Academic Access to Social Media Data for the Study of Political Online Safety

Andreu Casas, Georgia Dagher, and Ben O'Loughlin

A report from the *New Political Communication Unit*
Royal Holloway University of London.

newpolcom

ROYAL HOLLOWAY UNIVERSITY OF LONDON

Table of Contents

1. Introduction

On 26th October 2023 the Online Safety Act was enacted into law, with the goal of making internet services safer in the United Kingdom. The legislation puts forward a set of duties for internet service providers, such as social media platforms, and attributes new powers to the Office of Communication (Ofcom) to regulate online safety matters.

One key aspect of the new legislation is researchers' access to information. Independent academic research plays a key role in identifying, understanding, and addressing online harms — and in keeping the companies controlling our online environment in check. However, internet service providers, and social media companies in particular, are making it increasingly hard for academics to study online safety. For example, X (formerly Twitter) dismantled their academic API (application programming interface) in February 2023; and Meta discontinued Crowdtangle in August 2024, a tool used by many independent researchers and journalists to study the platform.

The goal of this report is to provide an overview of the kinds of data academics need in order to conduct independent research into political online safety matters on social media platforms, and the challenges we currently face. Additionally, we put forward ideas regarding novel governance structures that would enable high-quality independent research, while protecting users' rights and data privacy, in the United Kingdom.

This report is particularly timely. The political climate is increasingly polarised, mis/disinformation is becoming more sophisticated, and foreign information operations are on the rise — particularly during elections. In addition, there now is a window of opportunity for improving academic access to, and use of, social media data in the United Kingdom. In the context of the Online Safety Act, Ofcom will be assessing, by July 2025, academic needs and challenges for accessing and analysing platform data in the United Kingdom. Parliament is also working on other legislation related to the study of online harms, such as the Data (Use and Access) Bill and the Data Protection and Digital Information Bill.

This report is divided into five sections. First, we outline the key elements and threats to political online safety which motivate this report: mis/disinformation, toxicity, polarisation, extremism, biases in artificial intelligence tools, information operations, and political micro-targeting. We present three examples to illustrate the real-world impact of these elements. Second, we provide an overview of the types of data researchers need in order to study political online safety. Third, we describe the current state of data access for researchers on the main social media platforms. Fourth, we elaborate on the key values that should govern social media research. Finally, we argue for the establishment of an independent body in order to govern social media research.

1.1. Executive summary

In this report, we discuss the data needs of researchers studying political online safety, the challenges they currently face when accessing and analysing data from social media platforms, the values we believe should govern research on political online safety matters,

and we advocate for an independent intermediary body to manage and oversee academic access to (and analysis of) platform data moving forward.

Our analyses touches on the following 6 key points:

1. While there are clear political benefits to social media communication, such as allowing all voices to contribute to public discussions, there are also **potential political harms**. In particular, in this report we focus on the following: mis/disinformation, toxicity, echo chambers and polarisation, extremism and radicalisation, artificial intelligence (and its biases), information operations and election interference, and political micro-targeting. These can have negative attitudinal and behavioural consequences, such as undermining trust in democratic institutions, inciting (physical) violence, pushing some views in detriment of others, and manipulating public opinion.

2. Despite much high-quality research on these topics, we identify many **research gaps** to address. Some of these (at least partially) unanswered questions are: who believes and contributes to the spread of mis/disinformation? How can we detect increasingly realistic forms of misinformation? What is the role of platform algorithms versus real users in spreading misinformation, toxic content, pushing users into filter bubbles, and driving polarisation? How does exposure cross-cutting information online, or the lack thereof, shape political attitudes and behaviour? What are the key cross-platform differences in the moderation of extreme content, how effective are they, and can these have the unintended effect of radicalising individuals further? How can we identify and amend biases in platforms' recommendation and content moderation algorithms? What are the long-term and indirect effects of influence operations? What are the determinants and effects of political micro-targeting, and to what extent are online political ads persuasive?

3. In order to address these and many other pressing questions, researchers studying political online safety need access to the following types data from social media platforms: **account**-level information from elites, public and private accounts from ordinary citizens, and organisations; information about the **networks** in which these accounts are embedded; information about the content to which these accounts are **exposed**, including (political) advertising; information about the content **created** by these accounts; information about the content with which these accounts **engage**; and information about **platform governance** and **interventions**, such as content deletion and account suspensions. Additionally, researchers need access to **historical** and **live** data, to access and combine data from **multiple platforms**, and to combine platform data with other **external data** and **analytical models/tools** relevant to the research.

4. A detailed analysis of academic access to data from the five largest social media platforms in the United Kingdom (**Facebook**, **Youtube**, **Instagram**, **X**, and **TikTok**) reveals that it is **mostly unsatisfactory**. The most restrictive platform is X, which currently only allows for free data access to EU-based academics studying the threats described in Article 40 of the European Union Digital Service Act. The four other platforms provide some kind of access to platform data for researchers based in the United Kingdom. However, such access comes with many limitations. As a few examples, researchers have complained about: an overall lack of data documentation, the slow vetting process governed by the platforms

themselves (potentially threatening urgent research and research independence more generally), inconsistencies between the data shown on the platform and the data shared with researchers, difficulties using external data and analytical models to analyse platform data, and a lack of access to some crucial data, such as account and content information from private accounts, information about content to which accounts have been exposed, and information about platform governance, algorithms, and interventions.

5. The following key values must be at the core of political online safety research in order to strengthen its credibility, validity, and ethical standards: **trust**, **equality**, **transparency**, **reproducibility** and **replicability**, data **privacy** and **security**, and **consent**. Further details about each of them are provided in Section 6 of this report.

6. Finally, we argue for the establishment of an **independent intermediary body** to manage and oversee academic access and analysis of platform data. This body would act as an intermediary between platforms and researchers, facilitate data access in an ethical, equitable, and transparent manner, while also ensuring compliance with existing regulations. In particular, we argue that this body should undertake the following six functions. First, it should be in charge of the **vetting process**, developing a standardised set of rules and reviewing applications. Second, it should oversee and ensure the **quality of data** provided by platforms, by assessing the documentation and conducting data audits. Third, it should set up the necessary **infrastructure** for researchers to conduct their analyses. Fourth, it should **mediate disputes** that may arise between researchers and platforms. Fifth, it should provide **logistical and technical support** to platform researchers. Finally, the body should promote **stakeholder and public engagement** on political online safety matters.

2. Who we are. The *New Political Communication Unit* at Royal Holloway, University of London

The New Political Communication Unit is a research centre at Royal Holloway, University of London, launched in 2007. Our research agenda has three distinct but related foci:

- Digital communication technologies and the varied digital infrastructures across the world, and the assorted media forms and practices that make up those landscapes that define the contemporary era.
- New political behaviour, institutions, and policy challenges that shape and are shaped by the rapidly changing information and communication environment.
- New theoretical dilemmas, methodological techniques, and normative concerns that arise from the need to effectively research these rising phenomena.

From debates about the Internet's impact on citizen activism, political parties, and election campaigns, to concerns over international security, privacy, and surveillance; from the rise of blogging and social media as threats to traditional models of journalism, to controversies at the international level over whether, and if so how, internet media should be regulated and controlled; from the regulatory concerns created by powerful new media corporations, to massive programs of organisational reform taking place in the name of “big data”, what we term new political communication is continually in the headlines. There has been a steady stream of high-quality research in this field over the last decade.

The New Political Communication Unit is an international leader in this sphere, bringing together scholars with an interest in understanding the evolution of the information and communication environments which shape and are shaped by politics and policy-making. This combined focus — on theory, institutions, and practices, on the most interesting policy problems, and on multiple levels of analysis — will provide comprehensive coverage of some of the most important developments of our age.

3. **Motivation.** What is political online safety and why does it matter?

There are many relevant aspects to online safety. In this report we focus on improved data access for independent researchers studying online harms of a political nature: political online safety. Today social media plays a key role in politics. An increasing number of people rely on social media for consuming news, learning about, and engaging in politics. On the one hand, there are clear political benefits to the emergence of social media platforms. For example, it is easier for minorities and marginalised groups to have a voice in public debates. However, there are also clear political harms. As a few examples, people are more easily exposed to extremist views and to increasingly realistic fake news.

3.1. Political online safety

Political online safety refers to protecting people's political attitudes and behaviour, and political institutions and processes more generally, from online threats — including taking proactive actions that make the citizenry more capable of combating these threats, such as improved digital literacy. Some of the most pressing threats include, among others, misinformation, foreign operations, and hateful content. These have direct real-world political consequences, ranging from a loss of trust in political authorities, polarisation, extremism, physical violence, and democratic decay. In this section we provide an overview of some of these key threats, providing definitions, a description of existing academic findings, and a discussion of relevant open questions that future research must address.

3.1.1. Mis/disinformation

Definition. Misinformation refers to the dissemination of false information that can mislead the public (Jackson 2017; Sanders and Jones 2018; United Kingdom Digital, Culture, Media and Sport Committee 2019), and disinformation is false information that is deliberately propagated (Born and Edgington 2017; Tucker et al. 2018).

Consequences. Misinformation can have negative attitudinal and behavioural consequences, threatening democratic citizenry and institutions, and public safety more generally (Hendrickson and Galston 2019). For example, in the United Kingdom, high levels of online misinformation preceded recent violent mobilisations, such as protests against COVID-19 measures in 2021 and 2022 (Ahmed and Bales 2021; Wang et al. 2022; Wischerath et al. 2024) and violent far-right riots in the summer of 2024 (Institute for Strategic Dialogue 2024; Rowe and Mason 2024).

Prevalence. Empirical research shows that although only a small number of social media users are usually responsible for the creation of misinformation (Grinberg et al. 2019; Guess et al. 2019), misinformation spreads faster than other kinds of content (Vosoughi et al. 2018) and large numbers of users report having been exposed to some kind of misinformation (Lee et al. 2023). In the United Kingdom, Chadwick et al. (2018) found 9% of Twitter users to have deliberately shared disinformation, and another survey found 15% of respondents to have also deliberately engaged in the dissemination of false information on social media

(Deltapoll 2021). Additionally, in a recent survey, about 40% of adults in the United Kingdom reported having encountered misinformation online (Ofcom 2024).

Predictors. More ideologically extreme, particularly conservative and older individuals, are more likely to be exposed to and share misinformation (Grinberg et al. 2019; Guess et al. 2019). Other relevant predictors of misinformation exposure and sharing include gender (male), low media literacy, prior exposure to misinformation, distrust in the media, and a need for entertainment and socialisation (Sun and Xie 2024).

Effects. Regarding effects, research finds that exposure to misinformation leads to lower trust in the media and democratic institutions, and increased polarisation (Ognyanova et al. 2020; Azzimonti and Fernandes 2023).

Research gaps. Future research should look into the generalisability of findings, which are mostly based on data from the United States, including ideological asymmetries in believing and spreading false information (on the United States, see Grinberg et al. 2019; Guess et al. 2019; González-Bailón et al. 2023; Rossetti and Zaman 2023), determinants of public support for combating online misinformation and the role of governments, social media platforms, and individual users in this process (Jang et al. 2023), and the effectiveness (and potential unintentional effects) of different fact-checking and content moderation interventions (Pennycook et al. 2018; Tan 2022; Allen et al. 2024). Additionally, with the emergence of generative artificial intelligence, future research will need to adapt to studying misinformation that is increasingly realistic and sophisticated. For example, academics will need to develop better detection methods, design and test new digital literacy interventions for end users, and theorise about new potential implications.

3.1.2. Toxicity

Definition. Toxicity online includes incivility, intolerance, and violent threats. In its most extreme form, it can be characterised as hate speech, which refers to acts that advocate, incite, or justify discrimination and violence against a specific group, for example on the basis of race, religion, gender, or sexual orientation (United Nations Strategy and Plan of Action on Hate Speech 2019; Hietanen and Eddebo 2023).

Consequences. Toxicity can exacerbate harmful attitudes such as racism, misogyny, homophobia, and Islamophobia, which may incite physical violence over time. Empirical studies suggest that exposure to and sharing of hate and toxic speech on social media was in part to blame for the Finsbury Park Mosque terrorist attack (2017), for a spike in racially and religiously motivated hate crimes in the United Kingdom (such as in London in 2013-2014; Oxford in 2019; see Williams et al. 2020), and for the mainstreaming of sexist ideologies among the youth (Regehr et al. 2024). Toxic content also contributes to polarisation (Shcherbakova and Nikiforchuk 2022; Vasconcellos et al. 2023), and to the silencing of marginalised groups (Amnesty International 2017; Nadim and Fladmoe 2021; Koch et al. 2024).

Prevalence. Research finds that politicians are often the target of toxic speech. For example, between March and April 2022, politicians in the United Kingdom received more than 3,000 offensive tweets per day (Lynch et al. 2022). There seems to be an upper trend:

4.5% of the replies to candidates on Twitter during the 2019 election contained toxic language, compared to 3.3% during the 2017 election period (Gorell et al. 2020). On average, male candidates receive more political abuse in general, while women candidates receive more sexist comments and ethnic minority candidates receive more racist ones (Gorell et al. 2020). In the United States, there is evidence that 18% of tweets mentioning members of Congress in 2017-2018 contained uncivil language (Theocharis et al. 2020) and 1% of tweets mentioning Donald Trump and Hillary Clinton in 2016 contained extreme hate speech (Siegel et al. 2021). Regular users, not only politicians, are also the target of toxic speech. A 2021 survey reveals that in the United States, 20% of adults had experienced online harassment for different reasons. Men, white adults, and Republicans are more likely to report being harassed for their political views, women for their gender, ethnic minorities for their race, and LGBTQ individuals for their sexual orientation and gender (Vogels 2021).

Predictors. Platform-level predictors for the spread of toxicity online include platform affordances, such as the level of user anonymity (Barlett 2015; Zimmerman and Ybarra 2016; Moore et al. 2021). User-level predictors for engaging in the creation and dissemination of toxic content include higher tolerance for negativity and lower sensitivity to toxic content (Pradel et al. 2024; Pradel and Theocharis 2024), higher levels of polarisation (Saveski et al. 2021), and prior exposure to toxic comments online (Kim et al. 2021).

Effects. Over time, online hate and toxicity play a role in inciting real-world violence such as terrorist attacks, hate crimes, and harassment (Williams et al. 2020), and may have a silencing effect that limits targeted groups' freedom of speech online and leads them to exit online discourse (Nadim and Fladmoe 2021; Pradel and Theocharis 2024). Online toxicity directed at women politicians can negatively affect their decisions to re-run for elections (Gorell et al. 2020), perpetuating gender inequalities in political representation.

Research gaps. Future research should examine the nuances of toxic behaviour, as this comes in different formats, such as text, images, and cartoons (presented as entertainment; see Regehr et al. 2024). Additionally, we need further cross-platform research for a full picture of how toxic speech evolves and spreads across the social media ecosystem, and for a better understanding of the correlates between toxic behaviour and platform affordances (Munn 2020) and the extent to which platform policies or algorithms, versus human behaviour, are to blame for the spread of toxicity online (Munger and Philips 2019; Ledwich and Zaitsev 2020; Hokka 2021).

3.1.3. Echo chambers, filter bubbles, and (affective) polarisation

Definition. Echo chambers refer to people mostly following accounts and content on social media that reflect their own political views (Barberá 2020), and filter bubbles refer to the role of algorithms in sorting users into ideologically congruent networks (Barberá 2020). Ideological polarisation refers to people holding increasingly divergent political views, and affective polarisation refers to people holding increasingly negative views towards members of a political out-group (Kubin and von Sikorski 2021).

Consequences. A diversity of political views is key to democratic politics, yet high levels of ideological polarisation can lead to suboptimal outcomes such as policy gridlock (Jones 2001) and dissatisfaction with democracy (Wagner 2001). Additionally, affective polarisation

can contribute to undervalue and attack out-group members and counter-attitudinal views (Settle 2018), to a perception that the political environment is more polarised than it actually is, and to increased toxicity (Pascual-Ferrá et al. 2021; Hanscom et al. 2024). Social media echo chambers can potentially enhance ideological and affective polarisation (Sunstein 2017).

Prevalence. Findings based on the United States point to most social media users, particularly the politically-interested, being embedded in networks of like-minded accounts (Barberá et al. 2015, Barberá 2020; Nyhan et al. 2023; Wojcieszak et al. 2022). However, echo chambers seem to be weaker in European countries (Vaccari and Valeriani 2021). Additionally, cross-cutting interactions are not as rare as some public commentators argue (Sunstein 2017). For example, Barberá et al. (2015) find that for about 75% of Twitter users from Germany, Spain, and the United States, at least 25% of their following are accounts with a different ideology, and Eady et al. (2019) find a substantial overlap between the media accounts followed by liberals and conservatives on social media in the United States. Regarding polarisation, despite numerous evidence pointing to increased ideological and affective polarisation in the United States (Iyengar et al. 2019), these patterns do not always apply to other Western democracies (Boxell et al. 2024). In the United Kingdom, ideological polarisation has remained mostly stable (Boxell et al. 2024), with a few exceptions such as increasingly divergent views on immigration (Tipoe and Lee 2024). Affective polarisation, however, is slightly on the rise (Garzia et al. 2023), with group identities and public debates around Brexit playing a key role in this trend (Hobolt et al. 2020). Finally, polarising content such as anti-immigration posts, although mostly generated by a small number of users, travel faster than neutral content, in part thanks to these accounts being embedded in like-minded networks that contribute to the dissemination of these posts (Nasuto and Rowe 2024).

