REPHRAIN

Protecting citizens online

Making sense of the Twitter Takeover

REPHRAIN Synthesis Report on online safety, harms and wellbeing in social media

September 2023

UK Research and Innovation · University of Bristol · UCL · The University of Edinburgh · King's College London · University of Bath

Table of Contents

1. Introduction

Twitter has become an integral part of the social media landscape, enabling its users to communicate, share content and participate in debates on a wide range of topics. Over the past few years, however, the platform has been increasingly associated with online harms, including hate speech, harassment, and disinformation[1]. The recent acquisition of Twitter by Elon Musk in October 2022 reinvigorated these issues. As its new CEO, Elon Musk came under fire for dismissing its online safety teams and relaxing its approach to content moderation[2]. Some commentators claim that the future of Twitter is uncertain, as the company loses money, advertisers withdraw, and key servers lack staff to maintain them[3]. Further, between November 2022 – May 2023, over 2.5 million people were reported to set up new accounts on alternative social media providers, such as Mastodon[4].

The unfolding situation highlights the need for a regulatory response, as the current internal online safety policies are insufficient to tackle the scale and complexity of harms[5]. Policymakers in the UK must work towards incorporating social media platforms in their ongoing work on the upcoming flagship digital regulations such as the Online Safety Bill or the Data Protection and Digital Information Bill.

Furthermore, researchers and industry practitioners ought to take an active role in understanding and quantifying harmful phenomena as well as co-creating mitigations to detect misinformation, assist with content moderation or facilitate digital wellbeing in general. Research can also help policymakers to identify the most pressing issues and develop evidence-based policies to address them. Ultimately, the stakeholders face a timely opportunity to co-create social media ecosystems which benefit the society and facilitate well-informed, healthy, and civil discourse.

This report acts as a unified call of subject matter experts to govern and investigate the changing landscape of harms on social media. We are writing on behalf of REPHRAIN (the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online), the UK's world-leading interdisciplinary community focused on the protection of citizens online. As a UKRI-funded National Research Centre, we boast a critical mass of over 130 internationally leading experts at 17 UK institutions working across 46 diverse research projects and 23 founding industry, non-profit, government, law, regulation, and international research centre partners. As an interdisciplinary and engaged research group, we work collaboratively on addressing the three following missions:

- Delivering privacy at scale while mitigating its misuse to inflict harms;
- Minimising harms while maximising benefits from a sharing-driven digital economy;
- Balancing individual agency vs. social good.

[1] https://time.com/5482390/twitter-online-abuse-women-amnesty-international-study/
[2] https://www.bbc.co.uk/news/technology-64804007
[3] https://www.npr.org/2022/11/25/1139180002/twitter-loses-50-top-advertisers-elon-musk
[4] https://www.wired.com/story/the-mastodon-bump-is-now-a-slump/
[5] https://www.forbes.com/sites/christianstadler/2022/10/28/twitter-needs-more-regulation-not-less-for-elon-musk-to-advance-free-speech-and-help-humanity/?sh=fedd1e03fec4

Our researchers have extensive expertise in identifying, contextualising, and mitigating online harms in the context of social media and online communities spanning across disciplines like Computer Science, Psychology, Human-Computer Interactions, Science and Technology Studies, Politics, among the others.

In December 2022, we organised a workshop for the REPHRAIN researchers inviting them to respond to the evolving concerns about Twitter. In the following sections, we share the findings from the workshop, synthesise the expertise present across the research centre and bring attention to the remaining areas of contestation and intervention. Six months on, we reflect on sociotechnical complexities associated with moderation, verification, and online abuse. Finally, we call for an infrastructural approach to social media research and governance.

The aims of this report are to:

- outline the research landscape on online safety in social media, paying particular attention to Twitter and its potential alternatives;
- point to further research questions and priority areas;
- suggest evidence-based recommendations to policy makers and social media companies;
- signpost to the relevant and authoritative resources, such as peer-reviewed papers, datasets, and reports.

The target audience are practitioners and researchers of online harms, safety, and wellbeing on social media, including government bodies (e.g., Ofcom, the National Cyber Security Centre, National Crime Agency, DSIT), campaign and activist groups (e.g., Open Rights Group, the Tor Project, Childnet), social media companies (e.g., Twitter, Meta, Reddit), content moderators, fact checking services (e.g., Full Fact, BBC Reality Check), and software developers.

Specifically, the following researchers contributed to the formulation of this report (in alphabetical order): Prof Madeline Carr, Dr Ignacio Castro, Dr Alicia Cork, Dr Partha Das Chowdhury, Dr Andrés Domínguez Hernández, Prof Stephan Lewandowsky, Prof Corinne May-Chahal, Dr Inah Omoronyia, Dr Kopo Marvin Ramokapane, Prof Awais Rashid, Prof Massimo Renzo, Robert Schultz-Graham, Dr Laura Smith, Dr Gareth Tyson, Dr Mark Warner. The white paper was edited by Dr Ola Michalec (Policy Engagement Associate).

Please cite this report as: Michalec, O., Carr, M., Castro, I., Cork, A., Das Chowdhury, P., Domínguez Hernández, A., Lewandowsky, S., May-Chahal, C., Omoronyia, I., Ramokapane, K.M., Rashid, A., Renzo, M., Schultz-Graham, R., Smith, L., Tyson, G. Warner, W., (2023) Making Sense of the Twitter Takeover. REPHRAIN Synthesis Report on online safety, harms, and wellbeing in social media.

2. Relevant projects from the REPHRAIN Centre

The REPHRAIN Centre has mobilised expertise in the following areas related to online safety in social media environments:

- Project AUTAPP uses a combination of novel text mining and image/video analysis techniques that are able to flag a range of online harms on social media to explore the potential of automated harm detection methods (Dr Claudia Peersman, Dr Minhao Zhang, Dr Rohit Nautiyal).
- Project CLARITI is a multimodal machine learning based study of medical misinformation on social networks. The project researchers developed a model to understand the key roles of those promoting misinformation, how misinformation spreads and how to detect it in an automated way (Dr Ryan McConville, Dr Dan Saattrup Nielsen).
- Project DSNmod aims to tackle challenges of decentralised social networks by minimising online harms and exploring privacy-preserving federated moderation (Dr Ahmed M. Abdelmoniem, Dr Ignacio Castro, Dr Gareth Tyson).
- The HARM project provides a framework for categorising and anticipating online harms (Prof Adam Joinson, Prof Danaë Stanton-Fraser, Prof David Ellis, Dr Laura Smith, Dr Alicia Cork, Dr Othman Esoul).
- The INTERACT project builds on the taxonomy of harms to qualitatively define and quantitatively measure behavioural interactions as they relate to psychological harms. The researchers study how often individuals are exposed to harmful content online, who is exposed to harmful content, the types of harmful content that individuals see and the impact of the content. (Prof David Ellis, Prof Danaë Stanton-Fraser, Dr Othman Esoul).
- Project Key2Kindness investigates the role of adaptive (in the moment) hate speech awareness mechanisms to tackle online abuse (Dr Mark Warner, Dr Angelika Strohmayer, Dr Biju Issac, Prof Lynne Coventry).
- The MITIGATE project studies how various interventions (takedown, moderation, blocking accounts) can tackle online harm effectively and how to encourage community reporting of potential harmful content and interactions (Dr Laura Smith, Prof Adam Joinson, Dr Othman Esoul).
- The NEWS Project created a model predicting personality from news consumption online in order to feed into the design of interventions that can detect when political message content is matched for consumption by particular personalities and inform users when material they are reading is suspiciously tailored to their own personality (Dr Matthew Edwards, Prof Stephan Lewandowsky, Dr Barney Craggs, Dr Adam Sutton).
- Project PROACTIVE studies fringe communities to increase the understanding of the role they play in the online harms ecosystem with respect to the actions they cause on other, bigger platforms (Prof Emiliano De Cristofaro, Dr Kostantinos Papadamou).

