms continues unabated. Hate speech, mis- and disinformation, misogynistic and racist speech, images, memes and videos are all far

Suggested Citation (APA): Siapera, E., Bastos, M., Tuters, M., Walker, S., Rasti, S., Cunningham, P., Curley, C., Scally, B., Khan. M. (2022, November). *The Toxic Turn: Conceptual and Methodological Advances on Problematic/Toxic Contents on Social Media. P*anel presented at AoIR 2022: The 23rd Annual Conference of the Association of Internet Researchers. Dublin, Ireland: AoIR. Retrieved from http://spir.aoir.org.

too common on social media platforms and more broadly on the internet. While the diminishing popularity of populist politicians led to hopes for less social toxicity, the Covid-19 pandemic introduced new and more complex dimensions. Tensions have emerged around what constitutes problematic content and who gets to define it. Co-regulation models, such as for example the EC Code of Conduct against Illegal Hate Speech, focus on the legality of certain types of contents, while leaving other categories of problematic contents to be defined by platforms. In parallel, the social media ecosystem became more diverse, as new platforms with hands off moderation policies attracted users who felt too constrained by the policies of mainstream platforms. The proposed panel examines this complex and dynamic landscape by problematizing what is understood as toxic, deplatformed, removable and in general problematic content on platforms with the aim to probe the boundaries of what is constituted as acceptable discourse on platforms and to map its implications.

In particular, this panel discusses the broad definition of 'problematic content' employed by social media platforms, a catch-all term that cuts across hate speech and propaganda, including more politically topical content such as mal-, mis-, and disinformation, hyperpartisan and polarising content, but also abusive, misogynistic, racist, and homophobic discourse. The term is also employed to refer to spam and content that infringes upon the Terms of Service or the Community Standards of social media platforms. As such, it is a broad category that resists a narrower classification given the operational scope of its use. Defining what constitutes problematic content is a key operation of platform content moderation policies but is also the subject of intense debates (de Gregorio, 2020; Gillespie, 2018; Gillespie et al., 2020; Gorwa et al., 2020).

The panel interrogates the many definitions and applications of problematic content on social media platforms and applications through an empirically informed lens and focusing on deleted contents, complex mixed narratives, and grey areas, including hidden misinformation on voice applications. *Problematic Content according to Twitter Compliance API* presents ongoing work on the Twitter Compliance API and the Compliance Firehose, which allow researchers to identify content that has been deleted, deactivated, protected, or suspended from Twitter, a proxy for problematic content. In *Multi-Part Narratives on Telegram* Siapera presents ongoing research that probes the intersection between Covid-19 scepticism, far right and other political narratives in vaccine hesitant groups on Telegram. The third contribution, *What if Bill Gates really is evil, people? Investigating the infodemic's grey areas* discusses the conceptual and methodological definitions of problematic content in relation to work on anti-vax and other conspiratorial narratives on Instagram and on Twitter. All contributions foreground the difficulties and costs of identifying and dealing with problematic contents on social media.

The panel fits with theme of decolonization in two ways: firstly, because it is concerned with the tensions around how toxic/problematic contents are defined and who gets to define them; and secondly, because of its focus on neo-colonial discourses or justifications for colonialism in both narratives hosted by platforms and in platforms' attempts to regulate content. As some narratives are flagged for removal by social platforms, they also raise the question of who is deciding and what does problematic content mean, with far right discourses exploiting this tension and ironically denouncing any attempt to regulate the public discourse as ideological enforcement and justification for (neo)colonial practices performed by social media platforms. From this perspective, platforms' own claims about what constitutes acceptable content is uncomfortably close to colonial narratives of civilised discourse and brings to the fore the potential for neo-colonial narratives and practices in digital spaces.

References

De Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. Computer Law & Security Review, 36, 105374.

Gillespie, T. (2018). Custodians of the Internet. Yale University Press.

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., ... & West, S. M. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. Internet Policy Review, 9(4), Article-number.

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.