Predictors. Echo chambers on social media are mostly the result of users deciding to follow like-minded accounts and content. In particular, existing research points to age (older people) and ideology (liberals and conservatives, versus independents and moderates) as key predictors of slanted online media diets (Guess 2021). Yet, platform algorithms exacerbate these further, a phenomenon known as filter bubbles (Barberá 2020). For example, Bakshy et al. (2015) and González-Bailón et al. (2023) show that, in two different time periods, Facebook algorithmic ranking reduced exposure to cross-cutting content by about 15%. There are mix-findings regarding whether online echo chambers are predictive of increased polarisation. Boxell et al. (2017) find higher levels of polarisation among older generations who are the least likely to use social media, and in a deactivation experiment on Facebook, Nyhan et al. (2023) show no clear relationship between exposure to like-minded news sources on social media and ideological and affective polarisation.

Effects. Ideological polarisation can lead to legislative gridlock (Jones 2001) and increased dissatisfaction with democracy (Wagner 2001; Torcal and Magalhães 2022). Affective polarisation can lead to a more toxic online environment (Saveski et al. 2021), especially over time (Nelimarkka et al. 2018). Additionally, some findings link affective polarisation to increased support for political violence (Kalmoe and Mason 2022), although other work argues that this relationship has been overstated (Westwood et al. 2022).

Research gaps. Future research should investigate the role of platform algorithms, versus users, in driving polarisation — particularly given the rise of platforms such as TikTok, where recommendations play a much more crucial role in the curation of content. Further research should compare and explore the role of online versus offline networks: people may be embedded in like-minded networks online, but how different are these to their offline networks? Additionally, further research is needed to understand how exposure to cross-cutting information online, or the lack thereof, shapes political attitudes and behaviour.

3.1.4. Extremism and radicalisation

Definition. Extremism refers to holding extreme political views (such as far-left or far-right), and radicalisation to the process of one's views becoming more extreme.

Consequences. Extreme views expressed online can incite hate and violent behaviour, both online and offline. For example, far-right leaders across the world including the United Kingdom have used social media to spread extreme and often conspiratorial narratives, such as the rhetoric of invasion, to incite hatred against minorities and immigrants (Hope Not Hate 2019; Williams et al. 2020). Extreme views expressed on social media can inspire hate crimes and terrorist attacks. For example, the perpetrator of the 2019 Christchurch, New Zealand, terror attack posted a video ahead of the massacre where he asked viewers to "subscribe to PewDiePie", a social media star who has become known for his anti-Semitic comments and endorsements of white supremacist conspiracies (Chokshi 2019). A few days later, Swastikas with the words "sub 2 PewDiePie" were graffitied on a school wall in Oxford, United Kingdom (Williams et al. 2020). The terrorist also referenced Darren Osborne, who committed the Finsbury Park Mosque attack in 2017 and is known to have been influenced by social media communications ahead of the attack (Williams et al. 2020).

Prevalence. Extreme views on major social media platforms, such as Facebook, Instagram, TikTok, and X, are rare. Barberá et al. (2015) and Wojcieszak et al. (2022) show that the ideological distribution of ordinary users on X is fairly moderate, following a normal distribution. Although Bond and Messing (2015) and Eady et al. (2024) find ordinary users to follow a bimodal ideological distribution on Facebook and X, respectively, they also find extreme ideologists to be rare. In a dataset of tweets mentioning Donald Trump and Hillary Clinton during the 2016 United States election, Siegel et al. (2021) find white supremacist language in less than 0.02% of tweets. In a study of convicted terrorists in the United Kingdom, Gill et al. (2017) find white supremacists — particularly extreme right-wing terrorists — to be significantly more likely to learn and communicate online, compared to other users. Finally, extreme views seem more prevalent in niche platforms, such as Gab (Zannettou et al. 2018), 4chan (Colley and Moore 2022), Tumblr (Nagle 2017), Parler (Stevenson et al. 2023), Rumble, Odysee, and Telegram (Al-Rawi 2021) — although the popularity of these platforms is quite country-dependent (Juarez Miro and Toff 2023).

Predictors. Key predictors of online extremism and radicalisation mainly point to platform affordances, such as increased anonymity (Awan et al. 2019), softer content moderation policies (Zennettou et al. 2018; Colley and Moore 2022), and algorithmic biases that push users into rabbit holes (Tufekci 2018; Barnes 2022; Brown et al. 2022) — although the latter does not happen as frequently (Brown et al. 2022) as some public commentators argue (see Tufekci 2018; Roose 2019). Research also finds that users with extreme views are more

active and that their content spreads faster than that of moderate users (Wojcieszak 2010; Siegel et al. 2021; Wojcieszak et al. 2022; Eady et al. 2024).

Effects. Exposure to extreme views online can lead to ideological radicalisation (Koehler 2014; Magdy et al. 2016). Regarding heterogeneous effects, research points to younger citizens and first-time voters being particularly susceptible to radicalisation (Karl 2017). Online radicalisation can lead to violent offline behaviour. For example, Pauwels and Schils (2016) find higher self-reported political violence among young citizens that consume extreme content online.

Research gaps. Most social media research focuses on major social media platforms, while extreme and radicalising content often originates in smaller niche platforms. Future research should look into the wider social media ecosystem (Bovet and Grindrod 2022), paying closer attention to smaller platforms and mapping how content travels from those platforms to the mainstream ones (Buntain et al. 2021). Further research is also needed on the radicalising effects of social media, particularly given that these are likely to be cumulative and emerge over the long term. Additionally, further research should look into the effectiveness of platform content moderation policies: are there cross-platform differences in the moderation of extreme and radicalising content? What interventions are most effective at reducing the spread and prevalence of extreme views? Do extreme users radicalise further when expelled from mainstream platforms, by being pushed into niche platforms where exposure to extreme content is more present? Is extremist content more likely to be created and circulated on encrypted communication platforms, and if so, how can researchers and policy-makers access encrypted data to understand it and infer its possible effects on users?

3.1.5. Artificial intelligence (biases)

Definition. Artificial intelligence (AI) refers to automatic systems trained to perform a given task. AI bias refers to these automatic systems systematically underperforming in a way that results in unfair outcomes for particular groups of users (Ferrara 2023).

Consequences. Social media platforms increasingly rely on black-box AI tools to automate many tasks, such as moderating and recommending content (Gillespie 2020; Gorwa et al. 2020; Brown et al. 2022). AI-based systems can generate biased outcomes, mostly as a result of being trained on biased data (Ferrara 2023). In turn, social media platforms may unfairly (and inadvertently or intentionally) moderate and recommend particular political and ideological voices to the detriment of others — thus shaping relevant political conversations online.

Prevalence. Regarding content moderation, research identifies several AI tools as having a racial dialect bias (Davidson et al. 2019; Sap et al. 2019; Ball-Burack et al. 2021) and as silencing and censoring members of marginalised communities, including members of the LGBTQ community, women, and people of colour, even in cases where these individuals abide by platforms' rules (Haimson and Hoffman 2016; Cook 2019; Joseph 2019; Van Horne 2019; Are 2020; Haimson et al. 2021). AI tools designed to identify duplicates may be insensitive to the use of the same content in a different context (for example, terrorist propaganda being reposted in a journalistic context; Llansó 2019), and those designed to detect extremist content have the potential to eliminate crucial evidence of war crimes

(Kayyali and Althaibani 2017). Regarding recommendation algorithms, research finds platform algorithms to boost, at least to some extent, extreme content, misinformation, hate speech, radical ideologies, and conspiracy theories (Sunstein 2017; Tufekci 2018; Munn 2020; Ahmed and Bales 2021; Barnes 2022; Brown et al. 2022; Wang et al. 2022; Nasuto and Rowe 2024; Regehr et al. 2024; Wischerath et al. 2024). However, other research finds platform algorithms to actually reduce exposure to untrustworthy content (Guess et al. 2023b).

Determinants. Content from minorities and people with non-Western backgrounds is more likely to be unfairly moderated by AI systems (Haimson and Hoffman 2016; Cook 2019; Davidson et al. 2019; Joseph 2019; Sap et al. 2019; Van Horne 2019; Are 2020; Ball-Burack et al. 2021; Haimson et al. 2021; Casas 2024). Toxic content — such as extreme views, hate speech, as well as misinformation — is sometimes recommended at higher rates, compared to other kinds of content (Sunstein 2017; Tufekci 2018; Munn 2020; Ahmed and Bales 2021; Barnes 2022; Brown et al. 2022; Wang et al. 2022; Nasuto and Rowe 2024; Regehr et al. 2024; Wischerath et al. 2024).

Effects. The unfair moderation of content from minorities and users of non-Western culture and/or origin can lead to biased political conversations online (Van Horne 2019; Casas 2024; Webb-Williams et al. 2024). Biases in the recommendation of extreme views, misinformation, and harmful content at higher rates can lead to an increased toxic online environment, radicalisation, and violence (Magdy et al. 2016; Pauwels and Schills 2016).

Research gaps. There is a need for more audit-type research looking into potential social media biases in the moderation (Casas 2024; Mosleh et al. 2024) and recommendation of content (Brown et al. 2022), as well as other tasks performed by AI tools, particularly given the rise of generative AI in recent years. More research is also needed in developing and testing interventions for unbiasing biased systems and assessing how AI biases shape politically relevant conversations.

3.1.6. Information operations and election interference

Definition. Information operations, also known as influence operations, refer to states engaging in the collection and dissemination of information online to advance their (geopolitical) interests, by criticising an adversary and/or promoting their own narratives (Miskimmon et al. 2013; Golovchenko et al. 2020; Bergh 2024). Election interference is a type of information operation where the aim is to influence the outcome of an election, either by promoting/demoting a given candidate/party, polarising the electorate, or suppressing participation (Golovchenko et al. 2020; Bradshaw et al. 2021).

Consequences. Information operations and election interference aim to sway public opinion and voter behaviour. This undermines the democratic process, may facilitate the spread of false information, may increase polarisation, can lower public trust in the electoral process, and lead to accusations of electoral fraud.

Prevalence. While it is hard to quantify the prevalence of such operations, recent accounts point to more than 80 countries participating in some sort of information operation on social media platforms (Bradshaw et al. 2021), including Western democracies (Bradshaw et al.

2021; Earl et al. 2022; Casas 2024). A large body of academic literature documents information operations by the Russian Internet Research Agency, particularly in the context of the 2016 and 2020 United States Presidential elections (Lukito 2020; Tucker 2020), the 2016 Brexit referendum in the United Kingdom (Howard and Kollanyi 2016; Booth et al. 2017), and national elections in France, the Netherlands, and Germany in 2017 (Brattberg and Maurer 2018; Adler and Thakur 2021). These analyses reveal how the Internet Research Agency used both human-controlled and automatic bot accounts in its information operations, and how it coordinated actions across several social media platforms, such as Facebook, Twitter, and Reddit (Lukito 2020). Other information operations of relevance involve coordinated social media accounts amplifying pro-Chinese messages from China's diplomats in the United Kingdom (Schliebs et al. 2021).

Predictors. Predictors of being the target of influence operations vary by context, and include demographic characteristics and partisanship. The types of targeted messages also vary depending on the goals of the campaign. Evidence from the 2016 United States election finds that messages targeted at Republican voters emphasised immigration, race, and ethnicity, and messages targeted at African American voters emphasised structural inequalities, with the goal of encouraging them to boycott the election (Howard et al. 2018; Freelon et al. 2022).

Effects. Empirical evidence so far suggests that the operations conducted by the Russian Internet Research Agency did not significantly impact the outcome of the elections mentioned above. This is because exposure to influence operations and election-related mis/disinformation tends to be heavily concentrated towards specific groups of voters (Eady et al. 2023), to be less prevalent than content from domestic and trustworthy sources (Tucker 2020; Eady et al. 2023), and the number of accounts linked to these operations is generally small (Booth et al. 2017). Additionally, with the rising threat of interference, certain states have implemented prevention measures (Brattberg and Maurer 2018; Bateman and Jackson 2024). However, it remains difficult to assess the long-term impact of these operations, and they may have indirect effects, such as the spread of conspiracy theories following the vote and public unrest, particularly if domestic authentic networks become involved. The attack on the United States Capitol on January 6, 2021, shows this.

Research gaps. While the direct threat of these operations in influencing voting behaviour seems limited when focusing on single elections (François and Douek 2021; Eady et al. 2023), further research should investigate the longer-term effects of influence operations, as the strategies, tactics, and aims of these operations vary (Martin et al. 2020; Martin et al. 2023) and extend way beyond single elections. Social media platforms have made great efforts to fight influence operations (for example, through the Disinfodex database) but assessing the impact of takedowns is difficult, as threat actors behave differently and may change their behaviour over time, meaning that they may quickly adapt to punitive actions taken against them. Takedowns are only (a small) part of the solution to combat influence operations, and may be receiving too much attention in the public policy discourse. Further research must also examine the role of recommendation algorithms in spreading election-related misinformation; while one study focusing on Facebook and Instagram concluded that the recommendation algorithm was not a driver of exposure to misinformation in the 2020 United States election (Guess et al. 2023a), these findings have been disputed (Bagchi et al. 2024).

3.1.7. Political micro-targeting

Definition. Political micro-targeting refers to the collection of online behavioural data from individuals in order to deliver them tailored political advertising (Dobber et al. 2019). This phenomenon became widely known in 2016, when Cambridge Analytica misused data from millions of Facebook users in the United States to run tailored ads for Donald Trump's presidential campaign.

Consequences. Political micro-targeting can have societal and democratic benefits, such as increased political participation, providing voters with information about the issues they care about, and offering more cost-effective advertising for candidates (Zuiderveen Borgesius et al. 2018). However, political micro-targeting can also threaten democratic politics by invading people's (data) privacy, manipulating public opinion, boosting campaign spending, and attributing private social media platforms a key intermediary power in elections (Zuiderveen Borgesius et al. 2018).

Prevalence. In a study of political ads on Facebook and Instagram across 95 countries and 113 elections, Votta et al. (2024) show that political ads on social media platforms are now common across the globe, and that most ads target users based on a single or two criteria. Geographic location and socio-demographic characteristics are the most common targeting criteria, followed by interests and behaviour online (Votta el al. 2024). In the 2024 United Kingdom General Election, political parties spent more than 2.5 million pounds on Google Ads (Bishop-Froggatt 2024), and an average of about 1 million pounds per week on Meta ads (Who Targets 2024). These ads targeted particular audiences, with candidates from the Conservative party, for example, targeting older voters on Meta ads (Plevin 2024).

Predictors. Online micro-targeting is more present in Western democracies and wealthier countries, where parties and candidates have more resources to spend on online ads (Votta et al. 2024). In democratic countries, parties and candidates in countries with a proportional (versus majoritarian) electoral system, with limits on traditional media campaign spending, and with stricter data protection laws, are more likely to run targeted social media ads (Votta et al. 2024). Regarding the particular targeting criteria, right-leaning parties are more likely to target older men, and left-leaning parties are more likely to target younger and women voters (Votta et al. 2024).

Effects. In a study of the 2021 Dutch election campaign, Chu et al. (2023) found a sample of 505 participants to be exposed to about 9,000 ads from political parties on Facebook. The ads had an effect on both their propensity to vote and voting choice. In a meta analysis of campaign experiments conducted by the Democratic party in the 2018 and 2020 elections in the United States, Hewitt et al. (2024) find a small but meaningful variation in the persuasive effects of online political ads, and Tappin et al. (2023) find micro-targeting to be most effective at shifting policy views when based on a single (rather than multiple) individual characteristic.

Research gaps. The vast majority of research on political micro-targeting is based on a few countries (such as the United States, Netherlands). Future research should explore the prevalence, determinants, and effects of micro-targeting in other countries. Additionally, there is still a lot to learn about the prevalence of micro-targeted campaigns across

platforms, from which party-ideology, and the relative ability of online ads to persuade voters based on different individual characteristics.

3.2. Some illustrative examples

Recent cases of public and political unrest and individual harm have been attributed to the spread of unverified or intentionally false information and toxicity online. Misinformation and hate speech often spread in tandem, and through a cumulative process, they progressively trigger extremist attitudes, potentially leading to public disorder and real-world violence. This section elaborates on some of the threats mentioned above, showing how they often interact to jointly undermine the political and individual safety of citizens.

3.2.1. The UK riots: the capitalisation of anti-immigration attitudes and spread of misinformation

In July 2024, three girls were murdered in a knife attack in Southport, England. False reports that the attacker was a Muslim asylum seeker who had entered the United Kingdom illegally quickly spread on social media platforms (Adams 2024; Fung 2024; Fox 2024). While this false information was initially shared only by a few individuals, within a few hours related posts became some of the most widely circulated across multiple social media platforms.

This triggered the mobilisation of far-right groups online and eventually led to the eruption of violent anti-immigration and Islamophobic riots across several cities in the United Kingdom. Even after authorities disclosed the identity of the assailant — a British national born in the United Kingdom — false claims about his name and origins continued to spread online (Rowe and Mason 2024).

Analyses of users' behaviour on social media highlight the role of several platforms in fuelling the violent riots. Soon after the attack, the false name had received thousands of mentions, reached trending status on X, posts from far-right activists inciting hatred towards Muslims received close to a million views, hashtags related to the riots and racism became among the most popular on TikTok, and a YouTube livestream of the riots received hundreds of thousands of views with live comments showing support for the violence (Institute for Strategic Dialogue 2024; Rowe and Mason 2024).

Social media platforms failed to take down a range of toxic content and misinformation related to the attacker and the events. On the contrary, evidence points to recommendation algorithms amplifying far-right voices, misinformation, Islamophobic content, and hate speech (Institute for Strategic Dialogue 2024). In particular, platforms with relaxed content moderation policies, such as Telegram, played a key role in the dissemination of undesired content (Rowe and Mason 2024).

3.2.2. Misinformation during the COVID-19 pandemic threatening public safety

Throughout the COVID-19 pandemic, some individuals continuously opposed health and safety measures introduced by governments, such as lockdowns, mask mandates, and vaccination certificates. As these measures were implemented, online conspiracy-oriented

communities mobilised around fabricated or distorted evidence related to the virus (Ahmed and Bales 2021), in some cases encouraging public protests.