- Project PROM investigates the role of alternative online subcultures in the promotion of violence, through a data-driven comparative analysis of content across several (Chan) platforms (Dr Guillermo Suarez-Tangil, Dr José Such, Dr Ashwini Singh).
- Project MANIPU offers a philosophical analysis of online manipulation on social media platforms focusing on the disruption of democratic processes (Prof Massimo Renzo, Dr Kartik Upadhaya).
- The PhD project "Sensemaking and conspiracy theories" harnesses the concept of collective sense making to better understand how conspiratorial narratives and rumours develop at scale in online environments (Emily Godwin).
- The Responsible Innovation Strand established an interdisciplinary collaboration with project CLARITI to embed social and ethical considerations into the development of AI tools for combatting misinformation on social media (Dr Andrés Domínguez Hernández).

3. Current research landscape

3.1. Content moderation

3.1.1. Background

Content moderation on Twitter refers to the process of monitoring, removing, or flagging content that violates the platform's community guidelines. Twitter's guidelines prohibit a range of content, including violent speech, child sexual exploitation, abuse, hateful conduct, promotion of self-harm and facilitating transactions of illegal goods[6]. However, enforcing these guidelines has proved challenging, with Twitter often facing criticism for inconsistency in enforcing its policies[7].

One of the key debates surrounding content moderation on Twitter focuses on drawing the line between free speech and hate speech. Twitter states that "*You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease*"[8]. In practice, violations to community norms of this sort can be complicated to detect and enforce as the meaning and intention of communication depends on the context and culture and, therefore, requires well-resourced human moderation to complement computational tools. This is particularly pertinent in debates on the boundary between parody and offense.

Recent mass layoffs to Twitter's Trust and Safety teams have alerted the experts to the platform's limited capacity to moderate. The investigation by BBC reveals that since the change of leadership to Elon Musk in October 2022, there has been a 69% increase in new accounts following misogynistic and abusive profiles[9]. While Twitter rebuts the criticism, claiming it removed over 400 000 accounts to make the platform safer, there is a lot of uncertainty regarding the future moderation provisions as the company is undergoing internal restructuring[10]. Twitter's poor compliance with the EU Code of Practice on Disinformation has also attracted widespread criticism, with accusation of inaccurate reporting and plagiarism[11]. Finally, the ongoing work on the UK's Online Safety Bill plans to include content moderation, emphasising 'safety by design' and obligation to remove illegal content[12].

3.1.2. Our research on content moderation

REPHRAIN researchers have advocated to privacy-preserving content moderation in the context of child sexual abuse material (CSAM), which is often circulated via social media

[6] https://help.twitter.com/en/rules-and-policies/twitter-rules
[7] https://www.wired.co.uk/article/twitters-moderation-system-is-in-tatters
[8] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[9] https://www.bbc.co.uk/news/technology-64804007

[10] https://www.bbc.co.uk/news/technology-64804007
[11] https://techcrunch.com/2023/02/09/elon-musk-twitter-eu-disinformation-code-report, pers comm - Lewandowsky, S. (2023). Plagiarism analysis of Twitter's 2022 report of compliance
[12] https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/overview-of-expected-impact-of-changes-to-the-online-safety-bill

platforms and adjacent messaging services. There is an active debate on the appropriateness of government surveillance for safeguarding the most vulnerable members of our society. An example of a recent technological development is client-side scanning (CSS) which enables on-device analysis of data in the clear (Abelson et al., 2021). If targeted information were detected, its existence and, potentially, its source, would be revealed to the agencies; otherwise, little or no information would leave the client device. REPHRAIN researchers highlight that CSS creates security and privacy risks; it can easily be repurposed as mass surveillance tool and abused by adversaries ranging from hostile state actors to intimate partner abusers (Abelson et al., 2021). Ultimately, the introduction of this technology is inherently dangerous due to the pressure to expand its scope (i.e., from child abusers only to all citizens) and the associated chilling effect, where law-abiding citizens do not feel comfortable to express their opinions and communicate online (Abelson et al., 2021).

In order to address the criticisms aimed at detection and prevention of child abuse, REPHRAIN researchers created an evaluation framework on CSAM prevention and detection tools in the context of End-to-end encryption environments (Peersman et al., 2023). The framework is applicable to encrypted messaging services, commonly found in social media platforms, and recently trialled on Twitter (see this news announcement from May 2023[13]). The report applies evaluation criteria to a case study of five Proof-of-Concept (PoC) tools funded by the DCMS Safety Tech Challenge Fund.

The evaluation framework highlights the inherent difficulties in balancing the rights of all relevant parties (i.e., law-abiding users, (potential) CSAM victims, and perceived perpetrators, the Police forces) (Peersman et al., 2023).

3.1.3. Workshop findings

The discussions during the REPHRAIN workshop focused on the technical challenges of automated content moderation infrastructure. Researchers debated whether content moderation should be placed on device or in the networked infrastructure. Such automated moderation systems could be used to, for example, detect nudity online and be placed on devices used by children.

Current *on-device* moderation relies on permissions and direct memory access to the whole device, which is deemed a security risk. At the level of operating system, software processes should be siloed so that apps cannot access each other's data. REPHRAIN researchers explored this dilemma by researching mechanisms to moderate in an end-to-end encrypted fashion, such as supported by Signal or WhatsApp (Agarwal et al., 2022). These techniques allowed companies to offer users end-to-end encrypted message privacy, while simultaneously supporting the automatic detection of spam users. The work experimented with several models, building techniques that could operate both centrally (I.e., in WhatsApp's servers) and on-device (I.e., on users' phones).

[13] https://www.cnet.com/tech/services-and-software/twitter-now-offers-encrypted-dms-but-not-everyone-can-send-them/

Participants also noted that *networked* approach to moderation would support data protection capabilities better than on-device moderation as it facilitates data portability, removal and reporting to appropriate authorities where relevant.

In addition to technical considerations, REPHRAIN researchers also found that the public is broadly in favour of moderation and removal of problematic content. In the conjoint survey experiment study, the researchers systematically varied factors that could influence moral judgments and found that despite significant differences along political lines, most US citizens preferred quashing harmful misinformation over protecting free speech (Kozyreva et al., 2023).

3.2. Mis-/Disinformation

3.2.1. Background

Detection, downranking and banning incorrect claims are the key ways to tackle the spread of misinformation on social media. One way this could be achieved is by machine learning (ML) algorithms which are trained on authoritative corpus of data (such as Wikipedia, official documents or trustworthy fact-checking organisations) and either help human moderation or automate decisions around banning or downranking (see the outputs from CLARITI project).

It is important to distinguish between m/disinformation (which can be counteracted with moderation and reliance of authoritative sources, i.e., a claim that climate change is caused by sunspots) and complex and multifaceted debates (i.e., implementing effective climate action in particular areas while balancing it with justice concerns and available budget). With regards to the latter, it is crucial for social media platforms to curate discussion environments, where people can be challenged without aggression and where they can be afforded to change their minds without shame. These environments ought to embrace plurality, humility and opening up expertise to a wide range of stakeholders who would learn from each other's inherently partial perspectives[14]. One of the greatest threats to this goal is the fact that algorithms tend to recommend similar content to similar looking users, which leads to the creation of "filter bubbles". Instead of being challenged, the beliefs and behaviours of users are constantly reinforced because algorithms recommend content that is in line with their pre-existing beliefs. Moreover, some researchers have suggested that over time, filter bubbles push users towards increasingly extreme content[15].