PROBLEMATIC CONTENT ACCORDING TO TWITTER COMPLIANCE API

Marco Bastos
University College Dublin

Shawn Walker
Arizona State University

In this talk we discuss the Twitter batch compliance endpoint and the Compliance Firehose made available to researchers querying and monitoring the Covid-19 Twitter public stream. These endpoints allow developers and researchers to batch upload large amounts of Twitter data and receive information about the current state of the content on Twitter. While they were created to help developers and academic researchers keep their Twitter data sets in compliance with the Developer Agreement and Policy, these endpoints can be leveraged to identify removed tweets and user accounts, and therefore offer important information about content that has been deleted, blocked, suspended, or is otherwise no longer publicly available on Twitter.

The removal of social media content results from the enforcement of the community standards and guidelines put together by social media platforms in the past years. The systematic removal of content, however, has the problematic drawback of altering the record of social interactions. Unlike traditional media used to distribute propaganda, such as newspapers and posters in the 20th century (Sanders & Taylor, 1982), or pamphlets and leaflets going back as far as the 16th century (Raymond, 2003), the removal of social media content eliminates any trace of the event, thereby preventing forensic analysis and academic research on influence operations targeting social media platforms. While there have been attempts to create public archives of social media posts, these institutional efforts faced considerable challenges and failed to come to fruition (Zimmer, 2015). Similarly, although social platforms have at times offered archives of disinformation campaigns identified and removed by the very platforms (Elections Integrity, 2018), such sanctioned archives offer only a partial glimpse into the extent of influence operations and may prevent researchers from examining organic contexts of manipulation (Acker & Donovan, 2019).

Researchers using the batch compliance endpoint can thus submit a set of Tweet IDs or user IDs and receive the status of this content, which can be classified as deleted, suspended, protected, geo-information scrubbed, or deactivated—a set of information that until recently could only the retrieved through a labour-intensive process of querying the Search API and reverse-engineering the results. This is important information for researchers tracking problematic content on social media, including mis- and disinformation, which is the type of content that is more likely to be removed from the social platform due to infringements to the ToS. In our experience monitoring the Twitter Compliance Firehose, we estimate that the baseline of tweet deletion stands at around 15%, which is a substantial portion of the entire public conversation on Twitter.

With social media platforms rarely disclosing content that was flagged for removal, the compliance endpoints offer a much welcome glimpse into what is considered problematic

content by Twitter. It makes it possible, for instance, to study the politics of deletion on social platforms and to infer what the company classifies as ‘low-quality content,’ seeing that deleted but particularly suspended content can be used as proxies to tangible examples of users and tweets that were selected for removal by the social media platform.

References

Acker, A. and Donovan, J., (2019). Data craft: a theory/methods package for critical internet studies. *Information, Communication & Society*, 22(11), pp.1590-1609. https://doi.org/10.1080/1369118X.2019.1645194

Elections Integrity. (2018). Data archive https://about.twitter.com/en_us/values/elections-integrity.html

Raymond, J., (2006). *Pamphlets and pamphleteering in early modern Britain*. Cambridge University Press.

Sanders, M.L. and Taylor, P.M., (1982). *British Propaganda during the First World War, 1914–18*. Macmillan International Higher Education.

Zimmer, M., (2015). The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. *First Monday*.