In the early days of the pandemic, COVID-19 anti-vaccine messages were widespread in the United Kingdom and appeared to be amplified by social media, in particular targeted at young people who had a higher tendency to believe conspiracy theories about the vaccine (The Policy Institute 2020). COVID-19 related misinformation largely spread through society from the top down — from politicians, celebrities, and online influencers, particularly in the United States and the United Kingdom. A few months into the pandemic, only a dozen of individuals were found to be responsible for spreading the majority of COVID-19 related misinformation and anti-vaccine content on X and Facebook (Center for Countering Digital Hate 2021).

While most mainstream platforms eventually sought to implement stricter content moderation policies to fight misinformation, alternative platforms with lax content moderation policies, such as Telegram, grew in popularity. There is evidence that one of the largest conspiracy-oriented group chats in the United Kingdom on Telegram grew in size during and after the COVID-19 pandemic, and this increased interconnectivity coincided with more planning discussions for associated offline protests, which eventually occurred in response to lockdowns, vaccine rollouts, and governmental instability in the United Kingdom (Wischerath et al. 2024). While COVID-19 conspiracy theories may initially spread on mainstream platforms, as soon as these platforms implement measures to curb misinformation, individuals who believe them are quick to migrate to other platforms.

The implications of false information on the sources and measures to prevent a deadly virus are drastic, and groups that spread narratives to discredit legitimate public health measures may have contributed to preventable illness and death (Wang et al. 2022). Citizens who could have been protected were exposed to a virus that can have long term negative health effects (for example, see Office for National Statistics 2021).

3.2.3. The normalisation of misogyny

Two aspects of online misogyny which have dire political consequences are discussed here: the toxicity faced by women in politics and the role of social media in the mainstreaming of negative gender stereotypes.

Online toxicity towards women in politics includes not only threats and personal harassment, but also involves gender disinformation campaigns which build on gender stereotypes (Wilfore 2022; Di Meco 2023). Evidence from the United Kingdom shows that this has been on the rise, and that women politicians, compared to their male counterparts, are more likely to receive comments on social media that question their qualifications, their positions, or that are explicitly sexist (Greenwood et al. 2020; Ward and McLoughlin 2020; Southern and Harmer 2021; Esposito and Breeze 2022).

Apart from the personal harms this creates, direct political effects of facing such harassment include decisions to alter their online discourse, stand down, not re-run for office, or modify their campaigning activities in a way that hurts their electoral prospects (Collignon and Rüdig 2021). In addition, evidence from the United Kingdom shows that witnessing sexist

comments targeted at politicians on social media decreases the appeal of running for political office among women (Vrielink and van der Pas 2024). Toxicity targeted at women politicians therefore has broader, indirect political consequences, which threaten women's political participation and representation.

In parallel to the rise in toxicity targeted at women politicians and candidates, social media is enabling the spread of misogynistic ideologies among online communities. While social media gives a voice to human rights activists, it also provides a global platform for influencers with extremist ideologies. One example in the United Kingdom is Andrew Tate, a leader of the "manosphere" community, which promotes men's domination over women and regressive and violent gender stereotypes (Farrell et al. 2019; Bragg et al. 2022; Fazackerley 2023; Haslop et al. 2024; Pearson 2024).

This community is not fully isolated from the alt-right; many similar influencers embrace other hateful ideologies and conspiracy theories. Their followers then become at increased risk of falling into the extremist rabbit hole.

While Andrew Tate and similar influencers have been banned from mainstream platforms, they remain highly visible on them (Hall 2023), showing that social media platforms are failing to effectively contain the spread of extremist ideologies. Along with their fans at fear of being censored, they have moved to alternative platforms with more lenient content moderation policies, which have become a hub of alt-right groups. With time, their following may be lured into other alt-right communities such as conspiratorial or white-supremacist groups.

The most visible mobilising effect of these influencers has been in the United States, as shown during recent protests on university campuses (Del Rey 2024). In the United Kingdom, an increasing number of boys admire Andrew Tate, who is radicalising them into extreme misogyny (Evans 2023; Gillett 2024).

Apart from this online trend that started from the top-down, young boys are being radicalised through exposure to "softer" forms of misogynistic content. Misogynistic content online can go unnoticed as it comes in various formats, and is sometimes presented as entertainment that spreads with the help of recommendation algorithms. On TikTok, for example, a UK-based analysis of thousands of videos found a fourfold increase in the level of misogynistic content being recommended after only a week of usage (Regehr et al. 2024). The recommendation algorithm privileges more extreme material, and through increased usage, users are gradually exposed to more misogynistic ideologies which are presented as entertainment. As a result, these ideologies are normalised among young people and become embedded within mainstream youth cultures.

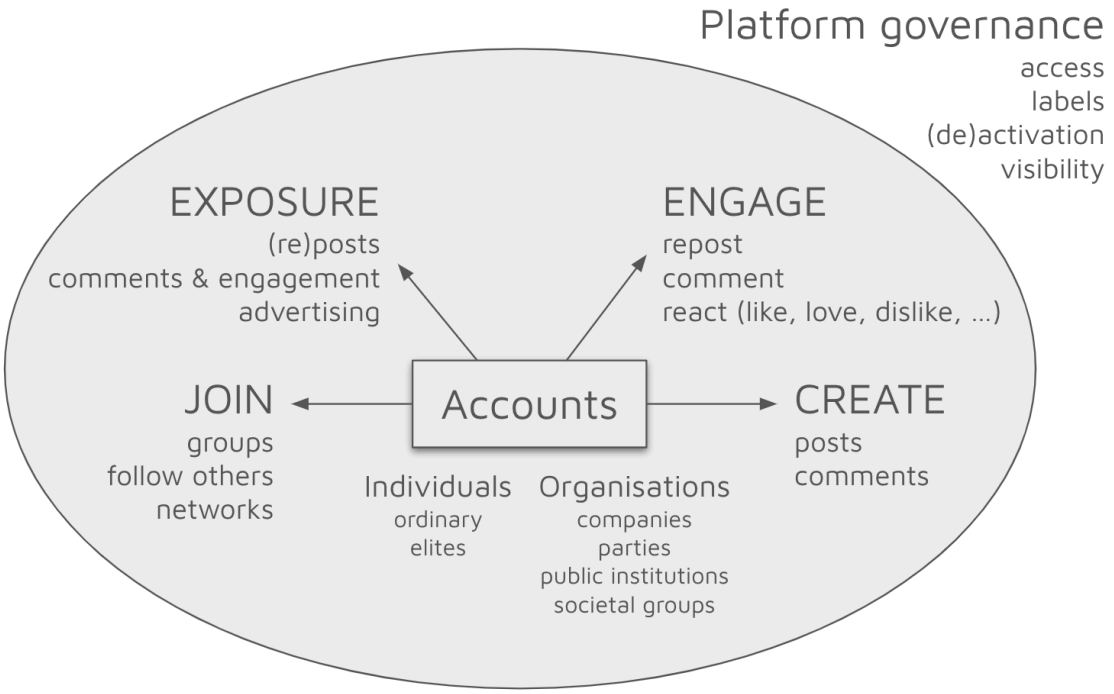
The implications of this are that women, as well as other marginalised groups, are at risk of facing increased discrimination offline, and long-term political risks include a backsliding on women's civil rights.

4. Data needs. What types of data do academics need to research political online safety?

In this section we provide an overview of the types of data that are crucial for conducting research on political online safety. Note that we do not necessarily mean for researchers to have direct unrestricted access to all of this data. In the last section of this report, we provide further discussion on how researchers can independently and transparently analyse social media data in a secure and safe manner.

Researchers need access to a variety of data in order to independently and accurately assess the many potential political harms that can emerge in the social media environment, as well as potential ways of addressing them. In *Figure 1* we provide a high-level overview of the different kinds of actors, actions, and interactions that are of interest. Next we elaborate on each of these data types and on specific data features related to them.

Figure 1. *Diagram of key social media actions and interactions.*



4.1. Accounts

At the core of social media communications are **accounts**. These are profiles that individuals and organisations have on social media platforms. The naming convention for accounts can vary depending on the platform: *accounts* on Instagram and Reddit; *profiles* on Facebook, X, and TikTok; and *channels* on YouTube.

Researchers studying political online safety are often interested in distinguishing between different kinds of individual and organisational accounts. For individual accounts, it is often relevant to distinguish between accounts from “ordinary” citizens and those from political

“elites” (Barberá et al. 2019; Eady et al. 2019) such as politicians and journalists. Some platforms also distinguish between “public” and “private” accounts (such as Facebook profiles versus public page or group) — where content of private accounts is only visible to other accounts vetted by the user. The distinction between, and access to all, these different kinds of accounts is relevant for a variety of political online safety research; for example, to address whether online polarisation is elite- or mass- driven, or to put forward a more accurate and useful picture of how (fake) news spread across the social media environment (McCarty et al. 2006; Fiorina and Abrams 2008; Eady et al. 2019).

Beyond access to the different types of social media accounts, access to the following account-level information is relevant for political online safety research.

- **Unique identifiers:** *the unique name and (alpha)numeric ID of the account*. Researchers need access to this information in order to select the data relevant for a given project (for example, based on a list of accounts pre-defined by the researcher), and to study the relationships between accounts (for example, who mentions or follows whom).

- **Description**: *self-reported description of an account*. This data is often crucial for identifying different kinds of accounts for analysis. For example, researchers can use a predefined set of keywords (such as “labour” or “conservative”) to identify politically engaged users, and/or supporters of different parties.

- **Tags**: *tags that define the type of account* (for example, verified blue check mark on X), or *type of content posted by the account* (for example, political, society, entertainment, etc. tags for YouTube channels) — either self-reported by the account holder, or given by the platform. This information is useful for selecting data for analysis (such as accounts posting about politics), and for identifying different types of accounts in larger datasets (such as exploring the role of entertainment-type accounts in the political ecosystem). These would also include any tags attributed to an account by a social media platform for marketing purposes.

- **Historical metadata**: such as *the creation date of the account and aggregate historical statistics* (such as the total number of posts/comments posted, number of followers, number of accounts followed). This information is particularly relevant to identify salient and influential accounts (posting more/less often, and with higher/lower reach). Additionally, previous research shows this information to also be particularly useful for identifying (bot and human) malicious accounts — for example, long-lasting accounts are less likely to be malicious.

- **Location**: *self-reported or platform-attributed location of an account.* This data is also important for selecting/sampling data for analysis. For example, this information can be relevant for researchers studying information operations and election interference.

4.2. Networks

Accounts **JOIN** and are embedded in networks, by (reciprocally) following other accounts, and/or joining groups, pages, and conversations. Political online safety researchers need

network-level information for a variety of purposes, such as to determine the ideological leaning of accounts (Barberá et al. 2019), to compare potential versus actual exposure to politically-relevant content, and to better understand the role of platform algorithms in information curation (Bashky et al. 2015; González-Bailón et al. 2023). Here is a list of relevant network-level data for political online safety research:

- **List of followers**: *list of accounts (unique identifiers) that follow a given account*. This information can be useful to study the kinds of users that follow untrustworthy media accounts (Guess et al. 2019). Additionally, timestamps for when each follower account started following a given account can be useful to determine the start and relevance of the relationship.

- **List of followees**: *list of accounts (unique identifiers) followed by a given account*. This information can be useful to identify politically-interested users, by for example pulling the list of followees of politicians and/or mainstream media accounts (Casas 2024). Additionally, timestamps for when a given account started following each of the followed accounts can be useful to determine the start and relevance of the relationship.

4.3. Posts

Accounts **CREATE** content by posting **public** messages visible to everyone, or **private** posts only visible to those vetted by the user. Many research projects may not need access to private posts, such as a study of whether candidates and parties contribute to the spread of misinformation during an election. However, other projects can enormously benefit from being able to study private posts, such as a study of the role of ordinary users and/or extreme ideologues, in the creation and spread of political misinformation. Additionally, political online safety researchers need access to (**historical**) content posted as far back as possible, to assess the presence of relevant patterns at and across different points in time.

- **Original and shared posts**: *the text, visual, and audio of original and shared posts*. Information about the content (re)posted by accounts is crucial to most political online safety research. Past research mainly focuses on text data from posts, yet social media is becoming increasingly audio-visual, which means that political online safety researchers need to be able to access and study at scale all the data modalities of social media posts. For example, recent research points to misinformation studies based on text data only (versus visual/multimodal data) substantially undercounting the amount of misinformation present on social media (Yang et al. 2023).

- **Expanded urls**: *full original (non-shortened) URL included in any (re)post*. Political online safety researchers often use information about URLs included in social media messages to identify links to (untrustworthy/fake) news sites (Guess et al. 2019; Grinber et al. 2019) and study the correlates of different kinds of media diets with political attitudes and behaviour (Guess 2021; Casas et al. 2023; Eady et al. 2024).

- **Account information for created/shared posts**: *unique identifier for accounts (re)posting a given post*. This data is crucial for mapping and studying which accounts contribute to the creation and dissemination of politically-relevant content.

- **Timestamp**: *day and time of the creation of original posts, and/or the time the original post was shared by an account of interest*. Temporal data is key to mapping and studying information diffusion for political online safety research.

4.4. Exposure

Accounts are **EXPOSED** to content generated by those in their network, recommended or ranked by the platforms, and/or by advertisers. This exposure is crucial to political online safety, as it can influence the political attitudes and behaviour of users online and offline. This data is also important to study the role of platform algorithms in the dissemination of information (for example, algorithms can up- or down-rank particular content and accounts), and the effects of political advertising and micro-targeting.

- **Post exposure**: ***number*** *and* ***list*** *of accounts exposed to a given (re)post*. In conjunction with **Network** information, this data can be used to assess the role of platform algorithms in content curation (Eady et al. 2019; Guess et al. 2023a; González-Bailón et al. 2023).

- **(Political) Ad exposure**: ***number*** *and* ***list*** *of accounts exposed to a given ad*. This information is crucial for studying the effects of (political) advertising and micro-targeting on people's political attitudes and behaviour. Political ads are particularly relevant, but so are those not labeled as such by the platform and/or advertiser, given that platforms sometimes do not label ads that are indeed political (Carolan 2024), and that non-political information (such as sports and gaming) can sometimes be linked to political attitudes.

4.5. Engagement

Accounts **ENGAGE** with content, by reacting (for example, liking), sharing, and commenting on content from other accounts. This data can help political online safety researchers identify the conditions under which toxic, extreme, and untrustworthy content reaches higher engagement levels, and whether platform algorithms boost ideologically congruent content and/or extreme views.

- **Post reactions**: ***number*** *of likes, dislikes, etc. of a (re)post*, *and* ***list*** *of accounts for each (re)post and engagement metric.* In conjunction with **Network** and **Post exposure** information, this data can also be used to assess the role of algorithms in curating and mediating content exposure and engagement (González-Bailón et al. 2023).

- **Reposts**: ***number*** *of reposts/shares of a given original post, and* ***list*** *of accounts that reposted the post.* This data is crucial to study the dissemination of politically-relevant information on social media.

- **Comments**: ***number*** *of comments on a given (re)post, textual-visual-audio* ***content*** *of each comment, and* ***account****-level information from commenters.* This information

is important for the study of information diffusion, echo chambers and polarisation, and toxic behaviour online, among other aspects of political online safety.

4.6. (Political) Advertising

Many are concerned about the potential harms of political micro-targeting (Zuiderveen Borgesius et al. 2018; Dobber et al. 2019). Political online safety researchers addressing these questions need information about which ads are being run on a given platform, when and by whom, and the user-level features targeted by the advertiser. Although political advertising is the main focus, information about other ads may be useful to disentangle whether platforms and/or advertisers are doing a good job at accurately labeling politically-relevant ads (Carolan 2024), and to study the conditions under which non-political ads (such as sports and gaming) can also influence users' political attitudes and behaviour (Wojcieszak et al. 2024)

- **Ad content**: *text, visual, and audio content of the (political) advertising*.

- **Advertiser**: *(self-reported) description of the advertiser running the ad,* such as the name of the candidate or political-party/organisation, the location, etc. This data is important for matching particular content and actions to the political groups under analysis.

- **Time period**: *start and end timestamps for when the ad was live on the platform*. This is crucial for mapping, identifying, and studying the effects of advertising.

- **Amount spent**: *monetary value the advertiser spent on running a given ad on the platform.* This data can be used to study the strategies of different political actors (for example, do conservative parties advertise on some platforms and left-leaning parties on others? Do they pay the same for advertising on a given platform?), and the conditions under which resources (inequalities) influence political outcomes.

- **Targeted features**: *the particular targeting and exclusion criteria (e.g. socio-demographic, location) specified by the advertiser for a given ad*. This information can be used to study under what conditions political advertisers target different segments of the online population, and their potential effects (Votta et al. 2024).

4.7. Platform interventions

Platforms play a key role in content curation. For example, platforms can decide to take down (Casas 2024; Mosleh et al. 2024) or reduce the visibility (Jaidka et al. 2023) of content and accounts, and prioritise the recommendation of particular content (Brown et al. 2022). Platforms often change their terms of service (Landi 2024), update their content moderation policies, and sometimes implement urgent "break glass" measures during sensitive times such as elections (Jackson 2024). Detailed information on the moderation and recommendation policies for political content is key to understanding whether observed dynamics are the output of user versus platform behaviour. Additionally, this information is

key to assessing the (unintended) effects of platform actions, and to designing improved interventions.

4.8. External data and models

Most quantitative studies on political online safety need to be able to: **(a)** combine social media data, with **other data** collected by the researchers (such as combining datasets of URLs mentioned on social media posts, with a list of untrustworthy news sites previously assembled by the researcher); and/or **(b)** use external and/or previously **trained (machine-learning) models for analysis** (such as a machine-learning model to automatically identify toxic content among a dataset of social media posts).