3.2.2. Our research on mis-/disinformation on Twitter

For the readers interested in the process of building ML to detect misinformation on Twitter, project CLARITI published a paper (Nielsen and McConville, 2022b), where the researchers released and described the dataset built to train a ML model to verify social media claims. The dataset contains a rich variety of social media information (tweets, replies, users, images, articles, hashtags), spans 21 million tweets belonging to 26 thousand Twitter threads, each of which have been semantically linked to 13 thousand fact-checked claims across dozens of

[14] https://philpapers.org/archive/harskt.pdf

[15] Pariser, E., 2012. The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think, Reprint edition. ed. Penguin BooksReprint edition.

topics, events, and domains, in 41 different languages, spanning more than a decade. The dataset itself can be accessed here for those who would like to continue research in that area.

Responsible innovation Strand researchers analysed the ethical and political issues related to the automated misinformation detection in a recent study. The findings of the study suggest that the data used to train ML moderation models can contain errors, biases, inaccuracies and uncertainties which are amplified by such systems. There are risks with the use of automation as ML models are susceptible not only to errors and uncertainty, but to implicit assumptions around the authoritativeness of knowledge sources. This is particularly the case when these systems are made opaque by technology companies. Oversight and human intervention are needed to avoid problems such as falsely categorising misinformation which could undermine people's trust in public information or reinforce false beliefs. There is currently limited scrutiny or self-reporting of how platforms use these tools and the design decisions that underpin them. The authors concluded with a series of recommendations for the responsible development of automated moderation tools (Domínguez Hernández et al., 2023).

Furthermore, researchers from the NEWS project (Lasser et al., 2022) explored alternative conceptions of honesty and facts by the U.S. politicians on social media. Here, they show that in the last decade, U.S. politicians' conception of truth has undergone a distinct shift, with authentic but evidence-free belief-speaking becoming more prominent and more differentiated from evidence-based truth seeking. The paper analyses communications by members of the U.S. Congress on Twitter between 2011 and 2022 and show that political speech has fractured into two distinct components related to belief-speaking and evidence-based truth-seeking, respectively, and that belief-speaking, but not truth-seeking, can be associated with the sharing of untrustworthy information. By contrast, increase in truth-seeking language in tweets and articles is associated with an increase in verifiable sources. The results support the hypothesis that the current dissemination of misinformation in political discourse is in part driven by a new understanding of truth and honesty that has replaced reliance on evidence with the invocation of subjective belief.

Researchers from the MANIPU project have considered a number of ways in which the design of social media platforms can be exploited by information manipulation campaigns (Renzo et al., 2023). This study does three things: it surveys how different features of new digital technologies have radically changed the scope, scale, and precision of information manipulation campaigns; it assesses the distinctive ways in which people’s beliefs, emotions, and attention are compromised by these campaigns; it outlines a range of measures that can be taken to address these issues.

Finally, we highlight that conducting large-scale studies on m/disinformation and the flow of communication on social media more generally depends on researchers’ access to usable, transparent and affordable Application Programming Interfaces (APIs), or third-party software which allows other programs to communicate with each other. The recent introduction of charges (with fees ranging from $42,000 to $210,000 per month, as of May 2023) for accessing third-party applications and analysis tools could severely impact the whole computational

social science research community and hinder research on major public issues, such as military propaganda, racial discrimination or radicalisation[16].

3.3. Polarising debates, trolls and enraging recommender algorithms

3.3.1. Background

Harm on Twitter can take many forms, including cyberbullying, hate speech, and doxing (the public release of personal, often embarrassing, information). The prevalence of harmful behaviour on the platform has been linked to a number of negative outcomes, including mental health problems, social isolation, and a decline in civil discourse[17].

Enraging recommender algorithms are designed to maximise user engagement by showing them content that is most likely to elicit a negative emotional response, such as anger or fear. These algorithms can also contribute to the spread of misinformation and extremist views. For example, if a user shows an interest in conspiracy theories, the algorithm may recommend increasingly extreme content, leading the user down a rabbit hole of false information and radical beliefs[18]. Campaign organisations, such as Open Rights Group, have called for greater transparency and oversight around the algorithms that power the platform's content recommendations, arguing that users have a right to know how they are being influenced[19]. In response, UK regulators are working towards frameworks for auditing algorithms against potential harms[20].

Twitter has long struggled to address the issue of trolls on its platform. While the company has taken steps to combat abusive behaviour, implementing reporting mechanisms and banning users who violate its policies, many users continue to experience harassment and abuse[21]. Trolls can be particularly damaging to people who are already marginalised, such as women, people of colour, and members of the LGBTQ+ community.

3.3.2. Our research on polarising debates on Twitter

REPHRAIN researchers (from NEWS and HARM projects) explored the drivers and language of political tribalism and polarisation on Twitter. For example, a recent paper by North et al. (2021) studies how distinct online communities ('Leavers' and 'Remainers') emerged in the aftermath of Brexit. The data used were 32 months of discussions (n = 9,027,822) on Twitter, where researchers used identity-based keywords as proxies for tribalism. The analysis finds that four group identity keywords are used more frequently over time, suggesting an increase in tribal interactions. There is also evidence of a relationship between real-life Brexit events and spikes in tribal responses online.

[16] https://www.washingtonpost.com/technology/2023/06/20/twitter-policy-elon-musk-api/
[17] https://dl.acm.org/doi/abs/10.1145/2818052.2869107
[18] https://www.computer.org/csdl/magazine/co/2014/12/mco2014120090/13rRUxYIMYR
[19] https://www.openrightsgroup.org/campaign/stop-data-discrimination/
[20] https://www.gov.uk/government/publications/algorithms-how-they-can-reduce-competition-and-harm-consumers/algorithms-how-they-can-reduce-competition-and-harm-consumers#the-role-of-regulators-in-addressing-these-harms
[21] https://www.bbc.co.uk/news/technology-64804007

Further, researchers argue that trolls (including foreign actors) use social media to sow discord among Americans through political polarisation (Simchon et al., 2022). They presented an open-source linguistic tool to gauge polarised discourse on social media and found that three distinct troll populations, which hold anti-American views, used polarised language more than the average American user. This research provides insights into the mechanism through which trolls function, and sheds light on the role of language and social media in politics online.

Finally, REPHRAIN researchers called for a shift from individual-level to group-level analysis, in order to trace the formation of radicalising social interactions and group identities (Smith et al., 2019). An example of such approach was a study by Brown et al (2022), investigating how different ideological groups justified and mobilised collective action online. Here, the researchers collected 6878 posts from the Twitter and Telegram accounts of pro-Black Lives Matter (n = 13) and anti-Black Lives Matter (n = 9) groups. They found that both groups perceived their action as 'system-challenging', with pro-BLM accounts focused more on outgroup actions to mobilise collective action, and anti-BLM accounts focused more on ingroup identity. The implications are that groups' ideology and socio-structural position should be accounted for when understanding differences in how and why groups mobilise through online interactions.

3.3.3. Workshop findings

During the workshop, REPHRAIN researchers highlighted the problem of low value 'wasteful' social media content, that is simultaneously highly visible and viral. This is heightened by the application of opaque recommender algorithm, which prioritises accounts with numerous followers at the expense of accounts presenting content that might be relevant to the user. While this type of content can be difficult to categorise as directly harmful, researchers cautioned against the potential for poor user experience (i.e., emerging creators struggling to gain visibility, difficulty in finding useful content) or even addiction.