MULTI-PART TOXIC NARRATIVES ON TELEGRAM : COVID-19 SCEPTICISM AND THE FAR RIGHT

Eugenia Siapera
University College Dublin

Sanaz Rasti
University College Dublin

Padraig Cunningham
University College Dublin

Cliona Curley
University College Dublin

Brendan Scally
University College Dublin

Mansura Khan
University College Dublin

The pressures on social media to clean up their platforms from hate speech and disinformation led to a mass 'deplatforming' of far right and conspiracy theory accounts (Rogers, 2020). One of the results of this deplatforming was the 'migration' of such accounts to Alt Tech platforms, whose moderation policies were less stringent. Telegram, was among those platforms that gained followers dissatisfied with what they perceived as 'censorship' or infringement of their freedom of speech (Guhl and Davey, 2020). Since this constitutes a displacement rather than eradication of what may generally be seen as toxic or problematic contents, there are important questions regarding the ways in which the relatively unconstrained space of Alt Tech platforms allows such contents to expand, intensify, proliferate and intermix with other kinds of contents. A case in point is the co-articulation of far right narratives with anti-vaxx and anti-Covid-19 state measures. This complex kind of toxic/problematic contents eludes simple classification and requires a more nuanced approach to identify it, unpack it and understand how its various parts meld together. The proposed contribution is focusing on these complex narratives on Telegram with the aim to understand how the 'multi-toxic' contents/narratives are constituted.

In particular, Covid-19 and misinformation policies pushed a number of Covid-related sceptics, linked by a general mistrust of health authorities but not necessarily ideological affinities, to platforms such as Telegram, because of less surveillance, more opportunity for the exchange of fringe views and less control over and/or removal of contents. While the presence and activities of the far right on Telegram are increasingly well documented (e.g. Urman and Katz, 2020), the combined narratives of vaccine hesitancy/anti-Covid-19 measures along with far right political beliefs are still not well understood. Some studies provide evidence that the far right is associated with Covid-19 scepticism: for example, focusing on Twitter, Caiani et al. (2021) found that in both Italy and the UK the far right has used the pandemic to foreground xenophobic and racist narratives, while Gunz and Schaller

(2022) found overt antisemitic narratives in conspiracy beliefs on Covid-19 in both YouTube and Telegram.

While these studies have outlined some connections between the far right, conspiracy theories and vaccine scepticism, questions still remain. The research questions that this article addresses therefore are: how are complex and multipart toxic narratives constituted? What is the role of political ideology and how does vaccine hesitancy become entangled in far right narratives? What other political or politicised beliefs are involved? To address these questions the present study focused on the Irish Telegram channels on Covid-19, beginning with a set of seed accounts and snowballing to a total of 531 channels (including groups, supergroups and channels). Using a purpose built crawler we have collected all posts and urls posted from March 2020 to early 2022 (data collection is ongoing). Methodologically, the present study relies on ethnographic methods (social listening and digital ethnography), along with supervised and unsupervised topic modelling in order to identify the main narratives on Covid-19 and their articulation with far right, conspiratorial and other kinds of political beliefs.

Initial findings of the topic modelling revealed the following opinion networks: Covid denialism (fake pandemic); anti vaxx daily news; Covid vaccine victims and families; Covid vaccine sceptics supports; Stop the lockdown; Vaccine medical manipulation. We conducted a ‘deep dive’ in each of these areas, using a digital ethnographic approach, including observation, fieldwork note taking, reading, watching and listening to the various multimedia used (including a number of voice notes left by Telegram users) with the objective to identify the areas where they overlap with far right exclusionary beliefs and with conspiratorial beliefs and to understand the nuances of the various ideas, beliefs and discourse circulating.

While the qualitative analysis is still underway, several important observations are already emerging. These include: (i) a deep anxiety expressed by the majority of those posting in these groups covering the economic impact of the lockdown, the impact on children, uncertainties over the vaccines and side effects. Importantly, a number of those posting felt that their anxieties were dismissed in more mainstream areas of the digital public sphere. (ii) The provision of social and psychological support offered by these groups who have become closely knit communities. Importantly, the supports are occasionally offered conditionally, based upon confirming with the beliefs of the groups, for example, for as long as the member is still not vaccinating and wearing masks. (iii) Political beliefs are of two kinds: firstly an ultranationalism, that focuses on a perceived core Irish identity excluding those seen as different, and which views Irish elites as corrupt and not serving the needs of the Irish people. This mobilises a populist rhetoric and some racist and conspiratorial tropes (antisemitic tropes, globalism, depopulation, anti-Agenda 21 etc). The second kind is about infringement of personal liberty and fundamental freedoms, including over health choice, freedom of expression and privacy. Covid conspiracies over population control, constant surveillance, lack of freedom and ‘communism’ are common here. The combination of anxiety with far right political beliefs that provide a context, an explanation and a solution, alongside community-building supports which enhance group cohesion can be linked to an intensification and escalation of critical comments on Covid-19 vaccines, government policies and elite actor actions and to political actions, such as the organisation of street protests.