4.9. Cross-platform research

Cross-platform research is crucial for understanding, for example, how (mis/dis)information travels from one platform to another (Buntain et al. 2021), the prevalence of different communities and political narratives in different platforms (Kakavand 2024), and comparing platform interventions and effects. Yet, mapping content and users across platforms remains challenging. Future approaches to platform research should aim to solve these challenges — for example, by building a single system where researchers can query and analyse data from all platforms.

4.10. Computing and storage resources

Political online safety researchers need **(a)** substantial amounts of **storage**, and **(b)** access to state-of-the-art **computational resources** (such as powerful graphics processing units [GPUs]), to analyse large quantities of multimodal social media data. The rise of visual-based social media platforms, such as Instagram and TikTok, means that data collections of millions of social media posts can take several terabytes of storage. Today researchers can fine-tune Large (Visual) Language Models (L(V)LMs) to train machine-learning models capable of accurately classifying these large amounts of social media data into theoretical quantities of interest (for example, their political topic, and identify hateful content and misinformation). Yet, researchers need access to several costly GPUs for training and deploying these L(V)LMs. These storage and computational needs must be taken into account when thinking about the best governance model for providing academics with access to social media data.

5. **Data access**. State of the art

In this section we provide an overview of the state of the art regarding data access for political online safety research for the five largest social media platforms in the United Kingdom, according to the latest Digital News Report (2024): Facebook (used by 63% of the British population), YouTube (53%), Instagram (38%), X (25%), and TikTok (15%).

5.1. Meta: Facebook and Instagram

5.1.1. Meta Content Library and API

The Meta Content Library provides independent researchers access to data from public Facebook, Instagram, and Threads accounts. Researchers can access the data via a point-and-click *User Interface* (UI), or programmatically via an *Application Programming Interface* (API).

Accessible data
Post and **engagement** data for the following types of public **accounts**:

- **Facebook**: public pages, public groups, public events, and public profiles with a verified badge or 1,000+ followers.
- **Instagram**: public business, public creators, personal accounts set to public with a verified badge or 1,000+ followers.
- **Threads**: public profiles with 1,000+ followers.

Inaccessible data

- **Posts, engagement,** and **account** data for less salient **public** accounts, or those with less than 1,000 followers.
- **Network** data for **public** or **private** accounts.
- **Exposure** data for **public** or **private** accounts. No exposure data for **(political) ads** either.
- **Posts** and **engagement** data from **private** accounts.

Data access

- **Who?** Researchers affiliated with a qualified academic or research institution, and journalists working for non-profit organisations (Fischer 2024).
- **How?**
    - An independent intermediary body, the Inter-university Consortium for Political Science Research (ICPSR, at the University of Michigan), manages access to the Meta Content Library.
    - ICPSR receives, independently reviews, and approves applications from independent researchers.
    - Applications must provide information about the principal investigator, collaborators, general research goals, and analytical skills and needs. Additionally, those applying to the API (rather than the UI version), must provide additional details regarding programming language and analytical tools needed, and coding/statistical experience; plus a Restricted Data Use

Agreement to be signed by the applicant (principal investigator) or their institution.

Data analysis

- **Controlled environment**. Meta and ICPSR allow researchers approved by ICPSR to access and analyse (accessible) data only in controlled environments hosted by Meta and ICPSR, to maximise data privacy and security.
    - *User Interface* (UI). Researchers using the UI are given access to a platform where they can easily query (accessible) data in a point-and-click fashion. This platform runs on Meta's servers, and researchers access the platform via a url in their browser, and using the credentials provided by ICPSR after having been granted access. Researchers can initially inspect the returned data using a grid or list view, and then visualize trends in a dashboard.
    - *Application Programming Interface* (API). Researchers using the API can programmatically query and analyse (accessible) data in a virtual machine (VM) hosted by ICPSR. Researchers access these VMs via a VPN and using the credentials provided by ICPSR; and they can fire up jupyter notebooks for programmatically querying and analysing the data, using the python and/or R programming languages. A key **limitation** is that researchers cannot bring any data and/or trained models into the environment, which limits the researcher's ability to incorporate external datasets into the analysis and to use machine-learning models to identify theoretical quantities of interest in the data (such as toxic language, extreme views, misinformation).
- **Downloadable data**. However, as of recently, Meta allows researchers to **download** data for further analysis to be conducted offline or in an environment of the researcher's choice. Yet the criteria for downloadable data is even more restrictive, being limited to **posts** and **engagement** for public Facebook, Instagram, and Threads **accounts** with 25,000+ followers. Researchers are able to bring external and additional datasets and machine-learning models into the analysis of this data.

5.1.2. Meta Ad Library

The Meta Ad Library is a dataset of ads people can see on Meta platforms. Similar to the Meta Content Library, researchers can access and search the general Meta Ad Library using a *User Interface* (UI) or an *Application Programming Interface* (API). Additionally, approved researchers can also access the Ad Targeting Dataset, with additional information on ads on social issues, elections, and politics (for 120+ countries, since August 2020).

Accessible data

- For ads labeled as **not** being about social issues, elections, or politics:
  (only for **active** ads, or any ad delivered in European Union territory in the last year)
    - Advertiser name.
    - Advertiser profile description.
    - Targeted features: location (only when targeting an EU territory), gender, age.

- For ads labeled as being about **social issues, elections, or politics**:
  (**active** ads, and inactive ads published in the last 7 years)
    - Advertiser name.

    - Advertiser detailed information (location, contact information).
    - Amount spent (range).
    - Impressions received (range).
    - Time period when the ad was active.
    - Targeted features: location, gender, age, users' interests.

Inaccessible data

- Inactive ads that are not labeled as being about social issues, elections, or politics (non-political ads).
- Advertiser location and contact information for non-political ads.
- Amount spent for non-political ads.
- Active time period for non-political ads.
- More detailed targeted features for non-political ads.
- **Exposure** information for political or non-political ads (what individual users have been exposed to a given ad).
- **Engagement** information for political or non-political ads (what individual users have clicked on the ad).

Data access

- **Who?**
    - General Meta Ad Library: everyone.
    - Ad Targeting Dataset: only approved researchers from academic or non-profit institutions.
- **How? (Ad Targeting Dataset)**
    - ICPSR receives, independently reviews, and approves applications from independent researchers.

Data analysis

- **General Meta Ad Library**
    - **User Interface:** researchers can query data based on location, broad topical category, and keywords. They then see the ads in a list format, and explore them by clicking on them one by one, making Large-N analysis extremely complicated, if not impossible.
    - **API**: researchers can query the dataset programmatically, using the same data filters: location, broad topical category, and keywords. The API returns a dataset with the variables/fields described above, that can be saved and downloaded locally, making it suitable for large-N analysis.
- **Ad Targeting Dataset**
    - **Controlled environment**. Researchers approved by ICPSR can only query data via Meta's Research Platform, a virtual machine hosted by Meta where researchers can run remote python or R notebooks to query and analyse the data. The returned data must be analysed in this controlled environment, and cannot be downloaded locally for further analysis. No external data can be taken into the environment.

5.1.3. URL shares (Facebook only)

A dataset of **all** URLs shared on Facebook in 46 countries, and information about how (different kinds of) users engaged with the URLs. The dataset is updated approximately every 6 to 12 months. The latest release (v10, from April 12, 2023) contains information about 68 million URLs, for over 3.1 million rows. The data is protected using differential privacy — some *noise* has been added to avoid revealing any (private) user-level information, yet the data can still be used for valid scientific inference.

Accessible data

- URL **metadata:** such as full URL, top domain, posting time, times reported as false news, and rating from third-party fact checkers.
- URL **exposure/engagement** data: number of views, clicks, shares (with and without click), comments, likes, loves, hahas, wows, sorrys, and angers. These counts can be obtained by: country, year-month, age bracket, gender, and ideological bracket.

Inaccessible data

- Full (original) post/s that included the URL.
- Account information for those who (re)posted the URL.
- Account, and other posts, exposure, engagement, and network information for those exposed to URLs.

Data access

- **Who?**
  - Researchers from academic institutions who want to study the effects of social media on democracy and elections, and only for academic purposes.
- **How?**
  - Social Science One (hosted by Harvard's Institute for Quantitative Social Science) reviews and approves applications for accessing the data. Applications are accepted/reviewed once a quarter.
  - Additionally, approved applications/researchers (and their research institutions) need to sign a Research Data Agreement with Meta.

Data analysis

- **Controlled environment**. Researchers must use Meta's Research Platform, which consists of VMs hosted by Meta and with no internet access, where researchers can run python or R notebooks for querying and analysing the data.

5.1.4. 2020 United States Election Study replication data

Meta partnered with 17 academic researchers to study the impact of social media (Facebook and Instagram) on people's political attitudes during the 2020 Presidential election in the United States. Several studies came out of this partnership, and replication datasets for each study are being made public after publication (5 of them are available so far).

Accessible data

- This varies by the publication to be replicated. These contain a mix of URL- and user-level information, with detailed information about exposure and engagement metrics,

as well as user's attitudes, for those users who participated in the studies. The extensive codebooks for each replication material are available at the following link: https://socialmediaarchive.org/search?c=US2020&cc=US2020&ln=en.

Inaccessible data

- Only replication data is available; no additional data can be collected or queried.

Data access

- **Who?**
  - Only approved researchers from academic or non-profit institutions.
- **How?**
  - ICPSR receives, independently reviews, and approves applications from independent researchers.
  - Sign "Restricted Data Agreement".

Data analysis

- **Controlled environment.** Researchers approved by ICPSR must analyse these replication datasets using jupyter or R notebooks running on protected servers hosted by ICPSR. Researchers cannot use any additional external data and models for analysing this data.

5.2. YouTube

5.2.1. YouTube Research Program API

Academics affiliated with an accredited high-learning institution can apply to the YouTube Research Program to use their API to collect data for their research projects.

Accessible data

- **Account** (channel) information: such as unique identifiers, description, tags/labels, number of subscribers, etc.
- **Post** (video) information: such as video URL, transcript/close-caption when available, creation date, etc.
- Partial **network** information: list of channels to which a given account is subscribed to.
- **Comments**: text of the comments posted to a given video, key information for the accounts posting the comment.
- **Engagement** (video-level): number of views, likes, and comments for a given video.

Inaccessible data

- **Exposure**: no list of videos and/or comments to which a given user/account has been exposed.
- **Engagement**: no list of videos with which a given user/account has engaged (liked, commented on, shared).
- Partial **network** information: no list of subscribers to a given channel provided.

- **Platform interventions**: no information about changes in moderation or recommendation policies, and lack of information on, for example, account and video suspension (often generic *Not Available* reason provided*)*.

Data access

- **Who?** Academics affiliated with an accredited high-learning institution.
- **How?** Applications reviewed internally by YouTube.

Data analysis

- Data queried and collected programmatically via the API.
- Easy to store locally, and/or to merge with other external dataset, and/or use external models in the analysis.

5.3. X

5.3.1. X DSA Research API Access

After Elon Musk took over X, the company dismantled the Research API, widely used by researchers and journalists to study the impact of social media on democracy and elections, among other topics. Currently, only EU researchers who fulfill all criteria described in Article 40 of the European Union Digital Services Act, and for the sole purpose of working on the *detection, identification and understanding of systemic risks in the Union*, can access data from X for free. All other researchers can only access data using one of the paid subscription plans, which cost several thousands a month, and are out of reach for practically the entirety of researchers.

5.4. TikTok

5.4.1. TikTok Research API

TikTok provides academics with access to their data via their Research API.

Accessible data

- **Account** information for **public** accounts (most TikTok accounts are public), such as unique names and identifiers, whether the account is verified, account description, number of followers and followees, number of videos posted, and number of likes received.
- **Post** information for videos posted by **public** accounts, such as unique post identifier, creation datetime, post title/description, video label/s, close-captions/descriptions for some videos, unique identifiers for the author account, region code, and engagement measures (number of views, shares, comments).
- **Comment** information for videos posted by **public** accounts, such as the text, the creation datetime, and the number of likes and replies.
- **Network** information for **public** accounts: list of followers and list of followees.
- **Engagement** information for **public** accounts: post/video information for the videos a given account has liked, pinned, and reposted.

Inaccessible data

- **Account, Post, Comment, Network, Exposure,** and **Engagement** data is not available for **private** accounts.
- **Exposure** information is not available for any account. This is particularly relevant given the critical role of algorithms in determining content exposure on TikTok.
- **Post** information does not include the actual video or video frames.
- Information about **platform interventions** is also not available.

Data access

- **Who?** Academic researchers based in the United States, United Kingdom, European Economic Area, and Switzerland, and only for the purpose of conducting academic non-commercial research.
- **How?** Applications must be submitted to and are reviewed by TikTok.

Data analysis

Researchers must access the data programmatically using the API. Researchers use the API to download the data locally, and they can use external data and models in the analysis of the data.

6. **Values** relevant for social media research

Here, we outline the key values that should govern social media research, for this research to be credible and conducted in an ethical manner. These values include: trust, equality, transparency, reproducibility and replicability, data privacy, data security, and consent.

6.1. Trust

Trust refers to the confidence individuals and institutions place on the quality and the findings of research. Trustworthy information on the role of social media in politics and society is crucial for having meaningful public discussions on the topic and for advancing policy-making in this area.

Trust in social media platforms is low across many countries: only around 22% of citizens find them trustworthy according to a 21-country report from Ipsos (2023). The report also points to 43% of respondents advocating for further regulation of the platforms. Social media companies have their own economic incentives: to monetise engagement for targeted advertising. Given that polarising content and misinformation often achieve high levels of engagement, the public may not trust the platforms to prioritise public and political safety over engagement.

In recent years, platforms have engaged in a variety of public research efforts on political online safety, such as publishing internal research in open academic outlets (Bakshy et al. 2015), publishing data reports on their websites (for example, X's report on Removal Requests), providing activity and compliance reports to public agencies (for example, Meta's reports on compliance with European Union legislation), making relevant datasets available for independent analysis (such as X releasing a dataset of information operations; see Gadde and Roth 2018), conducting research with external academics (such as Meta's academic collaboration during the 2020 United States election), and making platform data available via research APIs (such as Youtube or TikTok's API).

Nevertheless, the aforementioned lack of trust in the platforms can cast doubt in these efforts. For example, recent information about "break glass" measures that Meta inadvertently put in place during the 2020 United States election collaboration with academics has cast doubt about the generalisability of those findings (Guess et al. 2023a; Bagchi et al. 2024). Researchers have also reported concerning mismatch between data provided to academics via APIs and data shown for the same content on the platform (Pearson et al. 2024). Additionally, platforms can be subject to a change in philosophy or ownership at any time, as emphasised by Elon Musk's takeover of Twitter and discussions around TikTok ownership in the United States.

In contrast, British society has high trust in academics. A survey conducted by YouGov (2022) for the British Academy of Science indicates that 65% of British citizens find academics to be "knowledgeable", and 50% trust them to deliver information.

In turn, for research on social media and political online safety to be trustworthy, independent researchers, and/or an independent intermediary organisation with no

economic interests, must play a key role in facilitating academic access to data. Beyond providing access to the data requested by the independent researchers (following safe and secure protocols in line with other values discussed in this section), the platforms should detach from the sampling and analytical processes as much as possible.

6.2. Equality

Equality refers here to free access to social media data and to the necessary analytical tools in order to level the playing field between researchers with different resources and skills. This promotes inclusivity in research and allows for more diverse voices to contribute to knowledge.

Social media data is under the control of social media companies. While researchers have been able to access some of it free of charge, policies can change suddenly and unexpectedly. For example, X, which had historically given researchers much larger access to its API than other platforms, monetised the API following a change in ownership (Kupferschmidt 2023). Even prior to that, the company provided access to two different versions of the API — the free-of-charge Streaming API which provided a subset of all published content, and its Firehose API which provided access to all public tweets. While data collected through the latter would be complete, the Firehose API was not equally accessible to all researchers given its high cost and the large amount of computational resources to retain the data (Morstatter et al. 2013).

A second component of equality in research concerns engineering infrastructure for analysing the data. In order to study the complexities of the social media environment, and to obtain more precise estimates, researchers often aim to test their hypotheses with large amounts of social media data. However, querying, storing, and analysing large amounts of (multimodal) data can be expensive — researchers need to have enough fundings to pay for virtual machines for data collection, large storage volumes, and computing nodes for analysis. In particular, computing costs have skyrocketed in the last few years, given that sophisticated expensive GPUs are needed to train and deploy LLM-based machine-learning models.

To ensure that all researchers, independent of their resources and computational skills, are able to conduct political online safety research, all vetted researchers should be able to have free access to the relevant social media data for their projects and to the necessary analytical tools for analysing such data.

6.3. Transparency

In research, Transparency refers to the degree of detail and disclosure about the specific steps, decisions, and judgment calls made during a scientific study (Aguinis et al. 2024). Research transparency fosters trust, integrity, the credibility of research findings, and researcher accountability.

Transparency encompasses different dimensions (Moravcsik 2014). Data transparency refers to providing access to the evidence or data used in the research (Moravcsik 2014),

including detailed explanations on the data collection, such as sources, and sampling procedures (Wulff et al. 2023). Analytic transparency refers to providing access to the data analysis process (Moravcsik 2014), such as information on the softwares used and reproducible scripts. Production transparency refers to providing clear information about, and justifications of, the evidence, arguments, and methods that were selected out of all the possible choices (Moravcsik 2014; Aguinis et al. 2024).

All three dimensions are of key relevance, as together they allow the public and scholars to evaluate whether the data has been collected, sampled, analysed, and interpreted correctly, and to understand the process by which researchers make inferences. They also lessen the dangers of selection, confirmation, or publication biases, whether intentional or not (Miguel et al. 2014).

Transparency entails several practices, such as pre-registration and pre-analysis plans, sharing study data, reproducible scripts, protocols, and methodology, and publicly reporting research findings (Aczel et al. 2020; Toth et al. 2021; Aguinis et al. 2024). These practices allow for the reproducibility and replication of the work.