Researchers also discussed the appropriateness of metaphors used to describe Twitter and its users' capabilities to get involved in civic discourse. They highlighted that the commonly used metaphors of 'town hall' or 'public square' are not accurate as Twitter users are not all visible in an equitable or democratic way due to the unfair recommender algorithm. A 'town hall' model of social media would imply an access to the structured discussion with people who care about a particular community or a shared issue. On the other hand, too much structure could ultimately lead to the creation of information bubbles, where users aren't exposed to people with opposing views. REPHRAIN researchers recommended pilot studies fostering a positive debating environment and investigate how users can leave their information bubbles without being subjected to abuse (for example projects conducted in collaboration with social media companies).

3.4. Verification and anonymity

3.4.1. Background

Twitter has undergone significant changes in its approach to account verification. Historically, verification has been viewed as a way to validate the authenticity of public figures (e.g.,

journalists or politicians) on a platform, using 'active, notable and authentic' as key criteria[22]. This then shifted to verification 'blue check' becoming a symbol of prestige. In recent years, Twitter faced considerable criticism as its verification process has been perceived as opaque and arbitrary[23].

In a proactive response to those growing concerns, the platform announced in 2017 that it would be temporarily suspending its verification process while re-evaluating criteria for verification. More recently, in 2022, Twitter introduced a paid subscription-based service called 'Twitter blue', which is a combination of a verification service and a 'premium service'[24]. Under the new eligibility criteria, verification is available to all paid users who are active, show no signs of being misleading, display complete name and photo, and provide a confirmed phone number. In parallel, news outlets report that Twitter is developing a separate service for verifying organisations for $1, 000/month[25]. The ongoing changes are a subject of controversy, as critics claim they diminish the value of verification[26].

3.4.2. Workshop findings

The workshop brought forth a range of interesting insights on how verification is changing in its meaning and purpose. Participants highlighted five main aspects of verification, namely 1) as an anti-impersonation safety measure, 2) a charitable service to the community, 3) a service increasing the reputation of some customers, 4) a business proposition of a premium service and 5) a hypothetical widespread policy. Participants also noted that verification debates don't always travel the same way outside of the content of Twitter, i.e., verification is not a common feature in cases of Reddit[27] or email while it's obligatory for mainstream financial services. The view on verification depends on local norms and its necessity varies across different contexts.

The existing approach to verification has been criticised by participants for focusing only on the protection of selected minorities with large followings. Our experts raised concerns about the duty of care to vulnerable users who may be susceptible to bullying or false claims. Workshop discussants agreed that Twitter's early conceptualisation of verification was based on users who were authentic, active, and notable. However, verification has become a badge of popularity in recent years, and the criteria for who deserves or needs verification have become blurred.

The ever-changing business model of verification, with Twitter Blue users being treated as a priority for the recommender algorithms, has posed questions about the future of user engagement on the platform. Twitter's introduction of effectively new tiers of service has transformed verification from a safety measure to a premium service. Due to the rapid pace of changes, there is a risk of confusion and misinformation about the changing understanding of

[22] https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts
[23] https://www.theverge.com/2022/10/31/23432816/elon-musk-twitter-verification-subscription-charge-trust-problems
[24] https://blog.twitter.com/en_us/topics/product/2022/twitter-blue-update
[25] https://help.twitter.com/en/using-twitter/verified-organizations
[26] https://www.theverge.com/2022/10/31/23432816/elon-musk-twitter-verification-subscription-charge-trust-problems
[27] With the exception of 'Not Safe for Work' subreddits

verification, the potential harm of an unfair recommender algorithm and a lack of safety measures for those who need them.

The workshop also discussed the potential of verification as a widespread policy. In this case, verification could have negative consequences for freedom of speech in authoritarian regimes. There are practical and economic challenges associated with verification on a large scale, for example, while verification is important for preventing misinformation for those with a large audience, it may not be the most cost-effective policy for small accounts.

Participants acknowledged the debates on privacy and encryption as a right, cautioning against hasty decisions on this issue. Multiple identities, personas, and accounts are valid, and people use social media for different reasons and may want to compartmentalise their lives for their well-being. Anonymity is unlikely to go away from the internet, and new platforms prone to moving harmful content elsewhere may emerge as a result.

Finally, the workshop experts stressed that the introduction of the 'Blue Tick' verification system accelerated a shift in the common understanding of this feature. Researchers and practitioners need to intervene to deal with this shift in a timely manner. In particular, there is a need for verification-as-safety mechanisms to continue protecting users who stopped receiving support from Twitter.

3.5. Decentralised social media – an alternative to Twitter?

3.5.1. Background

Decentralised Social Networks (e.g., Mastodon, Pleroma) have become popular over the past few years. These services offer microblogging services similar to centralised applications like Twitter. More recently, the acquisition of Twitter has brought a large number of users to Mastodon, with 2.5 million new users registered between October 2022 and January 2023. While the future of decentralised social media is uncertain (for example, Mastodon is already reporting a drop in 'active users' in early 2023[28]), the evolving Twitter takeover has raised the profile of what previously were niche initiatives for open-source enthusiasts. What's more, Twitter itself has expressed interest in developing a decentralised social network protocol, with the Bluesky spin-out initiative founded in 2021[29].

It is important to note that decentralised social media are not only competitors to Twitter but also present an alternative vision for the future of the Internet. Their distinct offer lies in the infrastructure underpinning the services: a lack of centralised owner/server, reliance on voluntary moderation and maintenance, and grounding in the Free and Open-Source culture. A radically different infrastructure provides an opportunity to reinvent business models of social media, however, it also presents a set of new challenges pertaining to online harms, software engineering and politics.

[28] https://www.theguardian.com/news/datablog/2023/jan/08/elon-musk-drove-more-than-a-million-people-to-mastodon-but-many-arent-sticking-around
[29] https://blueskyweb.xyz/

3.5.2. Our research on decentralised social media: architecture

Despite the debates around Mastodon migration focusing primarily on the user interface (e.g., differences in terminology like 'tooting' and 'tweeting'), the key contrast between the Twitter platform and Mastodon (supported by the ActivityPub protocol) lies in the underlying infrastructure. A recent paper by DSNmod project researchers, outlines potential benefits and challenges to decentralised web architecture (Raman *et al*., 2019). While detailing the concepts of decentralised web is outside the scope of this report, it is worth highlighting that decentralised social media protocols have the following in-built mechanisms:

- Open source: anyone could set up an independent server ('instance') that users can sign up to;
- Transparency: data ownership can be more transparent, as users can setup and manage their own independent instances. This can provide them with more control over their data.
- Interoperability: servers build on top of federated protocols[30] so that they can work together, in a peer-to-peer fashion[31], appearing as a globally integrated service;
- Decentralisation: there is no single owner or controlling authority (Raman *et al*., 2019).

Decentralised architecture, however, carries inherent challenges. For example, it is unclear how these systems might scale up, how wide-ranging malicious actors (e.g., spam bots) could be detected or how users could be protected from data loss during outages (Raman et al., 2019). Finally, Mastodon's decentralised architecture displays tendencies towards centralisation, i.e., the top 25% most populous instances contain 96% of the users. This pressure is counterbalanced by the greater activity of the users on smaller instances. On average, users of single user instances post 121% more statuses than users on bigger instances (Bin Zia et al., 2023). The attacks facilitated due to malicious home servers can be an extended in case of locally managed servers[32].