In making sense of these findings we develop a conceptual framework that brings together Tilly's political opportunity structure and Cammaerts' media opportunity structure. We use these to argue that the way in which the political, media and platform mainstream reacted to the pandemic along with the emergence of Alt Tech platforms have created a political and mediation opportunity structure for the politicisation of Covid-19 sceptics, enabling the far right to opportunistically offer a political roof to these loose networks. Moreover, our initial findings indicate the existence and operation of a comprehensive alternative information ecosystem.

Since many channels emerged out of the need to find a safe space for discussion for organisation of protests and other forms of resistance, in a climate of deplatforming of Covid-19 misinformation, perceptions of censorship and dismissal of all concerns over Covid-19 related measures, our findings suggest that while deplatforming 'exiled' these from the mainstream, they have found an alternative space to flourish. The focus on content removal and more broadly the focus on controlling the circulation of what was deemed to be problematic/toxic content on Covid has ignored the context of production of these contents, contributing to the creation of a political and mediation opportunity structure for loosely connected and ideologically disparate beliefs to coalesce in the anti-vaxx movement. We conclude by arguing that more attention to contexts of production rather than only control of problematic contents may lead to more understanding and ultimately more effective content moderation policies.

References

Caiani, M., Carlotti B., & Padoan, E. (2021). Online Hate Speech and the Radical Right in Times of Pandemic: The Italian and English Cases, *Javnost - The Public*, 28:2, pp. 202-218, DOI: 10.1080/13183222.2021.1922191

Cammaerts, B. (2012). Protest logics and the mediation opportunity structure. *European Journal of Communication*, 27(2), pp. 117-134.

Guhl, J. and Davey, J., (2020). *A safe space to hate: White supremacist mobilisation on Telegram*. ISD Global.

Gunz, H., & Schaller, I. (2022). Antisemitic Narratives on YouTube and Telegram as Part of Conspiracy Beliefs about COVID-19. In *Antisemitism on Social Media* (pp. 129-150). Routledge.

Rogers, R. (2020). Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 35(3), pp. 213-229.

Tilly, C. (1978). From Mobilization to Revolution. Reading: Addison-Wes-ley.

Urman, A., & Katz, S. (2020). What they do in the shadows: examining the far-right networks on Telegram. *Information, Communication & Society*, 1-20.

WHAT IF BILL GATES REALLY *IS* EVIL, PEOPLE? INVESTIGATING THE INFODEMIC'S GREY AREAS

Marc Tuters
University of Amsterdam

One of the surprising developments over the course of the Covid-19 pandemic has been the emergence of political protests against mitigation measures that cut 'diagonally' across identity and class in the shared conviction that 'all power is conspiracy' (Callison & Slobodian, 2021). This talk offers insights into conceptual frameworks and methods for studying these conspiratorially-minded communities in large social media datasets.

With the pandemic, social media platforms have been recognized as a potential vector of information contagion; an 'infodemic'. While platforms acted decisively to remove much 'problematic content', some is much harder to demarcate. An example here is the 'Great Reset', a narrative that emerged to prominence in late 2020 following a meeting in Davos of the World Economic Forum. Confusingly, the Great Reset is thus at once the name for a shadowy baseless 'New World Order'-type narrative (Klein 2020) and, at the same time, the name of an actual agenda for global business leaders publicly promoting a vision of "stakeholder capitalism" (Schwaab, 2021).