Public research efforts from the platforms align with some of these transparency dimensions and practices, but not with others. For example, Meta committed to not restrict the publication of findings coming from the 2020 United States election academic partnership, all studies were pre-registered, and analytical datasets have been made openly available for reproducibility and further analysis via ICPSR (González-Bailón et al. 2023; Guess et al. 2023a; Guess et al. 2023b; Nyhan et al. 2023; Allcott et al. 2024). However, Meta was not transparent about some parallel interventions by the platform that could have affected the results of the studies (Bagchi et al. 2024), and other external researchers are not able to fully reproduce the studies because they lack the same level of data access as the authors. As another example, in the past, X has openly shared with researchers datasets of messages posted by accounts the company had identified as being involved in foreign information operations. However, little was known about the data selection process, that is, how the company had identified these accounts as malicious and chosen specifically these out of potentially many.

6.4. Reproducibility and replicability

Reproducibility and replicability are crucial to ensure the validity and integrity of data and research findings, improve the accuracy of existing hypotheses, and revisit theories based on real-world developments and in different contexts (Brodeur et al. 2024). Reproduction and replication in social media research face specific challenges as it typically relies on dynamic data obtained from third-parties.

Reproducibility refers to obtaining consistent results and conclusions based on the original data and code of a study. It allows us to confirm the validity and reliability of the original study. It has three types (Dreber and Johannesson 2023): computational reproducibility (same data and code from the original studies), recreate reproducibility (using the information in the original studies without access to the data and code), and robustness reproducibility (same data with alternative analytical decisions).

Replicability refers to testing whether results of a study are consistent when using different data or methodology to answer the same research question as the original study. Replication allows us to improve the general accuracy of results and test whether they hold when making alternative sampling or methodological decisions. Replication can be direct (using the original studies' research design and analysis with new data) or conceptual (testing the same hypothesis with an alternative research design/analysis and new data; Dreber and Johannesson 2023).

Reviews in several disciplines find that reproducibility and replicability rates are generally low (see Camerer et al. 2018; Gertler et al. 2018), even in cases with assistance from the original authors. Reasons for this include the data being private in nature, unavailable or incomplete computer code, flexibility in data collection and reporting, specification searching (known as "p-hacking"), and presenting post-hoc hypotheses (Simmons et al. 2011; Rubin 2017; Christensen and Miguel 2018; Gertler et al. 2018; Bryan et al. 2019; Broder et al. 2020; Chang and Li 2022).

Studies relying on social media data will likely never be fully reproducible, but they may be replicable. This is because such data and its availability is in constant change. Platforms change their API access, prevent researchers from retaining the datasets, provide only a sample of the available data with no information on how this data was selected, and may suspend users or take down certain posts; while users may switch their privacy settings, delete old posts, or delete their account altogether (Ruths and Pfeffer 2014; Hutton and Henderson 2015; Graham and Huffer 2020; Morstatter et al. 2021; Davidson et al. 2023; Corson and Pierri 2024; Knöpfle and Schatto-Eckrodt 2024; Pearson et al. 2024). In other words, a researcher collecting social media data at a certain point in time will hardly be able to collect the exact same dataset at a later point. Yet moving forward we should strike for a data governance structure that facilitates reproducibility.

Even when unable to share the raw original data, researchers can adopt transparency standards to support reproducibility and allow for replication. These include sharing the pre-analysis plan, data sampling strategy, methodology, ethical procedures that were followed (including whether users were informed; which personally identifiable information or sensitive data was removed), and the analysis code (Hutton and Henderson 2018). Documenting the workflow enables other researchers to understand the context in which the original research was conducted, how the results were produced, and overall improves the credibility of the research.

6.5. Data privacy

Data privacy can be defined as an individual's ability to control the amount and use of information gathered about them, including control over how the data is collected, stored, shared, and used (Malhotra et al. 2004; Mahmoodi et al. 2018), as well as their awareness of data privacy practices of social media companies and researchers.

Data privacy is a core component of ethical research that aims to protect individuals from any harms that would not occur in the absence of data collection — such as data breaches, cyber threats, and any other security concerns. When investigating components of political online safety, researchers must establish clear privacy standards to protect individuals'

identities, so as not to undermine their political rights. Data of interest for social media research includes, among others, survey data and data on individuals' social media activity. In the former case, data privacy standards would be covered by government regulations (such as the Data Protection Act in the United Kingdom) and research would in most cases require approval from Institutional Review Boards (McCarthy 2008).

There is less consensus on data privacy standards when it comes to social media data because most of it is public. Users do consent to having their data collected when they sign up and post on a platform, thus donating their personal data to platforms and making some of it accessible to external researchers, governments, and private companies (Lauterwasser and Nedzhvetskaya 2023). While some personal data is kept closed to the public, even individuals concerned about their privacy tend to disregard privacy policies (Acquisti and Gross 2006).

Overall, users unknowingly become research observations not only for platforms but also for external researchers without those users' explicit and informed consent. Any research relying on their patterns of behaviour online could violate their trust.

Users may be aware that their platform engagement is public and are given the option to change their privacy settings. However, they are not informed of how their data will be used, whom it will be shared with, and are not guaranteed that what they may wish to keep private (such as their personal information they are required to disclose when signing up on a platform) will be fully protected by social media companies. Platforms hold an enormous amount of sensitive personal information, including not only on-platform engagement, but also off-platform behaviour and geo-location information. This sensitive data can be potentially mishandled: the Cambridge Analytica scandal, for example, illustrates that private records of users can be obtained by actors aiming to influence public opinion and elections (Rosenberg et al. 2018).

Privacy violations can also occur unintentionally. For example, researchers can predict users' personal characteristics with high accuracy based on their public engagement online (Kosinski et al. 2013), which raises the ethical concerns of protecting one's identity and anonymity. A further consideration is that data that was made public at a certain point can quickly become private if a user changes their privacy settings later on (Kekulluoglu et al. 2022), but there is no guarantee that this data will not be used by external researchers who collected it prior to the change in settings.

6.6. Data security

Data security in research refers to maintaining control over the data used from the inception to well after a research project's completion (Wackenhut 2018). It is foundational to data privacy, and thus to ethical research, but is distinct from it because it is primarily concerned with the secure storing and sharing of data in order to prevent unauthorised access, disclosure, and alteration of data. In this way, data security serves the purpose of not only protecting subjects' anonymity and data privacy, but also ensuring that a research project itself remains uncompromised by preventing the risk of alterations.

When it comes to social media research, both social media companies and researchers have the obligation to keep individuals' data secure so as not to harm them. For the former, for example, recent data breaches and leaks aiming to influence election outcomes and enabling micro-targeting for the spread of misinformation and polarising content (such as the influence operations conducted by the Russian Internet Research Agency and the Cambridge Analytica scandal; Bateman and Jackson 2024) shed light on the need to invest in cybersecurity.

For researchers relying on users' social media data or any other type of data (such as surveys), security guidelines must be established before, throughout, and after the data collection and research process. Steps to minimise security threats include sharing the data only with those who need it and storing it in a secure way (on an external drive with encrypted data, rather than a cloud-storage service), as well as collecting the necessary data only and deleting any data that is not needed for the specific research project (Tankala 2022). Any personal data stored must be anonymised as required by the Data Protection Act in the United Kingdom. Personal data refers to information relating to an identified or (directly or indirectly) identifiable person, and anonymisation includes deleting any separate list in which a respondent's or user's reference number can be matched to an identifiable person.

One consideration that has been overlooked by political scientists is that the context in which data has been collected and the research project conceptualised can change over time. Researchers must look "beyond the field" (Knott 2019) and ensure that data is kept secure long after the analysis and publication of findings.

Until recently, most online safety researchers have collected, stored, and analysed (large amounts of) public social media data locally, and without asking for explicit consent from the users. One exception is the current data-sharing model from Meta, which only allows researchers to access and analyse their data in clean safe environments hosted by Meta or ICPSR. This governance model is an important step forward in terms of protecting users' privacy and data security. Future governance models should aim to build on this experience. Ideally, such clean safe environments would be hosted by a trusted third-party that would also regularly assess the quality and robustness of the data provided by the platforms.

6.7. Consent

Informed consent is the process through which research subjects agree to participate in a research project, and this agreement should be based on informing them of what data will be collected about them, how it will be used and by whom, how it will be secured, the nature of confidentiality and anonymity, and any potential harms that might arise from the research (Knott 2019).

The aim of obtaining informed consent is to establish the rights of individuals and empower them to make autonomous decisions about the risks of participating in research.

In social media research, where much of the data of interest is public, users' consent is implied but may not be informed. Using public data obtained from social media platforms does not typically require permission or informed consent, and an analysis of studies using

Twitter data showed that very few of these actually discuss ethics reviews (Zimmer and Proferes 2014).

However, from a users' perspective, their consent for the public to view their content is distinct from their consent for their content to be collected and analysed by researchers (Zimmer 2010). As an implication, although researchers remain exempt from legal obligations to obtain users' permission, they do have an ethical responsibility to inform users on the collection and use of their data. There is evidence of Twitter users believing that obtaining permission should be an ethical rule in research (Fiesler and Proferes 2018), which suggests that violating users' right to be informed and consent to their data being used could undermine trust in social media companies and researchers.

When it comes to social media platforms, they are granted the permission to collect, use, and share user data when users sign-up and post on them; however, there is evidence that users rarely read or fully understand the platforms' terms and conditions, which are overly long and complicated (Reidenberg et al. 2015). This has allowed social media companies to conduct controversial experiments on users; for example, in 2012, Facebook conducted a psychological experiment testing the effect of the newsfeed on users' emotions (Hill 2014; Kramer et al. 2014). This experiment created personal harm to users, who may not have fully understood the implications of Facebook's data use policy when they agreed to it.

As social media research continues to grow, ethical norms should be established even around the use of public data, with the primary goal of protecting users. Users are generally uncomfortable with their data being used by researchers without explicit permission, but this varies by context. Contextual factors include research discipline, who is collecting the data, how much and what type of data is being collected, what the purpose of the research is, as well as whether the data is anonymised (Williams et al. 2017; Fiesler and Proferes 2018; Gilbert et al. 2021). This highlights the importance of seeking users' consent for specific research projects, as they may be comfortable with sharing their data in certain cases but not in others. It also highlights the importance of thinking clearly about the purpose of a research project ahead of time, and to only use whatever data is needed for the project at hand.

7. A governance model for academic access to social media data

7.1. Towards an independent intermediary body

As suggested in previous sections, the current governance model is characterised by power asymmetries between social media companies and researchers. Social media companies are in control of the data that can be shared with researchers (which may not be complete), the vetting process (which threatens research independence), and have the resources to terminate any research project if viewed to be against their interests (either commercial or reputational).

In this context, recent legislation aims to facilitate social media research moving forward (Riley and Ness 2022). In the European Union, the Digital Service Act requires platforms to provide data access to vetted researchers upon request from a government enforcement authority. In the United States, the Platform Accountability and Transparency Act, which is still under consideration, would create a data access mandate along with a new government office and process dedicated to vetting researchers and overseeing their engagements with platforms. In the United Kingdom, the Online Safety Act gives authority to Ofcom to assess existing platform practices related to research.

Legislation is only a first step towards supporting research on the role and effect of social media platforms on politics, society, and democracy. To put them into practice, an independent body trusted by all parties must develop detailed standards that should govern social media research, ensure that these standards are followed, act as a mediator between social media companies and researchers, while also supporting regulators and promoting collaboration between different stakeholders — such as academic researchers from various disciplines, journalists, civil society actors, platforms, and regulators from different government agencies.

In this section, we outline what we believe should be the core functions of this independent body. We take into account data governance models from other industries, previous proposals from different organisations, current practices in social media research, and the values we believe should govern this research environment. We particularly build on the recent work of the European Digital Media Observatory (EDMO), which called for the establishment of an Independent Intermediary Body to oversee academic access to social media data in the European Union (EDMO 2023). This initiative was launched following recommendations from a previous EDMO working group that drafted a Code of Conduct on data access and compliance with the General Data Protection Regulation (under Article 40 of the Digital Service Act; EDMO 2022).

In particular, we will go over six key functions we believe this body should undertake. First, the independent body should be in charge of the vetting process of researchers and research projects, developing a standardised set of rules and reviewing applications. Second, this body should oversee and ensure the quality and type of data provided by platforms, by developing a common codebook and conducting periodic audits of the data.

Third, it should set up the necessary infrastructure for researchers to conduct their analyses, by hosting a secure virtual environment and contributing the necessary computational resources and tools to analyse data from multiple platforms. Fourth, the body should have the authority to mediate disputes that may arise between researchers and platforms. Fifth, the body should provide logistical and technical support to researchers. Finally, the body should engage all relevant stakeholders from its set-up, including regulators, policy-makers, platforms, researchers, civil society actors, and members of the general public, and promote collaboration to advance the field.

7.1.1. The vetting process

Under the current social media research regime, social media companies are in charge of the vetting process, with the exception of data from Meta offered through their partnership with ICPSR. In order to access the data they need, researchers must submit a detailed application to social media companies which lays out their research questions, the data they would like to access, the purpose, and their affiliation with an academic or other research institution.

This is problematic because platforms could be reluctant to share data for research projects that may threaten their reputation, undermining research independence. Additionally, researchers have complained about existing vetting processes being slow, about delays in finalising agreements with platforms, and delays in obtaining the data once an agreement is reached (Vogus 2022). This is particularly challenging for those whose work is time sensitive.

Research applications must also address researchers' commitments to data privacy and their plans to keep the data secure. However, currently, there is a lack of clear unified standards when it comes to ethical practices and data storage and sharing in social media research.

In order to address these issues, the independent body should take on the vetting and reviewing process. This should be based on a pre-defined and standardised set of rules applicable to every research project and data request, including standardised data protection expectations, and fitness for purpose test to ensure that the data requested is appropriate (Riley and Ness 2022). Having an independent body in charge of the vetting and reviewing process could also ensure adherence to strict timelines, which is an issue that has been identified by researchers (Vogus 2022). The EDMO (2023) Working Group also suggests that, for the European Union, the independent body could accredit other organisations, such as Institutional Review Boards in academic institutions, to conduct the vetting process themselves if they abide by the body's guidelines; while in cases where researchers do not have access to such an institution, the body would conduct the process itself, thus levelling the playing field when it comes to data access requests — as currently, non-academic researchers (such as those working for think tanks or media organisations) have raised the issue of having their applications rejected due to their lack of academic affiliation (Vogus 2022).

7.1.2. Data type and quality

Social media companies maintain control over the data they share with researchers. This data has been found sometimes to be incomplete, inaccurate, and unclear (Vogus 2022; Pearson et al. 2024). Additionally, there are stark inconsistencies regarding the kinds of data shared by each platform, and a lack of documentation for the shared data, undermining the quality and reliability of research conducted using this data. Moreover, there is no robust channel for researchers to have a discussion with platforms regarding access to additional (types of) data (Vogus 2022).

In order to resolve these issues, an independent body should develop a common codebook for all platforms (EDMO 2023). Inspired by the data needs described in Section 4 of this report, and other demands from the research community, this codebook should define the data that should be shared by platforms, provide a detailed description of the provided data and its structure (such as variables and categories), and request platforms to provide information on the data limitations (such as whether some data is unavailable, and for what reason). One starting point could be following a model similar to the Data Documentation Initiative, which provides a standard for describing metadata, and adapting it to social media data.

Given known inconsistencies between the data observed on social media platforms and the data provided to researchers (Pearson et al. 2024), the independent body should also conduct regular audits to assess the robustness and the quality of the provided data (Ausloos et al. 2020) and establish sanctions when the platforms fail to accurately provide the requested data.

Furthermore, data needs are not static and as new threats to political online safety emerge, researchers should be given the opportunity to reassess the codebooks and request additional data. Rather than approving requests on a case-by-case basis, the intermediary body should have regular calls for input from researchers to allow them to set (new) data priorities and needs. The body would then discuss with social media platforms on the feasibility of providing this data and amend the codebooks and protocols accordingly.

When setting the standards on the nature of the data that can be accessed, the intermediary body should ensure that such data is shared in an ethical manner, where users' privacy is protected and researchers and platforms comply with data protection laws. Building on the model established by the Meta-ICPSR partnership (discussed in Sections 5.1.1, 5.1.2, 5.1.4 of this report), in the next section we advocate for the independent intermediary body to host, support, and manage a secure research environment where vetted researchers can query and analyse data from all platforms; including bringing their own complementary datasets and computational models into the environment.

7.1.3. Research environment and infrastructure

Currently, in most social media research projects, researchers collect, store and analyse social media data locally in their own computers and servers (with the exception of Meta data queried and analysed via the partnership with ICPSR), putting data security and privacy in danger and potentially violating relevant ethical standards. The secure storing and

manipulation of data prevents unauthorised access, the alteration of data, and data leaks, all of which can create harm for individuals and compromise a research project.

To address these issues, we believe that the independent body should provide the necessary research infrastructure by hosting a safe virtual computing environment, where vetted researchers gain access to their requested data and conduct the data analysis directly, similar to the Meta-ICPSR model. The environment should provide access to data from all platforms, thus allowing cross-platform research (Klinger and Ohme 2023). Given that the data would not be stored by researchers locally, the environment should allow them to input additional data and models from other sources. This could be done, for example, following a review of the additional data to ensure its quality and relevance. Allowing the input of external data and models has large benefits, such as being able to study interactions between behaviour on social media and real-world events and conditions. For example, Williams et al. (2020) study the impact of online hate speech on offline hate crimes by combining administrative, survey, and social media data, and Wischerath et al. (2024) study the impact of COVID-19 misinformation by combining social media data with a dataset of public protests.