3.5.3. Our research on decentralised social media: Digital literacy

The rise of Mastodon's popularity has led to users' confusion about terminology, resulting in conflicts between established and new accounts[33]. In response, journalists offered numerous 'how to' guides[34]. Within the REPHRAIN Centre, researchers in the Responsible Innovation strand argue that further significant investment will be needed in digital and futures literacy in order to ensure safe adoption of decentralised web technologies[35]. The initiatives should include a variety of stakeholders such as users, non-users, academics, journalists, and policy makers. These interventions ought to address security, privacy protection, safeguarding and wider ethical concerns raised by the potential mass adoption of decentralised social media. Literacy interventions should focus on improving the ability to use new interfaces, increasing

[30] Here, federation refers to a pattern in network architecture that allows interoperability and information sharing between semi-autonomous and de-centrally organised applications or users

[31] A peer-to-peer (P2P) network is created when two or more computers are connected and share resources without going through a separate server.

[32] Albrecht, Martin R., et al. "Practically-exploitable cryptographic vulnerabilities in Matrix." *Cryptology ePrint Archive* (2023).

[33] https://www.indy100.com/science-tech/mastodon-harry-potter-hogwarts-legacy

[34] https://www.wired.com/story/how-to-get-started-use-mastodon/

[35] https://bpb-eu-w2.wpmucdn.com/blogs.bristol.ac.uk/dist/1/670/files/2023/03/Call-for-Papers-by-the-All-Party-Parliamentary-Group-REPHRAIN-Response.pdf

users' agency over their preferences, understanding of the potential risks, and anticipating individual and social harms. These actions could be taken up by experts in digital literacy and education, human computer interaction, usability, and developer communities in collaboration with policy stakeholders like Ofcom.

3.5.4. Our research on decentralised social media: Content moderation

Studies from DSNmod researchers explored the challenges of content moderation in decentralised web. A critical difference with Twitter is that decentralised social networks are composed of independent servers (i.e., instances) that are moderated by independent administrators. Users, however, can interact with each other regardless of the instance where their account is providing a similar service and perception to their centralised alternative (i.e., Twitter). This decentralisation introduces new challenges, among which DSNmod researchers identify: 1) a heavy moderation load on administrators (Anaobi et al, 2023), and 2) fundamental limitations in the current moderation tools such as blocking all users from an instance rather than just the offenders (Anaobi et al, 2021). As a result, the researchers find widespread presence of toxic content (Anaobi et al., 2021; Bin Zia et al., 2022).

The work also identifies avenues to address these challenges: 1) new streamlined user-driven policies that enable administrators to moderate on a per-user (rather than per-instance) basis in an easier and potentially semiautomated fashion (Anaobi et al., 2021), 2) streamlining moderation with the assistance of automated classifiers (a technique relying on machine learning to classify information as complying with moderation policy) (Bin Zia et al, 2022). The latter however is complicated by the decentralisation. This is because of the economies of scale of machine learning, where: classifiers become more effective when trained on larger pools of data, but differently from Twitter where all content can be aggregated, in decentralised social networks, this information is scattered across instances.

In order to overcome the problem of economies of scale in decentralisation, the researchers show how federated learning (i.e., machine learning techniques which train an algorithm across multiple decentralised servers holding local data samples, without exchanging them) allows multiple entities to train an algorithm without sharing the actual data. This privacy preserving approach can help reduce the barrier of entry for services where large economies of scale can deter new entrants and reduce competition. This is particularly the case for social networks, where economies of scale and network effects have frequently resulted in limited competition and monopolistic behaviours (Bin Zia et al., 2022).

3.5.5. Workshop findings

REPHRAIN researchers, while expressing openness and curiosity about the development of decentralised social media, expressed caution about their possible futures. Migration to Mastodon and other services is still a dynamic phenomenon, as are the evolving cultures and moderation norms. On a technical level, the dependencies between various fediverse services are not yet fully understood, which might ultimately impact the performance and usability of decentralised social media.

Regardless of the future developments in decentralised social networks, the period of uncertainty at Twitter opens new possibilities to reimagine what social media should look like. Researchers, practitioners, and lay users alike have a unique opportunity to voice their

preferences regarding digital safety, wellbeing and positive experiences on Twitter and its alternatives. The sections below will summarise workshop discussions regarding the future of decentralised social media, focusing on affordances, power, content moderation and server hosting.

Affordances

During the workshop, REPHRAIN experts discussed the different affordances of social media such as Twitter or Mastodon. First, it is important to note that these services historically had different target audiences, resulting in divergence in community norms (for example, content warnings for spoilers or trigger are more common on Mastodon than Twitter[36]). The current influx of new users of Mastodon may cause discomfort for its early adopters and require a restating of social norms and best practices.

Participants flagged that Mastodon has several different affordances to Twitter, including a chronologic timeline and the need to manually retweet (or reblog) to let it appear on other users' timelines. This contrasts with Twitter where 'liking' a post can contribute to its visibility on others' timelines due to the algorithmic recommender system. Despite this shift, Mastodon's interface is similar enough to Twitter that users do not necessarily think of accounts on other instances as separate (i.e., they have an option to both view the 'local' timeline of their instance and the 'home' timeline of all accounts they follow across instances).

According to the workshop discussants, currently, the network effect on Mastodon and other decentralised social media platforms is not yet sufficient for the users to enjoy the 'buzz' of information and opinions. This means that decentralised social media seem 'quieter' than traditional platforms such as Twitter.

Moving away from the discussion on user interfaces, participants emphasised that the key difference between Mastodon and Twitter lies in the matters of ownership and administration. The decentralised protocol has been dubbed a 'game changer'. Precisely, it's the interoperability between protocols which creates a so-called fediverse. The fediverse refers to the concept of plurality, open-source architecture, community ownership and interoperability.

Although decentralised social media bring promises of democratising communication, there are serious implications for content management, harms, and moderation. For example, one can create their own recommender algorithm or decide on their own moderation rules. This presents both opportunities and challenges for administrators of Mastodon instances.

Going forward, to promote innovation in social media environments, participants recommended that it should be easier to migrate across social media platforms with their own data. This should be mandated in regulations to prevent social media companies from holding excessive user data.

Power, centralisation and monopolisation

Decentralised social media are often touted as a solution to the issues of power and monopolisation that have plagued centralised platforms. However, as workshop participants noted, it is important to be wary of conflating decentralisation in a technical versus political

[36] https://www.dailydot.com/debug/mastadon-content-warnings-twitter/

sense. They pointed out there is a false dilemma between 'centralised' and 'decentralised' systems. In the words of the Responsible Innovation Strand researcher, Domínguez Hernández (2023), decentralisation is best understood as an in-flux dynamic concept that is negotiated between different actors.

According to workshop participants, the key question concerning power and (de)centralisation should be "how does one make a judgement about what a 'good' system design looks like?". Participants explained that in a classic software engineering approach, the process of system design would start with agreeing on the requirements for the system, user needs, and answering the question of whether the architecture satisfies user needs. The concept of needs should adequately include the opportunities users have to make use of a system[37].

Further, REPHRAIN experts working for the DSNmod project recalled their study about system properties in decentralised architecture and found that there are pressures towards centralisation if the system is not run well (Raman et al., 2019). The goal within decentralised social networks (e.g., Mastodon) has always been to better distribute power and control. However, the researchers found that, in practice, users tend to centralise on a small number of servers (rather than creating their own server or joining a smaller community). This runs the risk of simply creating large central players. Thus, their research found that most users converge on a few favourites, leading to a tendency towards centralisation and monopoly. Going further, there is a need for developing privacy-preserving techniques to reduce monopolisation and ensure that power is not concentrated in the hands of a few.