As a primary case study, the talk focuses on a figure connected both with the Davos agenda and the Great Reset conspiracy theory: Bill Gates. Long perceived as an archvillain within the open-source programming community, with the arrival of the pandemic Gates suddenly became the preeminent figure at the centre of multiple conspiratorial plots, including a plan to use vaccines to 'microchip' human populations (Shahsavari et al, 2020). While often demonstrably false, generally these narratives are based on kernels of truth—for example relating to Gates' many and varied philanthro-capitalist ventures, including funding pandemic wargames and the development of unique digital certificates embedded within vaccines amongst countless other speculative investments.

The talk offers a selective assessment of the relative engagement of some of these peculiar narratives—as well as an assessment of their moderation status—across two large datasets: 1M Instagram posts and 15M Tweets and from 2020, collected based on an expert list of 'conspiratorial' queries. Drawing on the method of 'digital hermeneutics' (Rommel and Furia, 2018), which combines data science methods with qualitative interpretation and theorization, the talk presents various techniques by which to potentially detect conspiratorial 'narrative convergence' through patterns in hashtags, language and image use over time. By comparing hashtag use in the Instragram dataset diachronically over separate quarters in 2020, preliminary research reveals various pre-existing hashtags (New World Order, Illuminati, etc.) increasingly overlapping with covid-related hashtags. At the same time we see the emergence of Gates as a shared antagonist across many of these hashtag communities. As this research is ongoing, it forms the basis for a hypothesis regarding the phenomenon of narrative convergence over the course of the first waves of the pandemic. The Instagram findings in turn inform queries into the much larger Twitter dataset, which seek to test this convergence hypothesis.

Conceptually, the talk begins by acknowledging Noortje Marres' (2018) claim that, in prioritising 'engaging' messages that circulate over objective knowledge, social media platforms—such as Instagram's parent company Facebook— have created 'a truth-less public sphere by design'. (423) Rather than pathologizing these truth-less narratives—as is so common in the literature on conspiracy theory (cf Sunstein and Vermeule, 2009)—the talk aims to considered whether they may be considered as engaged in forms of speculation, based on Aris Komporozos-Athanasiou's (2021) provocative claim that platforms like Instagram "afford underexplored spaces for the exercise of the speculative imagination" (9). Komporozos-Athanasiou's concept aims to update Benedict Anderson's 'imagined community' to the era of social media, by arguing that "[s]peculation has become the very practice around which modern societies coalesce, the vernacular through which we express our collective disbelief in the waning legitimacy of neoliberalism" (144).

By applying this conceptual lens to these data, the talk considers the extent to which infodemic narratives may plausibly be read as collective means of coping with uncertainty—for example over loss of control in the context of total media immersion and operations of globalised technocapitalism—as opposed to atavistic right-wing reaction? Ultimately, the talk proposes that while deplatforming the infodemic is arguably desirable from a public health perspective, some of what has been labelled as problematic content during the pandemic may be of significant interest to internet researchers—especially those interested in novel practices of knowledge production and community formation in times of great uncertainty.

References

Callison W, and Q Slobodian (2021) "Coronapolitics from the Reichstag to the Capitol". *Boston Review*.

Klein N (2020) 'The Great Reset Conspiracy Smoothie'. *The Intercept*

Komporozos-Athanasiou, A (2021). *Speculative Communities: Living with Uncertainty in a Financialized World*. Chicago: University of Chicago Press

Marres, N (2018). "Why We Can't Have Our Facts Back." *Engaging Science, Technology, and Society* 4.

Romele A, M Severo and P Furia (2020) "Digital hermeneutics: from interpreting with machines to interpretational machines". *AI & SOCIETY* 35(1): 73–86.

Schwab, K and P Vanham (2021) *Stakeholder Capitalism: A Global Economy that Works for Progress, People and Planet*. New York: Wiley

Shahsavari S, P Holur, T Wang, T Tangherlini and V Roychowdhury (2020) Conspiracy in the time of corona: automatic detection of emerging COVID-19 conspiracy theories in social media and the news. *Journal of Computational Social Science*: 1–39.

Sunstein, C R, and A Vermeule. (2009). "Conspiracy Theories: Causes and Cures." *Journal of Political Philosophy* 17 (2): 202–27.