Regarding the data analysis, the hosting environment should be accessible to researchers with varying skills in order to ensure inclusivity. The interface could perform data driven tasks, allowing researchers to aggregate, transform, and visualise data through a point-and-click system. But it should also allow researchers to query and analyse data using programming languages such as R and python (by for example providing access to jupyter or Rstudio notebooks/sessions), and provide access to novel machine-learning models such as those accessible via open-source platforms like HuggingFace. Additionally, researchers should be able to run their analyses in virtual environments with access to powerful GPUs, to be able to use state of the art machine-learning models when analysing large amounts of (multimodal) social media data.

7.1.4. Mediating disputes

During the research process, conflicts may arise between social media companies and researchers. Researchers may raise claims of insufficient data access provided by platforms, with the data shared not matching what was previously agreed upon; while platforms could claim that researchers are violating their protocols or data privacy laws (Riley and Ness 2022).

For example, Facebook routinely sends cease-and-desist letters to researchers conducting investigations of its platform, claiming that these investigations violate their terms of service, even when that is not the case (Abdo et al. 2022). One specific account is of a project monitoring Instagram’s newsfeed algorithm (Kayser-Bril, Algorithm Watch 2020) which had to be terminated following Facebook’s wrongful claims that it violated its terms of service and data protection requirements under the General Data Protection Regulation. Although the project did not violate any rules or regulations, Algorithm Watch terminated it and deleted all of its data under the threat of lawsuit.

In cases like these, an independent body could act as an arbiter and determine whether there was a breach of contract, violation of platforms’ policies, violation of existing laws on

the part of researchers, or whether the platforms' claims are unfounded. This body would establish clear procedures for receiving complaints and resolving disagreements that do not need legal enforcement, which would support researchers in appealing the decisions made by platforms, and support platforms by reviewing potential violations of their policies by researchers (EDMO 2023). The independent body would review the complaints and propose remedial measures, and in cases where these measures are rejected, the body could develop a complete record of the dispute with sufficient detail to allow regulators to proceed (Riley and Ness 2022).

7.1.5. Stakeholder engagement

When setting up the standards that should govern social media research, the independent body should engage with various stakeholders, including researchers from different disciplines, representatives of social media companies, regulators and government agencies, members of civil society organisations, and the public. A collaborative approach would ensure that all relevant actors are represented in setting up the governance model rather than leaving it under the exclusive control of social media companies or state authorities.

Additionally, a formal and stable discussion forum for the various stakeholders is necessary given the constantly evolving nature of social media communications, and their potential threats to political online safety. For researchers, data needs are constantly evolving, and allowing them to (re)define the data priorities would strengthen their ability to conduct relevant and timely research in conversation with regulatory authorities and platforms. For policy-makers, enabling interactions with researchers would facilitate evidence-based policy-making in this crucial area.

One challenge in social media research relates to obtaining user consent, as discussed in Section 6.7 of this report. Users may not be fully aware of the data that social media companies collect and share for research purposes. Engaging members of the general public would promote trust, transparency, and legitimacy when it comes to social media research. This research often serves users' interests because it investigates the (potentially negative) effects of social media. The intermediary body should also design an appropriate public outreach strategy, to make sure that the public is fully aware of research being conducted in this area, and the high ethical and data privacy standards guiding the research.

7.1.6. Technical support

In addition to setting up and maintaining the secure research environment discussed above, researchers from diverse backgrounds will require technical and engineering support in accessing and using the environment. Hence, we propose the independent body to invest in an IT department to support these needs.

First, the independent body should include an IT team to assist researchers with the onboarding once they gain access to the virtual environment, and to assist them with any technical issues they face while using the tool. In parallel, the body should provide training to researchers on how to use the environment, including making the data requests, saving their projects, importing additional data and models, and exporting their final results.

Additionally, researchers would benefit from this IT team to work on building material and tools that can be of use for several research projects and researchers, such as workshops on particular topics, sample jupyter and/or R notebooks/scripts, and computational tools, such as fine-tuned machine learning models for identifying different types of content.

We thank the reader for exploring our analysis and recommendations, and we refer them back to our Executive Summary for a statement of our six main arguments.

8. References

Abdo, A., Krishnan, R., Krent, S., Falcón, E. W., and Woods, A. K. (2022). A Safe Harbor for Platform Research. *Knight First Amendment Institute*. Available at: https://knightcolumbia.org/content/a-safe-harbor-for-platform-research (accessed on Jan. 16, 2025).

Acquisti, A. and Gross, R. (2006). Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In Danezis, G. and Golle, P. (eds) *Privacy Enhancing Technologies. PET 2006. Lecture Notes in Computer Science*, 4258, 36-58.

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, S., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M. A., Ioannidis, J. I., Johnson, E., Jonas, K., Kousta, S., Lilienfeld, S. O., Lindsay, S., Morey, C. C., Munafò, M., Newell, B. R., Pashler, H., Shanks, D. R., Simons, D. J., Wicherts, J. M., Albarracin, D., Anderson, N. D., Antonakis, J., Arkes, H. R., Back, M. D., Banks, G. C., Beevers, C., Bennett, A. A., Bleidorn, W., Boyer, T. W., Cacciari, C., Carter, A. S., Cesario, J., Clifton, C., Conroy, R. N., Cortese, M., Cosci, F., Cowan, N., Crawford, J., Crone, E. A., Curtin, J., Engle, R., Farrell, S., Fearon, P., Fichman, M., Frankenhuis, W., Freund, A. M., Gaskell, M. A., Giner-Sorolla, R., Green, D. P., Greene, R. L., Harlow, L. L., de la Guardia, F. H., Isaacowitz, D., Kolodner, J., Lieberman, D., Logan, G. D., Mendes, W. B., Moersdorf, L., Nyhan, B., Pollack, J., Sullivan, C., Vazire, S., and Wagenmakers, E. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, 4, 4-6.

Adams, T. (2024). 'It's OK, everyone else is doing it': How do we deal with role violence on social media played in UK riots? *The Guardian*. Article available at: https://www.theguardian.com/media/article/2024/aug/11/its-ok-everyone-else-is-doing-it-how-do-we-deal-with-role-violence-on-social-media-played-in-uk-riots (accessed on Jan. 16, 2025).

Adler, W. T. and Thakur, D. (2021). A Lie Can Travel: Election Disinformation in the United States, Brazil, and France. *Center for Democracy and Technology*. Report available at: https://cdt.org/wp-content/uploads/2021/12/2021-12-13-CDT-KAS-A-Lie-Can-Travel-Election-Disinformation-in-United-States-Brazil-France.pdf (accessed on Jan. 16, 2025).

Aguinis, H., Li, Z. A., Der Foo, M. (2024). The research transparency index. *The Leadership Quarterly*, 35(4), 101809.

Ahmed, N. and Bales, K. (2021). Inside the radicalised anti-vaxxer network 'influencing' government vaccine advisory panel. *Byline Times*. Report available at: https://bylinetimes.com/2021/10/01/inside-the-radicalised-anti-vaxxer-network-influencing-government-vaccine-advisory-panel/ (accessed on Jan. 16, 2025).

Allcott, H., Gentzkow, M., Mason, W., Wilkins, A., Barberá, P., Brown, T., Cisneros, J. C., Crespo-Tenorio, A., Dimmery, D., Freelon, D., González-Bailón, S., Guess, A. M., Kim, Y. M., Lazer, D., Malhotra, N., Moehler, D., Nair-Desai, S., Nait El Barj, H., Nyhan, B., Paixao de Queiroz, A. C., Pan, J., Settle, J., Thorson, E., Tromble, R., Velasco Rivera, C., Wittenbrink, B., Wojcieszak, M., Zahedian, S., Franco, A., Kiewiet de Jonge, C., Stroud, N. J., and Tucker, J. A. (2024). The effects of Facebook and Instagram on the 2020 election: A deactivation experiment. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 121(21), e2321584121.

Allen M. R., Desai, N., Namazi, A., Leas, E., Dredze, M., Smith, D. M., and Ayers, J. W. (2024). Characteristics of X (Formerly Twitter) Community Notes Addressing COVID-19 Vaccine Misinformation. *JAMA*, 331(19), 1670-1672.

Al-Rawi, A. (2021). Telegramming hate: Far-right themes on dark social media. *Canadian Journal of Communication*, 46(4), 821-851.

Amnesty International (2017). Amnesty reveals alarming impact of online abuse against women. Press release available at: https://www.amnesty.org/en/latest/press-release/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/ (accessed on Jan. 16, 2025).

Are, C. (2020). How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist Media Studies*, 20(5), 741-744.

Ausloos, J., Leerssen, P., and ten Thije, P. (2020). Operationalizing Research Access in Platform Governance What to learn from other industries? *Algorithm Watch*. Available at: https://algorithmwatch.org/en/wp-content/uploads/2020/06/GoverningPlatforms_IViR_study_June2020-AlgorithmWatch-2020-06-24.pdf (accessed on Jan. 16, 2025).

Awan, I., Sutch, H., and Carter, P. (2019). Extremism Online - Analysis of extremist material on social media. Report prepared for the *UK Commission for Countering Extremism*, available at: https://assets.publishing.service.gov.uk/media/5d8b7bd2e5274a08c8cc0d1b/Awan-Sutch-Carter-Extremism-Online.pdf (accessed on Jan. 16, 2025).

Azzimonti, M. and Fernandes, M. (2023). Social media networks, fake news, and polarization. *European Journal of Political Economy*, 76, 102256.

Bagchi, C., Menczer, F., Lundquist, J., Tarafdar, M., Paik, A., and Grabowicz, P. A. (2024). Social media algorithms can curb misinformation, but do they? arXiv:2409.18393.

Bakshy, E., Messing, S. and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130-1132.

Ball-Burack, A., Seng Ah Lee, M., Cobbe, J., and Singh, J. (2021). Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection. In Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3-10, 2021, Virtual Event, Canada. ACM, New York, NY, USA.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10), 1531-1542.

Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., and Tucker, J. A. (2019). Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 113(4), 883-901.

Barberá, P. (2020). Social media, echo chambers, and political polarization. In *Social media and democracy: The state of the field, prospects for reform*, Persily, N. and Tucker, J. A. (eds). Cambridge University Press.

Barlett, C. P. (2015). Anonymously hurting others online: The effect of anonymity on cyberbullying frequency. *Psychology of Popular Media Culture*, 4(2), 70-79.

Barnes, M. R. (2022). Online Extremism, AI, and (Human) Content Moderation. *Feminist Philosophy Quarterly*, 8(3/4), 1-28.

Bateman, J. and Jackson, D. (2024). Countering Disinformation Effectively: An Evidence-Based Policy Guide. *Carnegie Endowment for International Peace*. Report available at: https://carnegieendowment.org/research/2024/01/countering-disinformation-effectively-an-evidence-based-policy-guide (accessed on Jan. 16, 2025).

Bergh, A. (2024). Guide to social media training with Somulator – 2024. *FFI - Norwegian Defence Research Establishment*. Report available at: https://www.ffi.no/en/publications-archive/guide-to-social-media-training-with-somulator-2024 (accessed on Jan. 16, 2025).

Bishop-Froggatt, J. (2024). What did the big political parties spend on Google Ads in the 2024 UK Election? *8 Million Stories*. Article available at: https://8ms.com/blog/uk-election-google-ad-spend/ (accessed on Jan. 16, 2025).

Bond, R. and Messing, S. (2015). Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review*, 109(1), 62-78.

Booth, R., Weaver, M., Hern, A., Smith, S., and Walker, S. (2017). Russia used hundreds of fake accounts to tweet about Brexit, data shows. *The Guardian*. Article available at: https://www.theguardian.com/world/2017/nov/14/how-400-russia-run-fake-accounts-posted-bogus-brexit-tweets (accessed on Jan. 16, 2025).

Born, K. and Edgington, N. (2017). Analysis of Philanthropic Opportunities to Mitigate the Disinformation/Propaganda Problem. *Hewlett Foundation*. Report available at: https://hewlett.org/wp-content/uploads/2017/11/Hewlett-Disinformation-Propaganda-Report.pdf (accessed on Jan. 16, 2025).

Bovet, A., Grindrod, P. (2022). Organization and evolution of the UK far-right network on Telegram. *Applied Network Science*, 7(76).

Boxell, L., Gentzkow, M., and Shapiro., J.M. (2017). Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Science,* 114(40), 10612-10617.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2024). Cross-country trends in affective polarization. *Review of Economics and Statistics*, 106(2), 557-565.

Bradshaw, S., Bailey, H., and Howard, P. N. (2021). Industrialized disinformation: 2020 global inventory of organized social media manipulation. *Computational Propaganda Project, Oxford Internet Institute*. Report available at: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/01/CyberTroop-Report-2020-v.2.pdf (accessed on Jan. 16, 2025).

Bragg, S., Ringrose, J., Mohandas, S., Cambazoglu, I., Bartlett, D., Barker, G., Gupta, T., and Merriman, J. (2022). The state of UK boys: Understanding and Transforming Gender in the
Lives of UK Boys. *Global Boyhood Initiative*. Report available at: https://www.equimundo.org/wp-content/uploads/2022/12/State-of-UK-Boys-Long-Report.pdf (accessed on Jan. 16, 2025).

Brattberg, E. and Maurer, T. (2018). Russian Election Interference: Europe's Counter to Fake News and Cyber Attacks. *Carnegie Endowment for International Peace*. Report available at: https://carnegie-production-assets.s3.amazonaws.com/static/files/CP_333_BrattbergMaurer_Russia_Elections_Interference_FINAL.pdf (accessed on Jan. 16, 2025).

Brodeur, A., Cook, N., and Heyes, A. (2020). Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review*, 110 (11), 3634-3660.

Brodeur, A., Esterling, K., Ankel-Peters, J., Bueno, N. S., Desposato, S., Dreber, A., Genovese, F., Green, D. P., Hepplewhite, M., Hoces de la Guardia, F., Johannesson, M., Kotsadam, A., Miguel, E., Velez, Y. R., and Young, L. (2024). Promoting Reproducibility and Replicability in Political Science. *Research & Politics*, 11(1).

Brown, M., Bisbee, J., Lai, A., Bonneau, R., Nagler, J., and Tucker, J. A. (2022). Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users.

Bryan, C. J., Yeager, D. S., and O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, 116(51), 25535-25545.

Buntain, C., Bonneau, R., Nagler, J., and Tucker, J. A. (2021). YouTube Recommendations and Effects on Sharing Across Online Social Platforms. *Proceedings of the ACM on Human-Computer Interaction,* 5(CSCW1), 1-26.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Iman, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E., and Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2, 637-644.

Carolan, L. (2024). The three tech stories of GE24. *The Briefing*. Available at: https://www.thebriefing.ie/the-techstory-of-ge24/ (accessed on Jan. 16, 2025).

Casas, A., Menchen-Trevino, E., and Wojcieszak, M. (2023). Exposure to extremely partisan news from the other political side shows scarce boomerang effects. *Political Behavior*, 45(4), 1491-1530.

Casas, A. (2024). The Geopolitics of Deplatforming: A Study of Suspensions of Politically-Interested Iranian Accounts on Twitter. Political Communication, 41(3), 413-434.

Center for Countering Digital Hate. The disinformation dozen. *Center for Countering Digital Hate*. Report available at: https://counterhate.com/research/the-disinformation-dozen/ (accessed on Jan. 16, 2025).

Chadwick, A., Vaccari, C., and O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New Media & Society*, 20(11), 4255-4274.

Chang, A. C. and Li, P. (2022). Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not". *Critical Finance Review*, 11(1), 185-206.

Chokshi, N. (2019). PewDiePie in Spotlight After New Zealand Shooting. *The New York Times.* Article available at: https://www.nytimes.com/2019/03/15/technology/pewdiepie-new-zealand-shooting.html (accessed on Jan. 16, 2025).

Christensen, G. and Miguel, E. (2018). Transparency, Reproducibility, and the Credibility of Economics Research. *Journal of Economic Literature*, 56 (3): 920-80.

Chu, X., Vliegenthart, R., Otto, L., Lecheler, S., de Vreese, C., and Kruikemeier, S. (2023). Do Online Ads Sway Voters? Understanding the Persuasiveness of Online Political Ads. *Political Communication*, 41(2), 290-314.

Colley, T. and Moore, M. (2022). The challenges of studying 4chan and the Alt-Right: 'Come on in the water's fine'. *New Media & Society*, 24(1), 5-30.

Collignon, S. and Rüdig, W. (2021). Increasing the cost of female representation? The gendered effects of harassment, abuse and intimidation towards Parliamentary candidates in the UK. *Journal of Elections, Public Opinion and Parties*, 31(4), 429-449.

Cook, J. (2019). Instagram's shadow ban on vaguely "inappropriate" content is plainly sexist. *Huffington Post*. Article available at: https://www.huffpost.com/entry/instagram-shadow-ban-sexist_n_5cc72935e4b0537911491a4f (accessed on Jan. 16, 2025).

Corson, F. and Pierri, F. (2024). What we can learn from TikTok through its Research API. arXiv:2402.13855v2.

Data Documentation Initiative. *DDI Alliance*. Available at: https://ddialliance.org/ (accessed on Jan. 16, 2025).

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25-35. Association for Computational Linguistics.

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., van der Linden, D., Roscoe, J. F., Ayravainen, L., and Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, 7, 2054-2057.

Del Rey, M. (2024). Texas campus in uproar after protesters hold signs declaring 'women are property' on quad after Trump victory. Independent. Article available at: https://www.independent.co.uk/news/world/americas/us-politics/texas-university-trump-women-signs-b2643468.html (accessed on Jan. 16, 2025).

Deltapoll (2021). Climate Change and Disinformation. Commissioned by the *New Political Communication Unit*, Royal Holloway, University of London. Unpublished.

Di Meco, L. (2023). Monetizing Misogyny. Gendered Disinformation and the Undermining of Women's Rights and Democracy Globally. *She Persisted*. Report available at: https://she-persisted.org/wp-content/uploads/2023/02/ShePersisted_MonetizingMisogyny.pdf (accessed on Jan. 16, 2025).