Finally, workshop participants argued that we also need to understand how computer networks concentrate power in terms of acquisitions and unfair competition. These processes should be closely monitored and regulated by bodies such as the Competitions and Market Authority.

Content moderation

During the workshop, DSNmod researchers highlighted that the rise of decentralised social media platforms has brought the challenge of content moderation to the fore. In case of Mastodon, for example, this is happening on a 'per instance' rather than 'per user' basis. This challenge stems from the fact that moderators cannot control content that is exported from other instances. While instances can filter how they interact with other instances, this does not completely solve the problem of decentralised moderation. Additionally, automatic blocking of the whole instance can be problematic as it blocks all content, not just the content of a specific user.

On the other hand, one of the benefits of decentralised moderation is the plurality of moderation policies, as users can leave an instance if they disagree with the moderation policies and move to another instance that aligns with their values. This was dubbed a 'marketplace moderation' approach, as it encourages self-accountability and ownership of the rules, akin to the original analogy of social media as a town hall. However, scaling up self-accountability can be challenging, and it is important to account for the potentially harmful

[37] Chowdhury, Partha Das, et al. "From Utility to Capability: A Manifesto for Equitable Security and Privacy for All." (2023).

experiences of readers and lurkers, i.e., people who read social media but don't interact or write their own content.

When self-regulation fails then regulators need to step in. Setting policies at a higher layer needs to be backed with mechanisms to check compliance at a lower layer. A mesh of interconnected servers residing within diverse administrative boundaries makes verifying compliance complex. As a potential solution, participants suggested that prioritisation of content rather than moderation could be technically easier and cheaper, but this would come at the cost of community norms and standards. There is also a paradox of decent moderation as social media scales up, as increased moderation can result in a creation of information bubbles and the loss of the randomness of interactions. Additionally, complications with moderation arise from the variety of policies, which can result in wrongly banned accounts and false positives. Recognising the gaps in understanding and policy evidence, workshop experts called for further geopolitics and HCI research to determine legal accountability of moderation.

Hosting a server

Finally, workshop participants discussed the feasibility of organisations hosting Mastodon servers (or instances). This could be a practical solution for institutions (e.g., universities or governmental bodies) looking to establish self-managed and private social networks. However, this comes with a lot of labour-intensive tasks, such as dealing with user requests, complaints, and moderation. Without appropriate resourcing, organisations might find it difficult to handle these tasks efficiently, however, as one participant flagged, hiring a server can be an option[38].

Even if the challenges of labour and cost can be overcome, hosting a server requires careful consideration of reputational, privacy and security risks. On the one hand, running one's own Mastodon server can be a way to self-verify (as seen with European Commission instances[39]) or to manage internal communications only (similar to using Slack or Discord, which filter the audience). On the other hand, as decentralised web is still an emerging technology, there is a lack of clarity on host responsibilities with regards to monitoring and tackling harmful or false content.

3.6. A turn to infrastructural research and governance

3.6.1. Background

Social media platforms rely heavily on network infrastructure and system design to function properly. There are several issues that can arise in the context of their day-to-day operations and governance. One major challenge is the sheer volume of data that social media platforms generate. This data must be stored, processed, and analysed in real-time, which requires a significant amount of computing power and storage capacity. As a result, social media companies must maintain their infrastructure to ensure that they can handle the demands of their users.

While there has been considerable research concerning social media algorithms, the underlying infrastructure – networks, protocols etc. – have received considerably less attention. This is a critical omission as the architecture of social media platforms has political

[38] See, for example, https://masto.host/
[39] https://social.network.europa.eu/about

and ethical implications. The section below will outline the need for the 'turn to infrastructure' in social media governance and analysis.

3.6.2. Workshop findings

The REPHRAIN workshop pointed at a pressing need to re-conceptualise social media as global infrastructures, paying attention to how they support critical services, ensure public safety, and recover from incidents. This move could have significant implications for international politics, particularly during elections or social unrest periods when disruptions and outages can have severe consequences. Understanding social media as global communication infrastructures raises the questions of responsibility over content and societal dependence on access to good quality and timely information.

However, as REPHRAIN experts pointed out, there is an inherent difficulty in implementing these ideas into practice as major social media providers are private enterprises with little regulatory oversight. The lack of transparency of system architectures in social media platforms is particularly problematic[40]. Currently, it is not clear how third parties process data, who can see them, for what purposes, and how users can be informed about their data flows.

Workshop participants called for a new wave of user-centred controls over social media infrastructures – at the moment, the internal architecture of platforms like Twitter is obscure and, therefore, not conducive to trust. The need for user controls extends beyond having an option to reject or accept cookies. Users should be able to influence how data is being utilised. For example, usage of data for collective good like medical research might be preferred over usage of the same data for targeted marketing by health services companies. These findings underscore the importance of transparency and accountability of social media platforms like Twitter, and the need to balance public service responsibilities with private enterprise considerations.

3.6.3. Our research on systems design

Appropriate security policies and their effective implementation are intended to build robust and reliable software systems. However, studies of systems engineering over the last few years have demonstrated several outstanding challenges. REPHRAIN researchers showed that software developers do not fully understand the security implications of permissions they seek from users (Tahaei et al., 2023). The gaps lead to mechanisms that fail to meet the legitimate security and privacy expectations of end users. Further, Software Development Kit (SDK) monopolies do not clearly enumerate the data they collect and share. Lack of documentation, complex patch management among other things negatively affect developers to code securely (Das Chowdhury, 2021).

The paradigm of web decentralisation introduced some key changes to secure and reliable software development. Using the case of Mastodon, the process of federation created a functionality where instances can now collaborate and interact in an open-source and non-hierarchical manner, creating content without the need to cater to the recommender algorithm (Raman et al., 2019). However, the research from the REPHRAIN centre shows these efforts weren't sufficient so far in countering Mastodon's tendencies to centralise. For example, 10%

[40] https://www.opensourceforu.com/2023/04/its-a-red-herring-to-use-twitters-open-source-algorithm/

of instances are home to almost half of the users and three Autonomous Systems (e.g., Amazon or Cloudfare) host almost 2/3 of the users. This creates a potential point of failure a serves outage in a handful of instances could remove the majority of toots posted (Raman et al., 2019). This also acts as a security vulnerability, as it can become an easy target of malicious botnet attacks.

What do these observations mean for the governance of new micro-blogging platforms? The decentralised design of Mastodon means negotiating diverse security policies and trust assumptions across different system and administrative domains. Decentralisation also requires an anticipatory approach involving iterative user testing. Based on our illustrative examples, the decentralised design in one domain can lead to the experience of centralisation in another.

The above research findings have bearing on compliance. The European Union mandates that various platforms should be able to share content across them (see the Digital Markets Act[41]). Such a mandate statedly aims to allow users on various platforms to communicate with one another. This proposal mandates bigger players to allow smaller entities access their services and APIs. However, they leave the technical implementation details to the platforms. We highlight that such a proposal is fraught with security and privacy risks[42]. Among other things there is an issue of trust where a (possibly) independent service will have access to the content on user‘s device. Considerable work is required at the protocol stack to facilitate secure and private interoperability.

4. Policy and practice recommendations

Recommendations for the UK policymakers

- Include decentralised social media in algorithmic fairness and market regulation frameworks. Government bodies like the Competitions and Market Authority or Ofcom are currently designing policy frameworks aiming to detect and address monopolising effects and exclusionary behaviours in marketplaces and digital services. Emerging technologies, like decentralised social media should come under the scope of these initiatives.
- Improve literacy around the capabilities and risks of decentralised social media. Bodies like Ofcom and Department of Education should launch a public/school campaign highlighting online harms in these emerging technologies and ways to prevent or minimise them. The campaign ought to include differences in functionality of platforms vs protocols.

[41] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en

[42] Jenny Blessing & Ross Anderson, One Protocol to Rule Them All? On Securing Interoperable Messaging. Accepted/In press at the International Security Protocols Workshop, 2023

- Continue working with the category of 'legal but harmful content' in the creation of the Online Safety Bill.

Recommendations for social media services

- Twitter and other social media services ought to provide usable, affordable and transparent access to their APIs for researchers.
- Content moderation responsibilities of server hosts on Mastodon should be clarified.
- All social media services ought to follow the duty of care principle while considering diverse capabilities of internet users[43], prioritising protection of the most marginalised groups and maximising user's agency over their data, recommender algorithms and interactions with other accounts.

Overarching regulatory framework: Infrastructural approach

- Large social media companies ought to be regulated as digital infrastructures (e.g., by the EU NIS2 Directive and the upcoming UK equivalent),
- Regulating social media as infrastructure should enable: a) data portability – easy migration across services to disrupt network effects and prevent monopolisation; b) resilience in the face of network outages or large-scale cyber security attacks as social media services would have an operational incident response plan.
- Network effects of social media ought to be addressed under the EU competition law to prevent unfair mergers.

[43] https://bpb-eu-w2.wpmucdn.com/blogs.bristol.ac.uk/dist/1/670/files/2023/02/Capability-Approach-Manifesto.pdf

5. Recommended resources

REPHRAIN researchers boast a track record of high-quality scientific publications, datasets, reports, and preprints related to online harms, wellbeing, and safety in the context of social media. Below we offer a list of relevant outputs published across a variety of themes, such as content moderation, harm taxonomies, mis/dis- information, radicalisation, privacy, security.

We are happy to present these findings as policy roundtables or briefings upon request. Please contact us at rephrain-centre@bristol.ac.uk to explore your preferred methods of communication.

Content moderation

- Anaobi, I.H., Raman, A., Castro, I., Bin Zia, H., De Cristofaro, E., Sastry, N., and Tyson, G. (2021) Exploring content moderation in the decentralised web: The pleroma case. Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies, pp. 328-335.
- Anaobi, I.H., Raman, A., Castro, I., Bin Zia, H., Ibosiola , D. and Tyson, G. (2023) With Great Power comes Great Responsibility: Exploring Administration in Decentralized Social Networks. Proceedings of the Web Conference. 2023.
- Bin Zia, H., Raman, A., Castro, I., Anaobi, I.H.., De Cristofaro, E., Sastry, N., & Tyson, G. (2022). Toxicity in the decentralized web and the potential for model sharing. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 6(2), 1-25. https://dl.acm.org/doi/abs/10.1145/3530901
- Iqbal, W., Arshad, M. H., Tyson, G., & Castro, I. (2022). Exploring Crowdsourced Content Moderation Through Lens of Reddit during COVID-19. Proceedings of the 17th Asian Internet Engineering Conference (pp. 26-35). https://dl.acm.org/doi/abs/10.1145/3570748.3570753
- Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. Proceedings of the National Academy of Sciences, 120(7),
- Raman, A., Joglekar, S., Cristofaro, E. D., Sastry, N., & Tyson, G. (2019). Challenges in the decentralised web: The mastodon case. Proceedings of the internet measurement conference (pp. 217-229). https://doi.org/10.1145/3355369.3355572

Online harm taxonomies

- Cork, A., Smith, L. G., Ellis, D., Fraser, D. S., & Joinson, A. (2022). Rethinking Online Harm: A Psychological Model of Contextual Vulnerability. https://psyarxiv.com/z7re2/
- REPHRAIN Research Centre (2023) REPHRAIN Map of Online harms, risks, and vulnerabilities https://rephrain-map.co.uk/

- Hernández, A. D., Cork, A., Godwin, E. J., Michalec, O., Johnstone, E. K., Chowdhury, P. D., ... & Rashid, A. (2023). Co-creating a Transdisciplinary Map of Technology-mediated Harms, Risks and Vulnerabilities: Challenges, ambivalences and opportunities. The 2023 ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)

Mis/Dis-information

- Domínguez Hernández, A., Owen, R., Nielsen, D. S., & McConville, R. (2022). Addressing contingency in algorithmic (mis)information detection: Toward a responsible machine learning agenda. arXiv preprint. https://arxiv.org/abs/2210.09014
- Lasser, J., Aroyehun, S. T., Carrella, F., Simchon, A., Garcia, D., & Lewandowsky, S. (2022). New conceptions of truth foster misinformation in online public political discourse. arXiv preprint https://arxiv.org/abs/2208.10814
- Lasser, J., Aroyehun, S. T., Simchon, A., Carrella, F., Garcia, D., & Lewandowsky, S. (2022). Social media sharing of low-quality news sources by political elites. PNAS nexus, 1(4), https://pubmed.ncbi.nlm.nih.gov/36380855/
- Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2022). "It is just a flu": Assessing the Effect of Watch History on YouTube's Pseudoscientific Video Recommendations. Proceedings of the international AAAI conference on web and social media (Vol. 16, pp. 723-734). https://ojs.aaai.org/index.php/ICWSM/article/view/19329
- Renzo, M., Bradshaw, S. (2023) Social Media and Manipulation. In Routledge Handbook of Media Ethics, eds. C. Fox and J. Saunders
- Saattrup Nielsen, D., & McConville, R. (2022a). A Heterogeneous Graph Benchmark for Misinformation on Twitter. https://graph-learning-benchmarks.github.io/assets/papers/glb2022/A_Heterogeneous_Graph_Benchmark_for_Misinformation_on_Twitter.pdf
- Saattrup Nielsen, D., & McConville, R. (2022b). Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 3141-3153). https://dl.acm.org/doi/abs/10.1145/3477495.3531744 and https://mumin-dataset.github.io/
- Scott, L., Coventry, L., Cecchinato, M., & Warner, M. (2023). "I figured her feeling a little bit bad was worth it to not spread that kind of hate": Exploring how UK families discuss and challenge misinformation. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), Hamburg, Germany https://discovery.ucl.ac.uk/id/eprint/10163917/

Political tribalism, radicalisation, polarisation

- Aran, X. F., Van Nuenen, T., Criado, N., & Such, J. (2021). Discovering and Interpreting Biased Concepts in Online Communities. IEEE Transactions on Knowledge and Data Engineering. https://nms.kcl.ac.uk/hasp/pubs/ferrer2021discoveringtkde.pdf