Disinfodex database. Available at: https://disinfodex.org/ (accessed on Jan. 16, 2025).

Dobber, T., Ó Fathaigh, R. and Zuiderveen Borgesius, F. J. (2019). The regulation of online political micro-targeting in Europe. *Internet Policy Review*, 8(4), 1-20.

Dreber, A. and Johanneson, M. (2023). A Framework for Evaluating Reproducibility and Replicability in Economics. Available at SSRN: https://ssrn.com/abstract=4458153.

Eady, G., Nagler, J., Guess, A., Zilinsky, J., and Tucker, J. A. (2019). How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data. *Sage Open*, 9(1).

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., and Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14, 62.

Eady, G., Bonneau, R., Tucker, J. A., and Nagler, J. (2024). News Sharing on Social Media: Mapping the Ideology of News Media, Politicians, and the Mass Public. *Political Analysis*, 1-18.

Earl, J., Maher, T. V., and Pan, J. (2022). The digital repression of social movements, protest, and activism: A synthetic review. *Science Advances*, 8(10), eabl8198.

Esposito, E. and Breeze, R. (2022). Gender and politics in a digitalised world: Investigating online hostility against UK female MPs. *Discourse & Society*, 33(3), 303-323.

European Digital Media Observatory (2022). Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access. *Institute for Data, Democracy & Politics*. Available at: https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf (accessed on Jan. 16, 2025).

European Digital Media Observatory (2023). Core Tasks and Principles for an Independent Intermediary Body that Will Facilitate Researchers' Access to Platform Data. *Institute for Data, Democracy & Politics*. Available at: https://iddp.gwu.edu/sites/g/files/zaxdzs5791/files/2023-11/creating_an_independent_intermediary_body_to_facilitate_platform_research.pdf (accessed on Jan. 16, 2025).

Evans, A. (2023). Andrew Tate: How schools are tackling his influence. *BBC*. Article available at: https://www.bbc.com/news/education-64234568 (accessed on Jan. 16, 2025).

Farrell, T., Fernandez, M., Novotny, J., and Alani, H. (2019). Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. Association for Computing Machinery, 87-96.

Fazackerley, A. (2023). 'Vulnerable boys are drawn in': Schools fear spread of Andrew Tate's misogyny. *The Guardian*. Article available at: https://www.theguardian.com/society/2023/jan/07/andrew-tate-misogyny-schools-vulnerable-boys (accessed on Jan. 16, 2025).

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1), 3.

Fiesler, C. and Proferes, N. (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1).

Fiorina, M. P. and Abrams, S. J. (2008). Political polarization in the American public. Annual Review of Political Science, 11(1), 563-588.

Fischer, S. (2024). Meta shuts down data tool widely used by journalists. AXIOS. Available at: https://www.axios.com/2024/03/19/meta-shut-off-data-access-to-journalists (accessed on Jan. 16, 2025).

Fox, K. (2024). UK rocked by far-right riots fueled by online disinformation about Southport stabbings. *CNN*. Article available at: https://edition.cnn.com/2024/08/01/uk/southport-attack-disinformation-far-right-riots-intl-gbr (accessed on Jan. 16, 2025).

François, C. and Douek, E. (2021). The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations. *Journal of Online Trust and Safety*, 1(1), 1-30.

Freelon, D., Bossetta, M., Wells, C., Lukito, J., Xia, Y., and Adams, K. (2022). Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation. *Social Science Computer Review,* 40(3), 560-578.

Fung, B. (2024). UK riots show how social media can fuel real-life harm. It's only getting worse. *CNN*. Article available at: https://edition.cnn.com/2024/08/09/tech/uk-protests-social-media/index.html (accessed on Jan. 16, 2025).

Gadde, V. and Roth, Y. (2018). Enabling further research of information operations on Twitter. *X Blog*. Available at: https://blog.x.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter (accessed on Jan. 16, 2025).

Garzia, D., Ferreira da Silva, F., and Maye, S. (2023). Affective Polarization in Comparative and Longitudinal Perspective. *Public Opinion Quarterly*, 87(1), 219-231.

Gertler, P., Galiani, S., and Romero, M. (2018). How to make replication the norm. *Nature*, 554, 417-419.

Gilbert, S., Vitak, J., and Shilton, K. (2021). Measuring Americans' Comfort With Research Uses of Their Social Media Data. *Social Media + Society*, 7(3).

Gill, P., Corner, E., Conway, M., Thornton, A., Bloom, M., and Horgan, J. (2017). Terrorist Use of the Internet by the Numbers: Quantifying Behaviors, Patterns and Processes. *Criminology & Public Policy*, 16(1), 99-117.

Gillespie, T. (2020). Content Moderation, AI, and the Question of Scale. *Big Data & Society*, 7(2).

Gillett, F. (2024). Influencers driving extreme misogyny, say police. *BBC*. Article available at: https://www.bbc.com/news/articles/cne4vw1x83po (accessed on Jan. 16, 2025).

Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A., and Tucker, J. A. (2020). Cross-Platform State Propaganda: Russian Trolls on Twitter and YouTube during the 2016 U.S. Presidential Election. *The International Journal of Press/Politics*, 25(3), 357-389.

González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., Iyengar, S., Kim, Y. M.,

Malhotra, N., Moehler, D., Nyhan, B., Pan, J., Rivera, C. V., Settle, J., Thorson, E., Tromble, R., Wilkins, A., Wojcieszak, M., de Jonge, C. K., Franco, A., Mason, W., Stroud, N. J., and Tucker, J. A. (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, 381(6656), 392-398.

Gorrell, G., Bakir, M. E., Roberts, I., Greenwood, M. A., and Bontcheva, K. (2020). Which politicians receive abuse? Four factors illuminated in the UK general election 2019. *EPJ Data Science*, 9(1), 18.

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

Graham, S. and Huffer, D. (2020). Reproducibility, Replicability, and Revisiting the Insta-Dead and the Human Remains Trade. *Internet Archaeology*, 55.

Greenwood, M. A., Bakir, M. E., Gorrell, G., Song, X., Roberts, I., and Bontcheva, K. (2020). Online Abuse of UK MPs from 2015 to 2019: Working Paper. arXiv:1904.11230v1.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425), 374-378.

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1), eaau4586.

Guess, A. M. (2021). (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. *American Journal of Political Science*, 65, 1007-1022.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Mie Kim, Y., Lazer, D., Moehler, D., Nyhan, B., Velasco Rivera, C., Settle, J., Thomas, D. R., Thorson, E., Tromble, R., Wilkins, A., Wojcieszak, A., Xiong, B., de Jonge, C. K., Franco, A., Mason, W., Stroud, N. J., Tucker, J. A. (2023a). How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381, 398-404.

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Mie Kim, Y., Lazer, D., Moehler, D., Nyhan, B., Velasco Rivera, C., Settle, J., Thomas, D. R., Thorson, E., Tromble, R., Wilkins, A., Wojcieszak, A., Xiong, B., de Jonge, C. K., Franco, A., Mason, W., Stroud, N. J., Tucker, J. A. (2023b). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, 381, 404-408.

Haimson, O. L., Delmonaco, D., Nie, P., and Wegner, A. (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(2), 466.

Haimson, O. L. and Hoffmann, A. L. (2016). Constructing and enforcing “authentic” identity online: Facebook, real names, and non-normative identities. *First Monday*, 21(6).

Hall, H. (2023). The rise of the mega misogynists. *Cosmopolitan*. Article available at: https://www.endviolenceagainstwomen.org.uk/wp-content/uploads/2023/11/Mega-Misogynists-PDF.pdf (accessed on Jan. 16, 2025).

Hanscom, R., Silbergleit Lehman, T., Lv, Q., and Mishra, S. (2024). The Toxicity Phenomenon Across Social Media. arXiv:2410.21589v1.

Haslop, C., Ringrose, J., Cambazoglu, I., and Milne, B. (2024). Mainstreaming the Manosphere's Misogyny Through Affective Homosocial Currencies: Exploring How Teen Boys Navigate the Andrew Tate Effect. *Social Media + Society*, 10(1).

Hendrickson, C. and Galston, W. A. (2019). Big tech threats: Making sense of the backlash against online platforms. *The Brookings Institution*. Report available at: https://www.brookings.edu/articles/big-tech-threats-making-sense-of-the-backlash-against-online-platforms/ (accessed on Jan. 16, 2025).

Hewitt L., Broockman, D., Coppock, A., Tappin, B. M., Slezak, J., Coffman, V., Lubin, N., and Hamidian, M. (2024). How Experiments Help Campaigns Persuade Voters: Evidence from a Large Archive of Campaigns' Own Experiments. *American Political Science Review*, 118(4): 2021-2039.

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

Hill, K. (2014). Facebook Doesn't Understand The Fuss About Its Emotion Manipulation Study. *Forbes*. Article available at: https://www.forbes.com/sites/kashmirhill/2014/06/29/facebook-doesnt-understand-the-fuss-about-its-emotion-manipulation-study/ (accessed on Jan. 16, 2025).

Ho, D. E., King, J., Wald, R. C., and Wan, C. (2021). Building a National AI Research Resource: A Blueprint for the National Research Cloud. The Stanford Institute for Human-Centered Artificial Intelligence. Available at: https://hai.stanford.edu/sites/default/files/2022-01/HAI_NRCR_v17.pdf (accessed on Jan. 16, 2025).

Hobolt, S. B., Leeper, T. J., and Tilley, J. (2021). Divided by the Vote: Affective Polarization in the Wake of the Brexit Referendum. *British Journal of Political Science*, 51(4), 1476-1493.

Hokka, J. (2021) PewDiePie, racism and Youtube's neoliberalist interpretation of freedom of speech. *Convergence*, 27(1), 142-160.

Hope Not Hate (2019). State of Hate 2019. People vs. the elite? Report available at: https://hopenothate.org.uk/wp-content/uploads/2019/02/state-of-hate-2019-final-1.pdf (accessed on Jan. 16, 2025).

House of Commons (2018). Data Protection Act 2018. Available at: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted/data.htm (accessed on Jan. 16, 2025).

House of Commons, Digital, Culture, Media and Sport Committee (2019). Disinformation and 'fake news': Final Report, Eighth Report of Session 2017–19. Report available at: https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf (accessed on Jan. 16, 2025).

Howard, P. N. and Kollanyi, B. (2016). Bots, #StringerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum. Unpublished Research Note, Oxford University Press.

Howard, P. N., Ganesh, B., and Liotsiou, D. (2018). The IRA, Social Media and Political Polarization in the United States, 2012-2018. *Computational Propaganda Research Project*. Report available at:

https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2018/12/IRA-Report-2018.pdf (accessed on Jan. 16, 2025).

Hutton, L. and Henderson, T. (2018). Toward Reproducibility in Online Social Network Research. *Transactions on Emerging Topics in Computing*, 6(1), 156-167.

Hutton, L. and Henderson, T. (2021). Making Social Media Research Reproducible. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(4), 2-7.

Institute for Strategic Dialogue (2024). From rumours to riots: How online misinformation fuelled violence in the aftermath of the Southport attack. Report available at: https://www.isdglobal.org/digital_dispatches/from-rumours-to-riots-how-online-misinformation-fuelled-violence-in-the-aftermath-of-the-southport-attack/ (accessed on Jan. 16, 2025).

Ipsos (2023). Global Trustworthiness Monitor. Stability in an Unstable World. *Ipsos*. Available at: https://www.ipsos.com/sites/default/files/ct/publication/documents/2023-01/ipsos-global-trustworthiness-monitor-stability-in-an-unstable-world.pdf (accessed on Jan. 16, 2025).

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual review of political science*, 22(1), 129-146.

Jackson, D. (2017). Issue Brief: Distinguishing Disinformation from Propaganda, Misinformation and Fake News. *National Endowment for Democracy*. Brief available at: https://www.ned.org/issue-brief-distinguishing-disinformation-from-propaganda-misinformation-and-fake-news/ (accessed on Jan. 16, 2025).

Jackson, D. (2024). We Know a Little About Meta's "Break Glass" Measures. We Should Know More. *Tech Policy*. Available at: https://www.techpolicy.press/we-know-a-little-about-metas-break-glass-measures-we-should-know-more/ (accessed on Jan. 16, 2025).

Jaidka, K., Mukerjee, S., and Lelkes, Y. (2023). Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse. *Journal of Communication*, 73(2), 163-178.

Jang, H., Barrett, B., and McGregor, S. C. (2023). Social media policy in two dimensions: understanding the role of anti-establishment beliefs and political ideology in Americans' attribution of responsibility regarding online content. *Information, Communication & Society*, 27(6), 1047-1072.

Jones, D. R. (2001). Party polarization and legislative gridlock. *Political Research Quarterly*, 54(1), 125-141.

Joseph, C. (2019). Instagram's murky 'shadow bans' just serve to censor marginalised communities. *The Guardian*. Article available at: https://www.theguardian.com/commentisfree/2019/nov/08/instagramshadow-%20bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive (accessed on Jan. 16, 2025).

Juarez Miro, C. and Toff, B. (2023). How right-wing populists engage with cross-cutting news on online message boards: The case of ForoCoches and Vox in Spain. *The International Journal of Press/Politics*, 28(4), 770-790.

Kakavand, A. E. (2023). Far-right Social Media Communication in the Light of Technology Affordances: A Systematic Literature Review. *Annals of the International Communication Association*, 48(1), 37-56.

Kalmoe, N. P. and Mason, L. (2022). *Radical American partisanship: Mapping violent hostility, its causes, and the consequences for democracy*. University of Chicago Press.

Karl, P. (2017). Hungary's radical right 2.0. *Nationalities Papers*, 45(3), 345-355.

Kayser-Bril, N. (2021). AlgorithmWatch forced to shut down Instagram monitoring project after threats from Facebook. *Algorithm Watch*. Available at: https://algorithmwatch.org/en/instagram-research-shut-down-by-facebook/ (accessed on Jan. 16, 2025).

Kayyali, D. and Althaibani, R. (2017). Vital Human Rights Evidence in Syria is Disappearing from YouTube. *WITNESS*. Article available at: https://blog.witness.org/2017/08/vital-human-rights-evidence-syria-disappearing-youtube/ (accessed on Jan. 16, 2025).

Kekulluoglu, D., Vaniea, K., and Magdy, W. (2022). Understanding Privacy Switching Behaviour on Twitter. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 31, 1-14.

Kim, J. W., Guess, A., Nyhan, B., and Reifler, J. (2021). The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication*, 71(6), 922-946.

Klinger, U. and Ohme, J. (2023). What the Scientific Community Needs from Data Access under Art. 40 DSA. *Weizenbaum Institute for the Networked Society*. Available at: https://www.weizenbaum-library.de/server/api/core/bitstreams/db396630-94c5-4751-9eed-ebad7bd65f97/content (accessed on Jan. 16, 2025).

Knöpfle, P. and Schatto-Eckrodt, T. (2024). The Challenges of Replicating Volatile Platform-Data Studies: Replicating Schatto-Eckrodt et al. (2020). *Media and Communication*, 12, 7789.

Knott, E. (2019). Beyond the Field: Ethics after Fieldwork in Politically Dynamic Contexts. *Perspectives on Politics*, 17(1), 140–153.

Koch, L., Ghawi, R., Pfeffer, J., and Steinert, J. I. (2024). Online Misogyny Against Female Candidates in the 2022 Brazilian Elections: A Threat to Women's Political Representation? arXiv:2403.07523.

Koehler, D. (2014). The radical online: Individual radicalization processes and the role of the Internet. *The Journal for Deradicalization*, 1, 116-134.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802–5805.

Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788-8790.

Kubin, E. and von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3), 188-206.

Kupferschmidt, K. (2023). Twitter's plan to cut off free data access evokes 'fair amount of panic' among scientists. *Science*, 379(6633), 624-625.

Landi, M. (2024). Social media moderation: How does it work and what is set to change? *Independent*. Available at: https://www.independent.co.uk/news/uk/politics/ofcom-elon-musk-britain-twitter-b2592617.html (accessed on Jan. 16, 2025).

Lauterwasser, S. and Nedzhvetskaya, N. (2023). Privacy in Public?: The Ethics of Academic Research with Publicly Available Social Media Data. *Berkley Journal of Sociology*. Article available at: https://berkeleyjournal.org/2023/08/11/privacy-in-public/ (accessed on Jan. 16, 2025).

Ledwich, M. and Zaitsev, A. (2020). Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization. *First Monday*, 25(3).

Lee, S., Gil de Zúñiga, H., and Munger, K. (2023). Antecedents and consequences of fake news exposure: A two-panel study on how news use and different indicators of fake news exposure affect media trust. *Human Communication Research*, 49(4), 408-420.

Llansó, E. (2019) Platforms want centralized censorship. That should scare you. *Wired*. Article available at: https://www.wired.com/story/platforms-centralized-censorship/ (accessed on Jan. 16, 2025).

Lukito, J. (2020). Coordinating a multi-platform disinformation campaign: Internet Research Agency Activity on three US Social Media Platforms, 2015 to 2017. *Political Communication*, 37(2), 238-255.

Lynch, P., Sherlock, P., and Bradshaw, P. (2022). Scale of abuse of politicians on Twitter revealed. *BBC*. Article available at: https://www.bbc.co.uk/news/uk-63330885 (accessed on Jan. 16, 2025).

Magdy W., Darwish, K., Abokhodair, N., Rahimi, A., and Baldwin, T. (2016). #ISISisNotIslam or #DeportAllMuslims? Predicting unspoken views. *Proceedings of the 8th ACM Conference on Web Science (WebSci '16)*. Association for Computing Machinery, New York, NY, USA, 95-106.

Mahmoodi, J., Čurdová, J., Henking, C., Kunz, M., Matić, K., Mohr, P., and Vovko, M. (2018). Internet Users' Valuation of Enhanced Data Protection on Social Media: Which Aspects of Privacy Are Worth the Most? Frontiers in psychology, 9, 1516.

Malhotra N. K., Kim S. S., and Agarwal J. (2004). Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. Information Systems Research, 15(4), 336-355.

Martin, D. A., Shapiro, J. N., and Ilhardt, J. G. (2020). Trends in Online Influence Efforts. *Empirical Studies of Conflict*. Report available at: https://esoc.princeton.edu/publications/trends-online-influence-efforts (accessed on Jan. 16, 2025).

Martin, D. A., Shapiro, J. N., and Ilhardt, J. G. (2023). Introducing the Online Political Influence Efforts dataset. *Journal of Peace Research*, 60(5), 868-876.

McCarthy, C. R. (2008). The origins and policies that govern institutional review boards. In *The Oxford Textbook of Clinical Research Ethics*, Emanuel, E. J., Grady, C. C., Crouch, R. A., Lie, R. K., Miller, F. G., Wendler, D. D. (eds). Oxford University Press.

McCarty, N., Poole, K. T., and Rosenthal, H. (2016). *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press.

Meta. Regulatory and Other Transparency Reports. *Meta*. Available at: https://transparency.meta.com/reports/regulatory-transparency-reports/ (accessed on Jan. 16, 2025).

Meta. Research partnership to understand Facebook and Instagram's role in the U.S. 2020 election. *Meta*. Available at: https://research.facebook.com/2020-election-research/ (accessed on Jan. 16, 2025).

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., and Van der Laan, M. (2014). Social science. Promoting transparency in social science research. *Science*, 343(6166), 30-31.

Moore, A., Fredheim, R., Wyss, D., and Beste, S. (2021). Deliberation and identity rules: The effect of anonymity, pseudonyms and real-name requirements on the cognitive complexity of online news comments. *Political Studies*, 69(1), 45-65.

Moravcsik, A. (2014). Transparency: The Revolution in Qualitative Research. *PS: Political Science & Politics*, 47(1), 48-53.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. (2021). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1), 400-408.

Mosleh, M., Yang, Q., Zaman, T., Pennycook, G., and Rand, D. G. (2024). Differences in misinformation sharing can lead to politically asymmetric sanctions. *Nature*, 634, 609-616.

Munger, K. and Phillips, J. (2019). A Supply and Demand Framework for YouTube Politics. Penn State Political Science. Paper available at: https://osf.io/73jys/download (accessed on Jan. 16, 2025).

Munn, L. (2020). Angry by design: toxic communication and technical architectures. *Humanities and Social Sciences Communications*, 7, 53.

Nadim, M. and Fladmoe, A. (2021). Silencing Women? Gender and Online Harassment. *Social Science Computer Review*, 39(2), 245-258.

Nagle, A. (2017). *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right*. John Hunt Publishing.

Nasuto, A. and Rowe, F. (2024). Exposing Hate - Understanding Anti-Immigration Sentiment Spreading on Twitter. arXiv:2401.06658.

Nelimarkka, M., Laaksonen, S. M., and Semaan, B. (2018). Social Media Is Polarized, Social Media Is Polarized: Towards a New Design Agenda for Mitigating Polarization. In

Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18). Association for Computing Machinery, New York, NY, USA, 957–970.

Newman, N., Fletcher, R., Robertson, C. T., Arguedas, A. R., and Nielsen, R. K. (2024). Reuters Institute Digital News Report 2024. *Reuters Institute for the Study of Journalism*. Available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024 (accessed on Jan. 16, 2025).

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Guess, A. M., Kennedy, E., Kim, Y. M., Lazer, D., Malhotra, N., Moehler, D., Pan, J., Thomas, D. R., Tromble, R., Velasco Rivera, C., Wilkins, A., Xiong, B., de Jonge, C. K., Franco, A., Mason, W., Jomini Stroud, N., and Tucker, J. A. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, 620, 137-144.

Ofcom (2024) Understanding misinformation: an exploration of UK adults' behaviour and attitudes. Report available at: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/making-sense-of-media/dis-and-mis-information-research/mis-and-disinformation-report.pdf?v=386069 (accessed on Jan. 16, 2025).

Office for National Statistics. Prevalence of ongoing symptoms following coronavirus (Covid-19) infection in the UK. *Office for National Statistics*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/7october2021 (accessed on Jan. 16, 2025).

Ognyanova, K., Lazer, D., Robertson, R. E., and Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*, 1(4), 1-19.

Pascual-Ferrá, P., Alperstein, N., Barnett, D. J., and Rimal, R. N. (2021). Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic. *Big Data & Society*, 8(1).

Pauwels, L. and Schils, N. (2016). Differential online exposure to extremist content and political violence: Testing the relative strength of social learning and competing perspectives. *Terrorism and Political Violence*, 28(1), 1-29.

Pearson, E. (2024). *Extreme Britain: Gender, Masculinity and Radicalization*. Oxford and New York: Oxford University Press.

Pearson, G. D. H., Silver, N. A., Robinson, J. Y., Azadi, M., Schillo, B. A., and Kreslake, J. M. (2024). Beyond the margin of error: a systematic and replicable audit of the TikTok research API. *Information, Communication & Society*, 1-19.

Pennycook, G., Cannon, T., and Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology*, 147(12), 1865-1880.

Plevin, L. (2024). How have UK political parties utilised Paid Media in the 2024 General Election? *Propellernet*. Article available at: https://www.propellernet.co.uk/how-have-uk-political-parties-utilised-paid-media-in-the-2024-general-election/ (accessed on Jan. 16, 2025).

Pradel, F. and Theocharis, Y. (2024). Gender Differences in Demanding Moderation. Pre-print version available at: https://osf.io/preprints/osf/aw6tk.

Pradel, F., Zilinsky, J., Kosmidis, S., and Theocharis, Y. (2024). Toxic Speech and Limited Demand for Content Moderation on Social Media. *American Political Science Review*, 1-18.

Regehr, K., Shaughnessy, C., Zhao, M., and Shaughnessy, N. (2024). Safer Scrolling: How algorithms popularise and gamify online hate and misogyny for young people. University of London, University of Kent, and Association of School and College Leaders. Report available at: https://www.ascl.org.uk/ASCL/media/ASCL/Help%20and%20advice/Inclusion/Safer-scrolling.pdf (accessed on Jan. 16, 2025).

Reidenberg J. R., Breaux T., Cranor L. F., and French B. (2015). Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30(1), 39-68.

Riley, C. and Ness, S. (2022). A Module Playbook for Platform-to-Researcher Data Access. Tech Policy. Available at: https://www.techpolicy.press/a-module-playbook-for-platform-to-researcher-data-access/ (accessed on Jan. 16, 2025).

Roose, K. (2019). The Making of a YouTube Radical. *The New York Times*. Available at: https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html (accessed on Jan. 16, 2025).

Rosenberg, M., Confessor, N., and Cadwalladr, C. (2018). How Trump Consultants Exploited the Facebook Data of Millions. *The New York Times*. Article available at: https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html (accessed on Jan. 16, 2025).

Rossetti, M. and Zaman, T. (2023). Bots, disinformation, and the first impeachment of U.S. President Donald Trump. *PloS one*, 18(5), e0283971.

Rowe, F. and Mason, M. (2024). The role of social media and local deprivation in UK anti-immigration riots. Article available at: https://www.franciscorowe.com/post/uk-antimigration/ (accessed on Jan. 16, 2025).

Rubin, M. (2017). When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress. *Review of General Psychology*, 21(4), 308-320.

Ruths, D. and Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346,1063-1064.

Sanders, A. K. and Jones, R. (2018). Clicks at Any Cost: Why Regulation Won't Upend the Economics of Fake News. *The Business, Entrepreneurship & Tax Law Review*.

Sap, M., Dallas, C., Gabriel, S., Choi, Y., and Smith, N. (2019). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1668-1678.

Saveski, M., Roy, B., and Roy, D. (2021). The Structure of Toxic Conversations on Twitter. *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1086-1097.

Schliebs, M., Bailey, H., Bright, J., and Howard, P. N. (2021). China's inauthentic UK Twitter diplomacy: a coordinated network amplifying PRC diplomats. DemTech Working Paper. Programme on Democracy and Technology, Oxford University. Paper available at: https://ora.ox.ac.uk/objects/uuid:6c62aded-c0d0-41f5-887f-dd50cd43e467/files/s9019s333b (accessed on Jan. 16, 2025).

Settle, J. E. (2018). *Frenemies: How Social Media Polarizes America*. Cambridge University Press.

Shcherbakova, O. and Nikiforchuk, S. (2022). Social media and filter bubbles. *Scientific Journal of Polonia University*, 54(5), 81-88.

Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., and Tucker, J. A. (2021). Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. *Quarterly Journal of Political Science*, 16(1), 71-104.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, 22(11), 1359-1366.

Southern, R. and Harmer, E. (2021). Twitter, Incivility and 'Everyday' Gendered Othering: An Analysis of Tweets Sent to UK Members of Parliament. *Social Science Computer Review*, 39(2), 259-275.

Stevenson, J., Edwards, M., and Rashid, A. (2023). Analysing The Activities Of Far-Right Extremists On The Parler Social Network. *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM '23)*. Association for Computing Machinery, New York, NY, USA, 392-399.

Sun, Y. and Xie, J. (2024). Who shares misinformation on social media? A meta-analysis of individual traits related to misinformation sharing. *Computers in Human Behavior*, 158, 108271.

Sunstein, C. R. (2017), #Republic: Divided Democracy in the Age of Social Media. Princeton University Press.

Tan, C. (2022). Regulating disinformation on Twitter and Facebook. *Griffith Law Review*, 31(4), 513-536.

Tankala, S. (2022). Maintaining the Privacy and Security of Research Participants' Data. *NN/g*. Available at: https://www.nngroup.com/articles/privacy-and-security/ (accessed on Jan. 16, 2025).

Tappin, B. M., Wittenberg, C., Hewitt, L. B., Berinsky, A. J., and Rand, D. G. (2023). Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences*, 120(25), e2216261120.

The British Academy (2022). Academics top 'trust' list in British Academy poll. *The British Academy*. Available at: https://www.thebritishacademy.ac.uk/news/academics-top-trust-list-in-british-academy-poll/ (accessed on Jan. 16, 2025).

The Policy Institute, King's College London. (2020). Coronavirus: vaccine misinformation and the role of social media. *The Policy Institute*. Available at:

https://www.kcl.ac.uk/policy-institute/assets/coronavirus-vaccine-misinformation.pdf (accessed on Jan. 16, 2025).

Theocharis, Y., Barberá, P., Fazekas, Z., and Popa, S. A. (2020). The Dynamics of Political Incivility on Twitter. *Sage Open*, 10(2).

Tipoe, E. and Lee, I. (2024). Education and ideological polarization: Cross-country evidence and recommendations for higher education. *British Educational Research Journal*.

Torcal, M. and Magalhães, P. C. (2022). Ideological extremism, perceived party system polarization, and support for democracy. *European Political Science Review*, 14(2), 188-205.

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., DeHaven, A., Bochantin, J., and Borns, J. (2021). Study Preregistration: An Evaluation of a Method for Transparent Reporting. *Journal of Business and Psychology*, 36, 553-571.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. Paper prepared for the Hewlett Foundation. Available at: http://dx.doi.org/10.2139/ssrn.3144139 (accessed on Jan. 16, 2025).

Tucker, J. A. (2020). The Limited Room for Russian Troll Influence in 2016. *The Lawfare Institute*. Article available at: https://www.lawfaremedia.org/article/limited-room-russian-troll-influence-2016 (accessed on Jan. 16, 2025).

Tufekci, Z. (2018). YouTube, the great radicalizer. *The New York Times*. Article available at: https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html (accessed on Jan. 16, 2025).

United Nations. (2019). UN Strategy and Plan of Action on Hate Speech. Report available at: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf (accessed on Jan. 16, 2025).

Vaccari, C., and Valeriani, A. (2021). *Outside the Bubble: Social Media and Political Participation in Western Democracies*. Oxford Academic.

Van Horne, J. (2019). Shadowbanning is a Thing — and It's Hurting Trans and Disabled Advocates. *Salty World*. Article available at: https://saltyworld.net/shadowbanning-is-a-thing-and-its-hurting-trans-and-disabled-advocates/ (accessed on Jan. 16, 2025).

Vasconcellos, P. H. S., Lara, P. D. D. A., and Marques-Neto, H. T. (2023). Analyzing polarization and toxicity on political debate in brazilian TikTok videos transcriptions. *Proceedings of the 15th ACM Web Science Conference*. Association for Computing Machinery.

Vogels, E. (2021) The State of Online Harassment. *Pew Research Center*. Available at: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/ (accessed on Jan. 16, 2025).

Vogus, C. (2022). Improving Researcher Access to Digital Data. *Center for Democracy and Technology*. Available at:

https://cdt.org/wp-content/uploads/2022/08/2022-08-15-FX-RAtD-workshop-report-final-int.pdf (accessed on Jan. 16, 2025).

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Votta, F., Kruschinski, S., Hove, M., Helberger, N., Dobber, T., and de Vreese, C. (2024). Who Does(n't) Target You? Mapping the Worldwide Usage of Online Political Microtargeting. *Journal of Quantitative Description: Digital Media*, 4.

Vrielink, J. and van der Pas, D. J. (2024). Part of the Job? The Effect of Exposure to the Online Intimidation of Politicians on Political Ambition. *Political Studies Review*, 22(4), 1022-1041.

Wackenhut, A. F. (2018). Ethical Considerations and Dilemmas Before, During, and After Fieldwork in Less-Democratic Contexts: Some Reflections from Post-Uprising Egypt. *The American Sociologist*, 49(2), 242-257.

Wagner, M. (2021). Affective polarization in multiparty systems. *Electoral Studies*, 69, 1-13.

Wang, Y., Bye, J., Bales, K., Gurdasani, D., Mehta, A., Abba-Aji, M., Stuckler, D., and McKee, M. (2022). Understanding and neutralising covid-19 misinformation and disinformation. *BMJ*, 379, e070331.

Wanless, A. and Shapiro, J. N. (2022). A CERN Model for Studying the Information Environment. *Carnegie Endowment for International Peace*. Available at: https://carnegieendowment.org/research/2022/11/a-cern-model-for-studying-the-information-environment?lang=en (accessed on Jan. 16, 2025).

Ward, S. and McLoughlin, L. (2020) Turds, Traitors and Tossers: The Abuse of UK MPs Via Twitter. *The Journal of Legislative Studies*, 26(1), 47-73.

Webb-Williams, N., Casas, A., Aslett, K., and Wilkerson, J. D. (Forthcoming). When Conservatives See Red but Liberals Feel Blue: Labeler Characteristics and Variation in Content Annotation. *The Journal of Politics*.

Westwood, S. J., Grimmer, J., Tyler, M., and Nall, C. (2022). Current research overstates American support for political violence. *Proceedings of the National Academy of Sciences of the United States of America*, 119(12), e2116870119.

Who Targets. (2024) Available at: https://whotargets.me/en/uk-campaign-analysis-19th-25th-june/ (accessed on Jan. 16, 2025).

Wilfore, K. (2022). Security, Misogyny and Disinformation Undermining Women's Leadership. In *Gender and Security in Digital Space: Navigating Access, Harassment, and Disinformation*, Haciyakupoglu, G. and Wong, Y. (eds.). London: Routledge.

Williams M. L., Burnap P., and Sloan L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6), 1149-1168.

Williams, M. L., Burnap, P., Javed, A., Liu, H., and Ozalp, S. (2020). Hate in the machine: Anti-Black and Anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1), 93-117.

Wischerath, D., Godwin, E., Bocheva, D., Brown, O., Roscoe, J. F., and Davidson, B. I. (2024). Spreading the Word: Exploring a Network of Mobilizing Messages in a Telegram Conspiracy Group. *Conference on Human Factors in Computing Systems - Proceedings, Association for Computing Machinery*. New York, U. S. A., 2024 CHI Conference on Human Factors in Computing Systems, CHI EA 2024, Hybrid, Honolulu, USA United States, 11/05/24.

Wojcieszak, M. (2010). 'Don't talk to me': Effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *New Media & Society*, 12(4), 637-655.

Wojcieszak, M., Casas, A., Yu, X., Nagler, J., and Tucker, J. A. (2022). Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. *Science advances*, 8(39), eabn9418.

Wojcieszak, M., Menchen-Trevino, E., Clemm von Hohenberg, B., de Leeuw, S., Gonçalves, J., Davidson, S., and Gonçalves, A. (2024). Non-news websites expose people to more political content than news websites: Evidence from browsing data in three countries. *Political Communication*, 41(1), 129-151.

Wulff, J. N., Sajons, G. B., Pogrebna, G., Lonati, S., Bastardoz, N., Banks, G. C., and Antonakis, J. (2023). Common methodological mistakes. *The Leadership Quarterly*, 34(1), 101677.

X. Removal Requests. *X*. Available at: https://transparency.x.com/en/reports/removal-requests#2021-jul-dec (accessed on Jan. 16, 2025).

Yang, Y., Davis, T., and Hindman, M. (2023). Visual misinformation on Facebook. *Journal of Communication*, 73(4), 316-328.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringhini, G., and Blackburn, J. (2018). What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber? arXiv:1802.05287.

Zimmer M. (2010). Is it ethical to harvest public Twitter accounts without consent? Available at: http://www.michaelzimmer.org/2010/02/12/is-it-ethical-to-harvest-public-twitter-accounts-without-consent/ (accessed on Jan. 16, 2025).

Zimmer M. and Proferes N. J. (2014). A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66, 250-261.

Zimmerman, A. G. and Ybarra, G. J. (2016). Online aggression: The influences of anonymity and social modeling. *Psychology of Popular Media Culture*, 5(2), 181-193.

Zuiderveen Borgesius, F. J., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B., and de Vreese, C. (2018) Online Political Microtargeting: Promises and Threats for Democracy. *Utrecht Law Review*, 14(1), 82-96.