- Bliuc, A. M., Smith, L. G. E., & Moynihan, T. (2020). “You wouldn’t celebrate September 11”: Testing online polarisation between opposing ideological camps on YouTube. Group Processes & Intergroup Relations, 23(6), 827-844. https://journals.sagepub.com/doi/pdf/10.1177/1368430220942567
- Brown, O., Lowery, C., & Smith, L. G. E. (2022). How opposing ideological groups use online interactions to justify and mobilise collective action. European Journal of Social Psychology, 52(7), 1082-1110. https://doi.org/https://doi.org/10.1002/ejsp.2886
- Crawford, B., Keen, F., & Suarez-Tangil, G. (2020). Memetic irony and the promotion of violence within chan cultures. Report https://kclpure.kcl.ac.uk/portal/en/publications/memetic-irony-and-the-promotion-of-violence-within-chan-cultures(51e9a948-2191-4ba9-8657-2c4491b2c0dd).html
- Crawford, B., Keen, F., & Suarez-Tangil, G. (2021). Memes, radicalisation, and the promotion of violence on chan sites. In Proceedings of the international AAAI conference on web and social media (Vol. 15, pp. 982-991). https://ojs.aaai.org/index.php/ICWSM/article/view/18121
- North, S., Piwek, L., & Joinson, A. (2021). Battle for Britain: Analyzing events as drivers of political tribalism in Twitter discussions of Brexit. Policy & Internet, 13(2), 185-208. https://onlinelibrary.wiley.com/doi/full/10.1002/poi3.247
- Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & Sirivianos, M. (2021). " How over is it?" Understanding the Incel Community on YouTube. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-25. https://dl.acm.org/doi/pdf/10.1145/3479556
- Simchon, A., Brady, W. J., & Van Bavel, J. J. (2022). Troll and divide: The language of online polarization. PNAS nexus, 1(1), https://research-information.bris.ac.uk/ws/portalfiles/portal/320764306/Full_text_PDF_final_published_version_.pdf
- Smith, L. G. E., Blackwood, L. & Thomas, E. F. (2020). The Need to Refocus on the Group as the Site of Radicalization. Perspectives on Psychological Science 15, 327-352, doi:10.1177/1745691619885870.
- Smith, L. G. E., Gavin, J., & Sharp, E. (2015). Social identity formation during the emergence of the Occupy movement. European Journal of Social Psychology, 45(7), 818-832. https://doi.org/10.1002/ejsp.2150
- Smith, L. G. E., McGarty, C. & Thomas, E. F. (2018). After Aylan Kurdi: How Tweeting About Death, Threat, and Harm Predict Increased Expressions of Solidarity with Refugees Over Time. Psychological Science 29, 623-634, doi:10.1177/0956797617741107.
- Smith, L. G. E., Piwek, L., Hinds, J., Brown, O. & Joinson, A. (in press). Digital traces of offline mobilization. Journal of Personality and Social Psychology, doi:10.1037/pspa0000338.
- Smith, L. G. E., Wakeford, L., Cribbin, T. F., Barnett, J. & Hou, W. K. (2020). Detecting psychological change through mobilizing interactions and changes in extremist linguistic style. Computers in Human Behavior, 108, 106298, doi:https://doi.org/10.1016/j.chb.2020.106298.

Censorship and self-censorship

- Marder, B., Joinson, A., Shankar, A., & Houghton, D. (2016). The extended 'chilling' effect of Facebook: The cold reality of ubiquitous social networking. Computers in Human Behavior, 60, 582-592. https://www.sciencedirect.com/science/article/pii/S0747563216301601
- Warner, M., & Wang, V. (2019). Self-censorship in social networking sites (SNSs)–privacy concerns, privacy awareness, perceived vulnerability and information management. Journal of Information, Communication and Ethics in Society. https://www.emerald.com/insight/content/doi/10.1108/JICES-07-2018-0060/full/html
- Stsiampkouskaya, K., Joinson, A., Piwek, L., & Ahlbom, C. P. (2021). Emotional responses to likes and comments regulate posting frequency and content change behaviour on social media: An experimental study and mediation model. Computers in Human Behavior, 124, https://doi.org/10.1016/j.chb.2021.106940
- Stsiampkouskaya, K., Joinson, A., Piwek, L., & Stevens, L. (2021). Imagined audiences, emotions, and feedback expectations in social media photo sharing. Social Media+ Society, 7(3), https://journals.sagepub.com/doi/full/10.1177/20563051211035692
- Stsiampkouskaya, K., Joinson, A., & Piwek, L. (2023). To Like or Not to Like? An Experimental Study on Relational Closeness, Social Grooming, Reciprocity, and Emotions in Social Media Liking. Journal of Computer-Mediated Communication, 28(2), https://academic.oup.com/jcmc/article/28/2/zmac036/6987873
- Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., & Stringhini, G. (2019). " You Know What to Do" Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-21. https://dl.acm.org/doi/abs/10.1145/3359309

Social media privacy & security

- Anthonysamy, P., Greenwood, P., & Rashid, A. (2014). A method for analysing traceability between privacy policies and privacy controls of online social networks. In *Privacy Technologies and Policy: First Annual Privacy Forum, APF 2012, Limassol, Cyprus, October 10-11, 2012, Revised Selected Papers 1* (pp. 187-202). Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-54069-1_12
- Hinds, J., Williams, E. J., & Joinson, A. N. (2020). "It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal. International Journal of Human-Computer Studies, 143, https://www.sciencedirect.com/science/article/pii/S1071581920301002
- Kökciyan, N., & Yolum, P. (2016). PriGuard: A semantic approach to detect privacy violations in online social networks. *IEEE Transactions on Knowledge and Data Engineering*, *28*(10), 2724-2737.
- Peersman, C. (2018). Detecting deceptive behaviour in the wild: text mining for online child protection in the presence of noisy and adversarial social media communications. Lancaster University (Thesis).

- Such, J. M., Porter, J., Preibusch, S., & Joinson, A. (2017). Photo privacy conflicts in social media: A large-scale empirical study. In Proceedings of the 2017 CHI conference on human factors in computing systems (pp. 3821-3832). https://dl.acm.org/doi/10.1145/3025453.3025668

Other references

- Abelson, H., Anderson, R., Bellovin, S. M., Benaloh, J., Blaze, M., Callas, J., ... & Troncoso, C. (2021). Bugs in our pockets: The risks of client-side scanning. *arXiv preprint https://arxiv.org/pdf/2110.07450.pdf*
- Agarwal, P., Raman, A., Ibosiola, D., Sastry, N., Tyson, G., & Garimella, K. (2022). Jettisoning Junk Messaging in the Era of End-to-End Encryption: A Case Study of WhatsApp. In *Proceedings of the ACM Web Conference 2022* (pp. 2582-2591).
- Bin Zia, H., He, J., Raman, A., Castro, I., Sastry, N. and Tyson, G., 2023. Flocking to mastodon: Tracking the great twitter migration. *arXiv preprint https://arxiv.org/abs/2302.14294*
- Brown, O., Smith, L. G. E., Davidson, B. I., & Ellis, D. A. (2022). The problem with the internet: An affordance-based approach for psychological research on networked technologies. Acta Psychologica, 228, 103650. https://doi.org/https://doi.org/10.1016/j.actpsy.2022.103650
- Chowdhury, P. D., Hallett, J., Patnaik, N., Tahaei, M., & Rashid, A. (2021). Developers are neither enemies nor users: they are collaborators. In 2021 IEEE Secure Development Conference (SecDev) (pp. 47-55). IEEE.
- Domínguez Hernández, A. (2023) "Self-Updating Prophecies: An Inquiry into Imagining and Building Decentralized Sensor Networks." *Science, Technology, & Human Values*, March. SAGE Publications Inc, 01622439231160466. doi:10.1177/01622439231160466.
- Peersman, C., Llanos, J. T., May-Chahal, C., McConville, R., Chowdhury, P. D., & De Cristofaro, E. (2023). Towards a Framework for Evaluating CSAM Prevention and Detection Tools in the Context of End-to-end-encryption Environments: a Case Study.
- Tahaei, M., Abu-Salma, R. and Rashid, A. (2023) "Stuck in the Permissions with You: Developer & End-User Perspectives on App Permissions & Their Privacy Ramifications." Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), Hamburg, Germany